B. S. H. MITHRANDIR

PHYSICAL OCEANOGRAPHY (MOSTLY) BY DRAWING PICTURES

RESEARCH INSTITUTE OF VALINOR

PUBLISHED BY RESEARCH INSTITUTE OF VALINOR

/

Last compiled: April 2021

Foreward

When I was first approached to deliver a first undergraduate course on physical oceanography, but that I should refrain from using maths and calculus because of reasons, my first (somewhat knee jerk) reaction was to laugh in the senior professor's face and said "*no, tell the students to suck it up, you can't learn physics things without maths*". The senior, being a more professional and less confrontational type than myself, asked me to sit on it a bit and have a try. Long story short the result is this set of notes, which is to complement a course on descriptive physical oceanography with a focus on the *dynamical processes* that lead to some of the phenomena we observe in the ocean. The bias is on things I can explain by drawing pictures and particularly on *dynamics* (because that's the thing that interests me, and I make no apologies for this focus). Some of my own arguments for and against this kind of descriptive course are as follows:

✓ yes, you actually can learn a lot (about oceanography and the physical arguments) just by drawing pictures!

× no, drawing pictures is only qualitative and gives you very little "teeth" if you are wanting to e.g. make predictions or quantitative statements

- e.g. pictorial/descriptive arguments can tell you why you might have Western intensification, but to get the *Stommel boundary layer* dependence you need some maths (although the pictures might suggest what should be involved)

✓ it does help develop intuition (speaking as a maths person by training, I learnt a lot by writing this course)

? it highlights why to take physical oceanography further you really do need the maths, because intuition can sometimes be misleading (the devil really is in the details)

The selling points (or deficiencies depending on your point of view) of the content here are:

- ocean phenomena motivated, but united by *dynamics*[1]
- skimpy on facts and things to memorise, focus more on *logical deductions* (partly because I am very lazy and hate remembering things)
- broad but relatively little *depth*, though enough concepts to start on the other books in the literature to take it further maybe

[1] Personal gripe: Prior editions of *Descriptive Physical Oceanography* (Pickard & Emery, 5th edn. and before) do not do this, and in my opinion it makes the topics more complicated/disconnected than it needs to be.

This set of notes presents a pictorial/geometric way of looking at ocean dynamics (or, more generally, aspects of *geophysical fluid dynamics*) that might be beneficial for those wanting some exposure to concepts in physical oceanography without necessarily specialising in it (e.g. marine biologists, ocean engineers, ecologists etc.), and those who are already familiar with the maths wanting to see things in another way (e.g. symbol worshippers like myself). The notes are relatively self-contained and don't really assume any background really; relevant concepts in physics and maths will be recalled here but in a very skimpy manner, and only on the concepts that are directly used. Calculus is employed here but the focus is going to be on their *geometric* meaning. Symbol manipulations and calculations are almost non-existent, limited to order of magnitude estimates and working out some signs (e.g. negative times negative is positive); see the exercises for the more involved calculations. The notes here are not meant to be cover everything in physical oceanography, and there are many other better resources out there with a more comprehensive ocean focus (e.g. Talley et al. [2011], Williams and Follows [2011]) and/or with more focus on dynamics (e.g. Vallis [2006]).

As you may have also gathered, the writing style here is deliberately conversational/informal/unprofessional, and the drawings are cartoon-ish/coarse. The former is by choice, the latter is because (*a*) I am artistically challenged, (*b*) I want to highlight that these drawings are things you can (and should) do yourselves to convince yourselves of the logic and arguments behind them, and (*c*) I am lazy (this will be a recurring theme...) Some pictures and animations I did make using Python (via Jupyter notebooks), and the codes will be available on a GitHub repository (again, no promises on clean or Pythonic code, again laziness). Do what you like with the material, just keep it open source and non-commercial. If you find any errors, have suggestions or even want to contribute (!), feel free to open an issue or pull request on the GitHub repository. The document is prepared using the the Tufte-LaTeXclass[2], modified from the files[3] of Jody Klymak (University of Victoria), with additional editing based on the excellent Finite Element Methods course notes[4] of Patrick Farrell (University of Oxford).

[2] https://github.com/Tufte-LaTeX/tufte-latex/

[3] https://github.com/jklymak/Eos314Text

[4] https://pefarrell.org/teaching/

Contents

1 *"Big picture"*

This section goes through some facts and observations about the ocean, partly to argue why you might care about the ocean (if you need convincing), and partly to highlight the phenomena we will try to explain in the subsequent sections using pictorial arguments. Since the course attached to this set of notes was designed somewhat for people with minimal physics/maths background, there are two subsections at the en of this chapter that will highlight the bare minimum of the concepts that are needed to link the symbols used with the pictures drawn. The convention of symbols and a cheat sheet of sorts is given at the end of this chapter. Taking artistic license for setting the scene for telling a story, and attempting to strike a balance between an attempt to whet one's appetite but also (attempt as best as possible) remain somewhat technically accurate, concepts and terminology will be used here in this chapter, but not elaborated on in detail until later chapters.

Of course as the reader it is absolutely your prerogative to use/skip/reject the propaganda presented here as you see fit. Don't just assume everything is right either! I will have invariably made tpyos, or worst, *thinko*'s (I first saw the term in Vallis [2006]), so you should convince yourself the ideas and arguments actually have some value. If you can spare the time please report any tpyos, thinkos or suggested improvements on the GitHub repository where you got this document as an issue or, if you like, as a pull request.

1.1 *Oceanography, and physical oceanography*

The two general questions regarding the ocean to me are:

- *what* does the ocean look like?
- *why* does it look like the way it does?

The former encompasses *observations* (Ch. 7), not just for the physical quantities such as temperature, salinity and ocean currents, but also for biogeochemistry or other quantities such as phytoplankton concentration, oxygen content, plastic distribution, ecosystem behaviours, and so forth. The latter involves understanding the processes underlying the physical, but also chemical, biological, ecological processes, to name but a few. The two are not mutually exclusive of each other: that *what* requires the *why* for explanation, the *why* sometimes tells you *what* to expect, but the *what* acts as a constraint on the explanations of the *why*. Oceanographers tend to have a particular focus, on theory, modelling, observations, physics, biology etc., but, fundamentally, most things are intrinsically linked

one way or another, and the differing approaches compliment each other.

Physical oceanography focuses on the physics aspect, and in a nutshell (to me anyway) is the study of *how water in the ocean moves around*, i.e. *dynamics*. Making this "simplification" in a sense, we are in a slightly privileged position to know the governing equations for the phenomena (e.g. Ch. 1.4.3), which itself is a modification of Newton's equations in classical mechanics, from which we can in principle derive everything else. Of course we generally can't that in the general case, so we employ a variety of tools, such as approximations, numerical methods, and observation techniques, to help us in this venture.

The general approach here is that we will first highlight the features observed in the ocean to set the scene, but mostly talk about *dynamics*[1], with a focus on *intuition* by going through *pictorial* arguments, to convince you why the arguments might be true (again, the devil is really in the details, but we will not touch on those too much here). Fundamentally, physical oceanography is an *interdisciplinary* science at its core, benefiting from multiple approaches in order to chip away at the overall problem, such as that depicted in Fig. 1.1.

[1] The study of rotating stratified fluid dynamics is generally grouped under *Geophysical Fluid Dynamics* (GFD), which is applicable to studies of planetary atmospheres and/or oceans. It also has extensions to stellar atmospheres and interiors with suitable extensions (e.g. include magnetic effects).

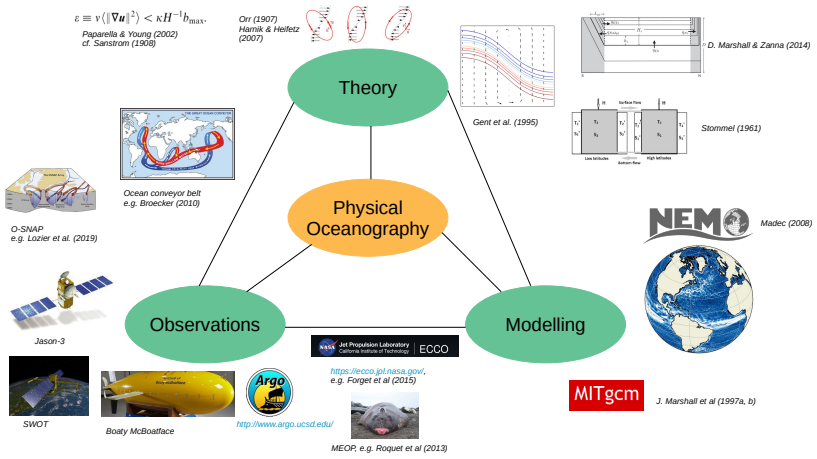


Figure 1.1: A schematic of how I see physical oceanography, like three sides of the triforce (if you know that reference), and all three parts are needed to make a whole (not going to comment which one represents wisdom, courage and power...) There are more developing branches coming out that doesn't quite fit neatly (e.g. data oceanography, though that could be somewhere between observation and modelling), and in reality most people are somewhere in between.

1.1.1 *Motivation: climate*

Here I give my own spin[2] as to why I think the marine environment and particularly why the physics is important. The study of the Earth's *climate* from a holistic point of view requires an understanding of the several "spheres", split for example as the *lithosphere* (solid Earth related things), *biosphere* (living things), *cryosphere* (ice), *atmosphere* (air), *hydrosphere* (the water stuff), and *anthroposphere*

[2] This is just one spin of why you might care about marine environments, and it's not the only one. A useful exercise would be to create your own depending on your interests, e.g. if you want to focus on stable isotopes.

(human things). These are not isolated subsystems of course: there is life in the ocean, ocean interacts with the atmosphere and ice and vice-versa, land boundaries and movements constraints and drives ocean dynamics, and so on. We are going to focus on the hydrosphere.

Fig. 1.2 shows a map of the globe but with an ocean focus rather than the usual land focus. From this point of view it is perhaps more convincing that the oceans actually covers around 70% of the Earth's surface. Two more attributes about the ocean:

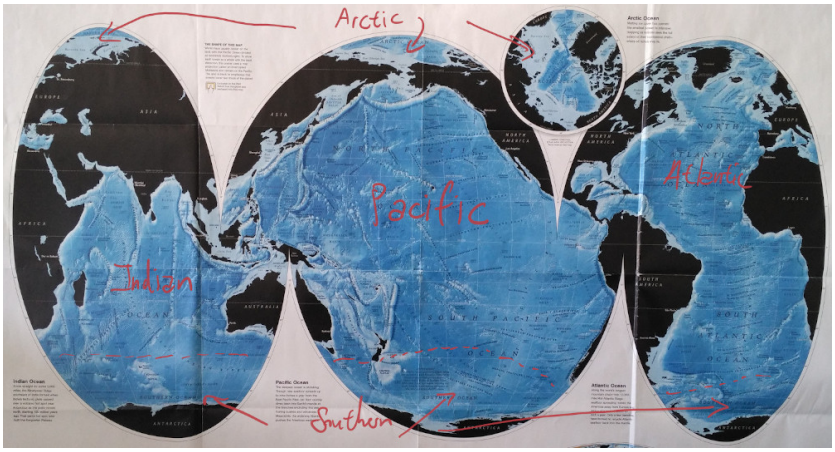


Figure 1.2: Ocean map focusing on the five major oceans by splitting the land. From National Geographic at some point.

1. the ocean holds around 50 times more *carbon*[3] than the atmosphere,

2. the upper 2.5 m holds as much *heat* as the atmosphere.

With these two observations, there are multiple points we can make for the marine environment being a central part of the climate system:

- The marine environment is a crucial component of the *carbon cycle*, and we care about that because carbon dioxide in the atmosphere is a *greenhouse gas*, and has consequences for the *energy balance* of the Earth. While we view the ocean largely as a sink for atmospheric carbon particularly for long-term storage (by the *biological* or the *physical pump*). This does not have to be the case and we want to know how the ocean evolves in the future, and what impacts this has for the carbon cycle.

- Related to above, the ocean is an important component of the *energy balance* of the Earth system acting largely as a heat reservoir absorbing a large amount of excess heat (see Fig. 1.3), partly

[3] Some inorganic (carbonates etc.) but largely organic, since most life on Earth is carbon based.

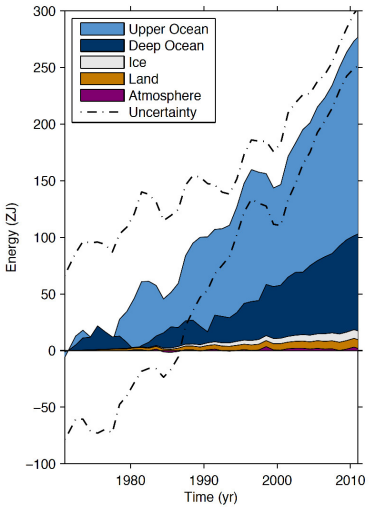


Figure 1.3: Figure 3.1 from the IPCC AR5 WG1 report, showing the destination of the excess energy received by the Earth. Most of it goes into the ocean.

because seawater has a much larger *heat capacity* than air and land (see Ch. 2). The *Meridional Overturning Circulation* (MOC, see Ch. 5) plays an important role in the transport of energy around the globe, and again we want to know how that evolves in time.

- If the ocean is warming because it is absorbing excess energy, then this has consequences for the following:
  - the density *stratification* and the MOC, which feeds back onto the the ability of the ocean in transporting/absorbing excess heat
  - *biogeochemical* cycles and its content, because surface warming is expected to lead to a strengthening of the stratification in the upper ocean, affect nutrient supply from the deep (arising from upwelling of colder, nutrient rich waters at depth), the ability of the water to hold chemicals (outgassing of dissolved oxygen and carbon dioxide because warmer water holds less dissolved gases), and others
  - general rise in *sea level*, since seawater larger than around 4° C expands when it is warmed up[4]
  - effects on the ocean ecology by physical stressors (such as increasing temperature) and/or biogeochemical stressors (such as decrease in nutrient supply, acidification, oxygen depletion[5])
  - consequences for economy via fisheries (food web consequences), shipping (changes in climate affecting the viability of certain route, or opening new routes in e.g. the Arctic), energy, ...

[4] Known as *thermosteric sea level*, sea level arising from thermal expansion. We revisit the point about 4° C in Ch. 2.

[5] Known as *hypoxia* and *anoxia*.

The points made for the importance of the ocean is non-exhaustive, but hopefully that provides a sample of the linkages to motivate the study of the physical aspects to do with marine environments.

1.2 Oceans

While the hydrosphere covers all things to do with water, for the purposes here we are going to mostly focus on the largest bodies of salty water (e.g. not touching on rivers and lakes here).

We first focus on **oceans**, which are taken to be the largest bodies of salty water that are bounded by **continental land masses**. These are the *Pacific*, *Atlantic*, *Indian*, *Southern* and *Arctic* ocean, sorted by surface area; these have been marked on the map given in Fig. 1.2. The first thing to highlight is that the oceans tend to be fairly deep, with an average depth of $H = 4000$ m over a significant area; contrast this to *seas* where the average depth is around 1000 m, even if they

can get very deep at a few locations (see Ch. 1.3). A schematic of oceans vs. non-oceans is given in Fig. 1.4, where we show a slice outlining the regions being referred to.

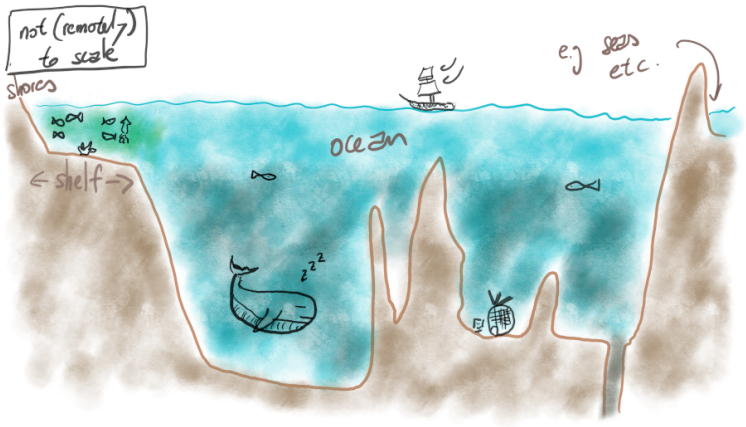


Figure 1.4: Schematic showing the oceans, shelf seas and smaller bodies of water. based on Figure 2.2 of Pickard and Emery [1990].

In the coean there is a massive discrepancy with the horizontal and vertical length scales: we are talking horizontal length-scales of at least $L = 1,000$ km (note the units!), so the **aspect ratio** H/L is small. The smallness of the aspect ratio has important dynamical consequences, which we will revisit in the subsequent discussion relating to dynamics. In some sense large-scale ocean dynamics turns out not to be so dissimilar to large-scale atmospheric dynamics, bar the important difference on lateral boundaries provided by the land[6]. The effects of boundaries lead to phenomena unique to the ocean that we will highlight and explain in due course (Ch. 4 and 5).

[6] Except in the case of the Southern Ocean (Ch. 5.1) where there are open latitudes that are unblocked.

Before we go on to talk a little about the five oceans, going to quickly introduce some terminology. The land features are normally referred to as **topography** and/or **bathymetry**. Technically the former refers to features above land and the latter refers to those features below sea level, but sometimes both terms are used in the ocean community. Some notable *bathymetric features* are labelled in Fig. 1.5. There are somewhat technical (though by no means completely universal) definitions for these features, but I am of the opinion that most of the definitions are not hugely important to the narrative here[7], so I refer the reader to other sources (e.g. Wikipedia, Pickard and Emery [1990], Talley et al. [2011]). The main point is that there are mountains and hills underneath the ocean that has significant influences on the dynamics, much like the case of the atmosphere, and we will need to take into account of the boundaries when talking about ocean dynamics and circulation.

[7] We will however say a bit about shelves and continental slopes.

Most of the following discussions focuses on highlighting dynam-

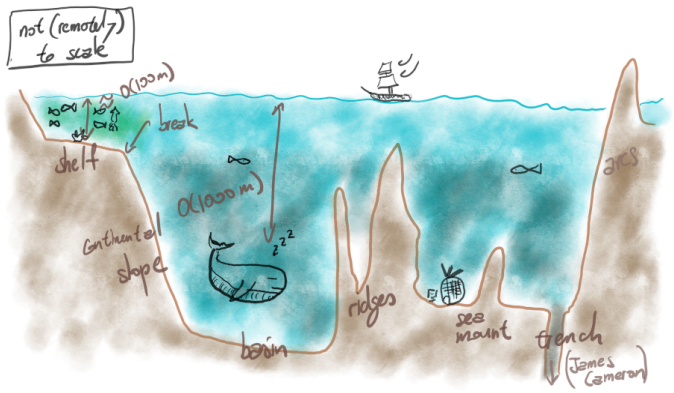


Figure 1.5: Some bathymetric features of note in the ocean. Figure based on Figure 2.2 of Pickard and Emery [1990].

ical features rather than going into details such as how many square km the particular ocean covers or what is the average temperature etc., unless they contribute to the narrative (these are covered in other books and can be searched for in Wikipedia).

1.2.1 *Atlantic Ocean*

We are actually going to start first with the Atlantic even if the Pacific is the biggest by surface area coverage. The Atlantic has traditionally received much more attention than the Pacific perhaps for the following reasons: ocean science was first systematically studied in Europe and North America; the Atlantic is perhaps scientifically more interesting because of the role of the *Atlantic Meridional Overturning Circulation*, its links with the global circulation, and its contribution to the weather/climate; the Atlantic is easier to observe and navigate simply because it is narrower than the Pacific. The Atlantic neighbours Europe, Africa, and the Americas, and is connected to the Arctic ocean to the North, the Southern Ocean to the south, and can be seen in Fig. 1.6.

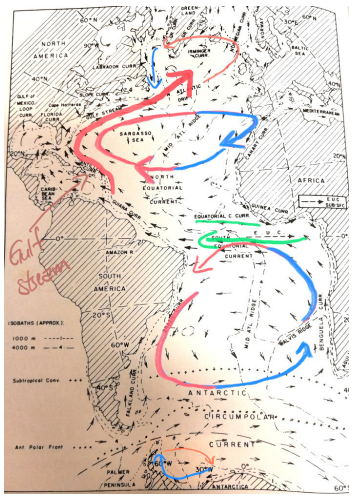


Figure 1.6: A marked up figure of the Atlantic ocean detailing the gyres and the Western Boundary Current. Modified Figure 7.9 from Pickard and Emery [1990].

Also marked onto Fig. 1.6 are some of the surface circulation features of note, such as the *equatorial currents* (the green arrows near the equator), the *gyres* (the recirculating currents), and the *Western Boundary Current* known as the *Gulf Stream* (the big red arrow on the America side going from Equator towards the North Pole). The **equatorial currents** are fairly fast east to west flowing currents in the low latitudes ($\pm 20^\circ$ N/S) that are largely driven by the *trade winds*, though there is a equatorial counter current that is normally slightly north of the equator, but goes the other way (against the wind).

The two big **gyres** immediately north and south of the Equator are known as the **subtropical gyres**, and these rotate clockwise and anti-clockwise in the Northern and Southern Hemisphere respec-

tively. The smaller gyres at higher latitudes are the **subpolar gyres**; technically in the Atlantic the subpolar gyre is in the north, since the subpolar gyre in the south belongs to the Southern Ocean. The Atlantic subpolar gyre rotates anti-clockwise[8]. The gyres acts to bring warm equatorial waters to higher latitudes and returning colder waters towards the equator.

[8] In both hemispheres the subtropical gyres rotate *anti-cyclonically* while the subpolar gyres rotate *cyclonically*; see Ch. 3.2 and 4.

The **Western Boundary Current** (WBC) in the Atlantic is called the *Gulf Stream* and is a particularly intense current going north-eastwards from the Gulf of Mexico towards the poles, and thus transporting warm equatorial waters towards the Western Europe. The Gulf Stream has a particularly important role for shipping during the Age of Discovery when the Europeans were colonising America: going from Europe to the Americas, it is normally preferable to sail down to Africa before going across (making use of the intense Westward *trade winds* at the equator), while coming back with goods (loot?) it makes sense to make use of the Gulf Stream to speed up the journey.

The Gulf Stream is a fairly narrow current (typically identified by the warm water it carries around), with an across-stream extent of around 100 km and extends down to around 1000 m depth. The speed of the current decreases with depth but at the surface it can get up to 2.5 m s^{-1}. This may not sound much, especially relative to atmospheric winds, but note that depth-average speeds of mean flows in the ocean are normally measured in cm s^{-1}, so the Gulf Stream surface speeds may be two orders of magnitude faster than average ocean flows. The Gulf Stream transports around 30 Sv[9], which is particularly notable given that the current takes up relatively little volume (cf. the *Antarctic Circumpolar Current* in the Southern Ocean).

[9] 1 Sv $= 10^6$ m^3 s^{-1}. The unit of Sverdrup is named after the Norwegian oceanographer Harald Sverdrup (1988-1957). For reference, the Amazon river has the world's largest discharge of freshwater into the ocean and the average discharge rate is around 0.2 Sv.

Since seawater is better than air in retaining heat due to water having a larger heat capacity, and that the Gulf Stream transports a significant amount of warm water towards Western Europe, the Gulf Stream has a significant impact on the weather and climate in Western Europe. The trivia to note is that while London is about 10° higher in latitude than New York (around 51° N and 41° N respectively), we might have expected London to be quite a bit colder especially in the winters (because the higher latitudes receives less sunlight). In fact it almost barely ever snows in London, yet New York gets particularly nasty snow storms and the average temperature in the winter months are sub-zero. The warmer Gulf Stream water carries heat with it as it transverses the Atlantic, and releases a portion of its heat to the atmosphere as it reaches the colder higher latitude air, leading to the more temperate climate observed in Western Europe.

The Gulf Stream should in some sense be seen as *averages/means* with a strong instantaneous signal. In reality the current meanders around, evolves over multiple time-scales (seasonally with the winds and solar forcing, longer time-scales with intrinsic variability), and on top of the mean signal there are smaller-scale fluctuations. A snap shot of the Gulf Stream is given in Fig. 1.7 showing sea surface temperature. Within the figure we can clearly identify meanders within the main current (as the main body of red), as well as loss of coherency towards the North East (cf. a water hose that is splattering around). The breaking of the jets arises from *instabilities* (cf. Ch. 6), which leads to what is referred to as **eddies**, for the moment to refer to mean closed regions of re-circulation (e.g. the blobs of green and red to the south and north of the mean current respectively)[10]. These eddies, which are effectively the ocean analogue of atmospheric high and low pressures, trap water from its location of generation and carry the water with it as it moves around, and can themselves interact with each other and lead to additional phenomena (e.g. merging, further breaking, forcing of the mean state).

So what drives the gyres and the WBCs? Why is it *Western* and not *Eastern* boundary currents? Given the WBC it seems there is a net transport at the surface to the poles, so where is the return flow? What leads to eddies and what do/can they do? Answers to these questions are sketched out in Ch. 3, 4, 5, and 6.

[10] There is some argument as to what really should be called 'eddies'. To quote Ryan Abernathey (or at least that's where I first heard it described this way), should we treat *eddy* as a 'noun' (the circular-esque blobs) or a 'verb' (the fluctuations about the mean)? I would say this distinction is not as clear as it should be when the term 'eddy' is used in the community. I personally take the latter view (the 'verb' includes the 'noun' but not vice-versa).

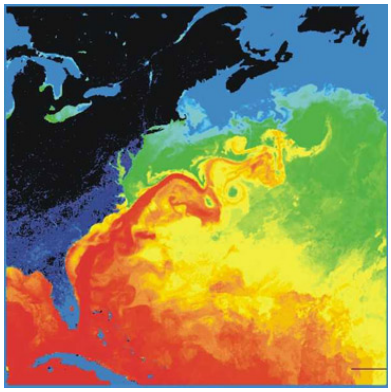


Figure 1.7: NASA observation of the Gulf Stream, showing the sea surface temperature (warm waters in red and cooler waters in blue). Image taken from Wikipedia.

1.2.2 *Pacific Ocean*

The Pacific is the largest ocean in the world and covers around 46% of the Earth's surface; historically this has meant it was difficult to chart and navigate. The Pacific mainly neighbours the Southern Ocean, but is also connected to the Indian ocean via the Indonesian Archipelago, and to an even lesser extent the Arctic ocean via the Bering strait to the north; the Bering strait is very shallow and water transport between the two oceans is limited.

The surface circulation features are highlighted in Fig. 1.8, and largely similar to the Atlantic ocean, in that there are equatorial currents, subtropical gyres at lower latitudes, subpolar gyres at higher latitude in Northern Hemisphere (the Southern one belongs to the Southern Ocean), and the **Kuroshio** WBC off the coast of Japan. One aspect we do highlight is the **Eastern Boundary Upwelling System** (EBUS) associated with the *Peru current*[11]; the analogous one in the Northern Hemisphere in the Pacific is called the *California current*. The difference with WBCs are that EBUS currents tend to be relatively shallow, slower flowing, and take colder water from the poles towards the Equator. These currents are associated with

[11] Sometimes the *Hulmboldt current*, after the Prussian naturalist Alexander von Humboldt (1769-1859).

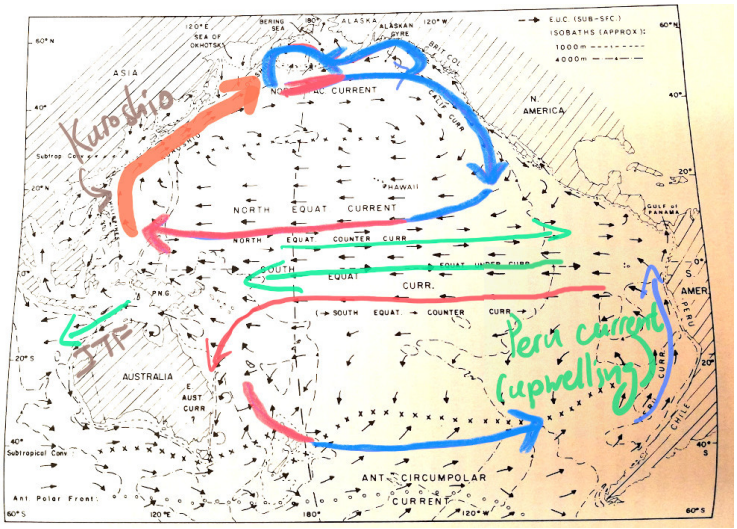


Figure 1.8: A marked up figure of the Pacific ocean detailing the gyres, the Western Boundary Current and the Peru current, which is part of a Eastern Boundary Upwelling System. Modified Figure 7.31 from Pickard and Emery [1990].

Ekman upwelling (Ch. 3.3.3), bringing cold nutrient-rich waters up towards the surface, and thus have important consequences for marine ecology and fisheries.

In the Pacific a particularly interesting climate phenomenon called the *El-Niño Southern Oscillation* (ENSO)[12] occurs in the tropical Pacific. This phenomenon is fundamentally one of the atmospheric-ocean coupled system[13]. El-Niño events occurs between two to seven years, generally seen with the reversal of the trade winds in the Pacific and an increased warming of the sea surface temperature in the region bordering the East Pacific. Associated with El-Niño events are increased precipitation (which can lead to severe landslides) and reduction in fishery yields (deepening of the *thermocline* in the Eastern Pacific; Ch. 2.2.1), while the Western Pacific gets much less precipitation, which can lead to drought-like conditions. El-Niño also appears to have an influence on global weather patterns, and has been argued to have contributed to:

- human sacrificial practices in the Aztecs
- the French Revolution (1789-1799; e.g. Grove [1998], and references within)
- the spread of Christianity in Qing dynasty China (associated with the Great North China Famine in 1876-1879)
- the Arab spring (2010-2012; Kelley et al. [2015])

These theories are quite entertaining to hear about and there are perhaps some plausibility to thems, but of course *correlation does not imply causation*, and I will leave it at that. This document does not touch on a mechanistic rationalisation of the ENSO phenomenon

[12] *El Niño* literally means *the boy* in Spanish, a reference to the Christ child, since El Niño events tend to happen around Christmas time.

[13] Via the *Bjerknes feedback*, named after the Norwegian-American meteorologist Jacob Bjerknes (1897-1975). Not to be confused with his father the Norweigan meteorologist Vilheim Bjerknes (1862-1951), who is known for his contributions to weather forecasting.

(it's generally agreed *waves* are important in the mechanism; more references in Ch. 6.1).

1.2.3 *Indian Ocean*

The Indian ocean borders mainly the Southern Ocean, and has connections with the Pacific and the Atlantic through the leakages from the *Aghulas current* (which is suggested to be the world's largest WBC with transports of around 70 Sv). There is only really a subtropical gyre in the Indian ocean because of land boundaries, and that associated subpolar gyre belongs to the Southern Ocean. Note that there are equatorial currents and smaller gyres towards the coast of Indian and Persian Gulf, which reversing seasonally because of changes in the wind forcing from the seasonally varying *monsoon winds* (Ch. 3.3).

One interesting aspect to note is that the Atlantic is seen to be saltier than the other oceans (e.g. look forward to Fig. 2.9), and part of this is attributed to the leakage of the Aghulas current, leading to a transfer of warm, salty water through the Southern tip of Africa into the Atlantic (see e.g. Fig. 1.10). This is perhaps interesting if you think about it: the Aghulas current is a WBC so the flow goes to the East, and there is the neighbouring *Antractic Circumpolar Current* which is also going to the East, so the scenario is stacked against Westward transfer, yet it exists. As can be seen from Fig. 1.10, these *Aghulas rings* can maintain its coherency and travel quite far into the Atlantic basin. This route of transfer is known as the *warm route* (as opposed to the longer *cold route* going eastward all the way round the globe to reach the Atlantic), and is a component with the global MOC (more in Ch. 5).

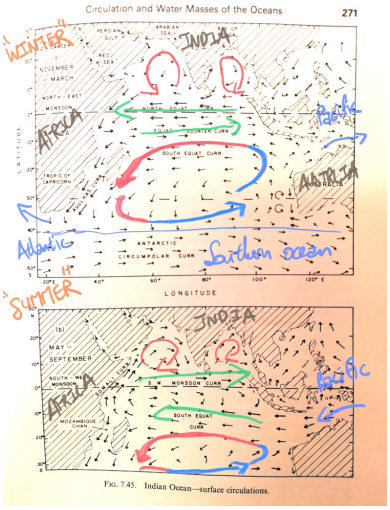


Figure 1.9: A marked up figure of the Indian ocean detailing the gyres and the equatorial currents in Summer and Winter. Modified Figure 7.45 from Pickard and Emery [1990].

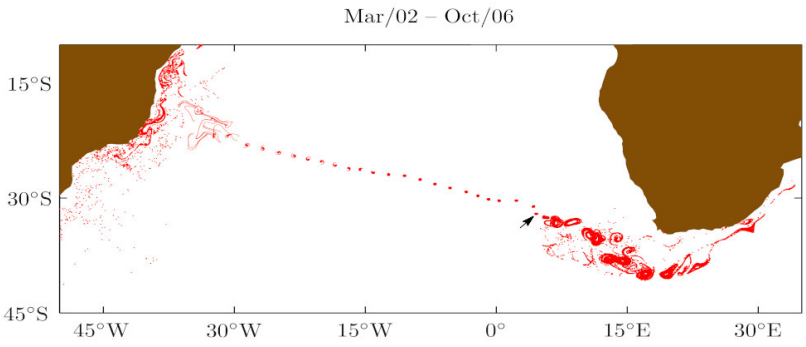


Figure 1.10: Tracking of Aghulas eddies. Image from Yan Wang (HKUST).

1.2.4 *Arctic Ocean*

The Arctic ocean is one of the smaller oceans (some people call it a margin sea to the Atlantic), and neighbours the Atlantic, and marginally connected to the Pacific through the Bering strait between North America and Asia.

The main thing in the Arctic is of course the presence of *sea ice*, which has thermodynamic as well as mechanical consequences for the dynamics. When sea ice forms because it is cold, the ice rejects the salt (*brine rejection*), leading to an increase of salinity in the region below the ice, which may have dynamical consequences (cf. *double diffusion* in Ch. 6.2). The presence of ice provides some shielding of the ocean from direct wind forcing, which reduces the momentum transfer to the water column by the atmosphere, but also provides an extra surface for the already flowing water underneath to rub against[14]. On the other hand, when ice breaks up or melts, the ocean is exposed to the cold temperatures and the winds in the atmosphere, leading to a loss of ocean heat and thus buoyancy *loss* (i.e., water gets colder and more dense), as well as momentum transfer into the ocean leading to stronger currents. The seasonal changes in the wind as well as the ice cover leads to interesting dynamics in the Arctic region as well as circulation features such as the *Beaufort gyre*.

There has been recent talk that, with the decrease in sea ice cover in the Arctic, new shipping routes could open (e.g. Aksenov et al. [2017]). The existing shipping routes that use the Bering strait to go between the Atlantic and the Pacific go along the coast of Russia, and the route is somewhat hazardous. If however the sea ice cover decreases enough there could be some incentive to just go straight through the Arctic. Although just because we could doesn't mean we should: while this is of course sound from an economic point of view, the associated consequences of loss of sea ice cover presumably will be catastrophic, that any gain will probably seem insignificant in comparison...

Figure 1.11: A marked up figure of the Arctic ocean detailing the Beaufort gyre and some of the currents. Modified Figure 7.26 from Pickard and Emery [1990].

[14] Referred to as the *ice-ocean governor mechanism*; Meneghello et al. [2018].

1.2.5 *Southern Ocean*

Last but certainly not least is the Southern Ocean, which is connected to all the major oceans except the Arctic. The boundary of the Southern Ocean is somewhat ill-defined, but is usually done by *fronts* separating regions with different *watermass properties*, i.e. somewhat sharp boundaries that separates water types with different temperature, salinity, or other *tracer* properties (see Ch. 2 and 5.2.2). The Southern Ocean experiences some of the strongest wind forcing in the world, and is influenced somewhat by the fact that there is sea and land ice extruding from Antarctica. Since the Southern

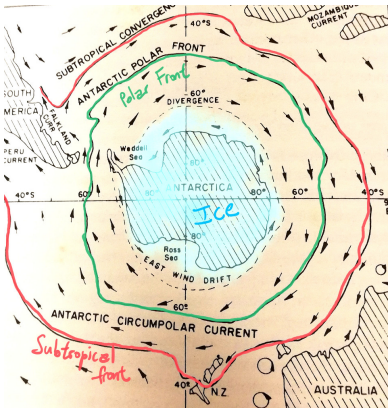


Figure 1.12: A marked up figure of the Southern Ocean, highlighting two of the fronts that roughly denote boundaries with significant difference in watermass properties. Modified Figure 7.45 from Pickard and Emery [1990].

Ocean is connected to all the major ocean basins, it is sometimes regarded as the center of the global ocean, and plays a central role in the global MOC (more in Ch. 5). For example, the subpolar gyres exist in the Southern Ocean (e.g. the *Ross* and the *Weddell gyre*), and these interact with the abyssal water forming regions responsible for producing the densest waters that fill the deep and abyssal regions of the ocean. It is an ongoing question as to who the interplay between wind, ice, thermodynamic forcing affect the intrinsic dynamics in the Southern Ocean, and its subsequent impacts on the global MOC.

Probably the most significant difference between the Southern Ocean and other oceans is the presence of *open latitudes* where there is no north-south land boundaries. The dynamical balances are noticeably different to the gyres, and the resulting dynamics within the Southern Ocean actually bears resemblance to atmospheric dynamics (see Ch. 3 and 5). Partly because of the open latitudes the Southern Ocean possesses the largest current in the world, the *Antarctic Circumpolar Current* (ACC), with a transport of around 130 Sv. While the current is not as intense as the Gulf Stream, it is much larger in terms of cross-stream and vertical extent, leading to the much larger transport. The ACC travels through a choke point at the *Drake passage* (between the tip of South America and Antarctica), then lurches northward and joins the *Brazil current* (a weaker WBC coming off South America), before proceeding around the globe whilst being steered by the bathymetry (such as the *Kerguelen plateau*, between Africa and Australia, closer to Africa). The ACC, being a strong current and subject to strong wind forcing, is a very dynamic and turbulent current, possessing surface *waves* that can have very large amplitudes[15] as well as being susceptible to lots of *instabilities*, which in turn has consequences for things like *air-sea exchanges*, *momentum/energy transfers* and, in turn, the global MOC. We will revisit these in Ch. 5 and 6.

[15] We are talking wave heights measured in meters; see for example `https://youtu.be/WQUXbkAdZhg` one of these surface waves battering a naval ship.

1.3 *Not oceans*

1.3.1 *Some terminology*

In contrast to oceans, the smaller bodies of water are more ambiguous to define (because there always seems to be exceptions to the rule), so for the purposes here are simply to be referred to as "not oceans", to include *seas*, *shelf regions*, *estuaries*, *lakes* etc. What is fairly unambiguous is the regions separating the ocean that are on average fairly deep, to the shallower non-ocean regions that have an average depth of 1000 m or less, which are the **continental slopes**. While Fig. 1.13 draws these slopes as very steep, you have to remember

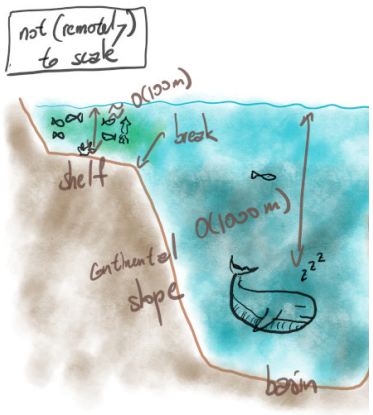


Figure 1.13: Zoomed in version of Fig. 1.4, focusing on the shelf regions. Based on Figure 2.2 of Pickard and Emery [1990].

the aspect ratio is massively exaggerated, the *gradients* (the slopes; Ch. 1.5.2) are actually numerically small, at around 1/20 (or around 3° is you want think of angles). While we might expect that the aspect ratio H/L is smaller in not-oceans because H is smaller, this is in fact not true since the horizontal extent L of these non-ocean regions are significantly smaller. The aspect ratio is still small but no longer as small, which has some consequences for the dynamics (not really touched on here).

Seas are smaller bodies of salty water that typically have a connection with the ocean and are partially enclosed by land (**semi-enclosed seas**), but of course exist, such as the *Black sea*, which may be regarded as essentially covered by land (so an **inland sea**). **Shelf Seas** are seas over shelf regions, which are particularly important for biogeochemistry and ocean ecology, due to the presence of nutrients such as from river runoff and/or *Ekman upwelling* (Ch. 3.3.3), and the fact that the region is relatively shallow and light penetration provides the necessary ingredients for *primary production*. A statistic to note is that while shelf seas cover around 8% of the global area, they account for around 15 to 20% of the ocean's primary productivity, and is a particular reason why all major fisheries in the world are in the shelf sea regions.

Fun trivia: the following are not 'seas' even though they have 'sea' in their name (they are technically lakes because they lie on land):

- *Sea of Galilee* (of the biblical fame) in northern Israel near the border of Jordan, which is a freshwater lake and is sometimes referred to as *Lake Tiberias*
- *Dead Sea* (of the cosmetics fame?) between Israel and Jordan, which is a salt lake (near the lowest point on land, see Fig. 1.14)
- *Caspian Sea* (of the caviar fame) north of Iran and south-east of Ukraine, which is also a salty lake

Figure 1.14: Picture taken near the Dead Sea (visit in 2014).

Figure 1.15: (left) Estuary in the Amazon; from Science Photo Library. (right) Pearl River Delta showing Humen bridge; from Wikipedia, user Tung Wu.

Estuaries are usually defined to be areas around the river mouths (e.g. Fig. 1.15) where there is influence from both *freshwater runoff* from the rivers and the salty water from the seas/oceans. These regions tend be very shallow and the effects of *tides* are very prominent, leading to *tidal excursions* that result in a particularly notable signal in the salinity (more precisely the boundary between freshwater with low salinity and water with high salinity). Notably these regions are generally near settlements, and thus experience forcing from human activity such as nitrogen loading, pollutants, heavy metal input, soil erosion, and so forth, which puts pressure on the ecological activity that exists in these regions. There are several ways to classify these based on the geometry as well as the watermass

properties (see Chapter exercise of Ch. 2 or just do it here).

Just a few more. **Fjords** is one of the famous Scandinavian exports to the English language. These are inlets with steep sides carved by *glaciers*, so fjords are normally only found in higher latitudes. Fjord waters are usually salty, for example, the many fjords that are dotted around the coast of Norway, although there are ones in the US Great Lakes that are not. Much like the use of the term 'seas', the use of 'fjords' is not entirely uniform either, and the Scandinavian use of the term is much more liberal. **Lakes** as mentioned above are bodies of water that are on land, and most of these are freshwater ones (e.g. the famous one in Scotland given in Fig. 1.16) but again with some famous exceptions already noted above.

There are other terms such as *lagoons*, *gulf*, *coves* etc., and we can go on all day, but perhaps lets go to some case studies instead of throwing words around... Below we give some case studies relating to not-oceans to highlight some features of interest in these smaller bodies of water, paying particular attention to the physical aspects. The list is non-exhaustive and clearly biased (e.g. I've ignored *corals* partly because I don't know much about it).

Figure 1.16: What lake am I? Pictures originally from Royal Caribbean website and TopPNG.

1.3.2 *Case study 1: Mediterranean Sea*

The Med Sea (I'm going to be lazy) is the big body of water between Europe and Africa, and is connected to the Atlantic via the Strait of Gibraltar (a **strait** is a narrow passage of water connecting two bodies of water, e.g. Bering Strait for Atlantic and Pacific, Strait of Malacca for South China Sea and Indian ocean). The Med Sea has traditionally be a very important area for travellers, warmongers, and merchants, enabling transport between Europe, Africa and to the Middle East, as well as providing an important source of food for the civilisations around the area.

The Med Sea has an average depth of around 1500 m but can

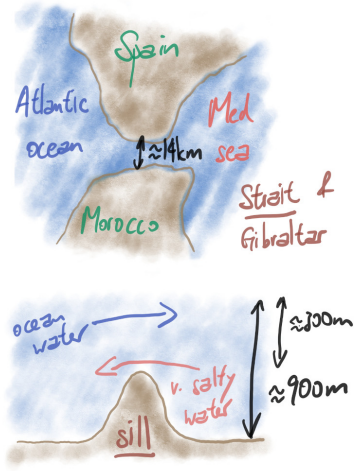


Figure 1.17: Schematic of the Med sea and its connection with the Atlantic via the Strait of Gibraltar. Schematic demonstrating the sill and resulting constraints on water exchanges.

get up to several thousand meters deep. However note that the connection with the Atlantic is choked at the Strait of Gibraltar, horizontally because the strait is only 14 km wide, and vertically because there is a **sill** that protrudes from the ocean bottom, leading to a **sill depth** (the depth between ocean surface to the top of the sill) of around 300 m. This limiting factor in the exchange results in the Med Sea waters having a particular signature (the *watermass property*). The Med Sea, being located just north of the equator, falls within the region of the *subtropical high* on the edge of the *Hadley cell*[16]. What this means is that this is a region of regular high atmospheric pressure, which suppresses atmospheric convection, the formation of clouds, and thus limits precipitation. Given this area is also warm, the Med Sea typically experiences strong *evaporation*, leading to water that is generally warm, but also very salty.

[16] A component of the atmospheric overturning circulation. More in Ch. 3.3

Occasionally there will be a cold burst of continental air coming from Europe, which leads to a cooling of the Med Sea water. Now, slightly cooler water and very salty water is very dense (Ch. 2.3) has the capability to sink, and the salty water starts filling up the bottom of the Med Sea. While the water may circulate around within the Med Sea, the fact that there is a sill at the Strait of Gibraltar means this water really has nowhere else to go[17], and piles up behind the sill. The presence of the sill allows for the relatively warm and very salty water characteristic of the Med Sea to build up, until it spills over into the Atlantic in the form of *overflows* (think under water waterfalls). Relative to the less salty and cooler water of the Atlantic, the Med Sea water is more dense, and ends up contributing somewhat to the *North Atlantic Deep Water* that sits at the mid-depths of the Atlantic[18].

[17] The Med Sea, being shielded by land, means tidal effects are weak, and the associated internal waves and diapycnal mixing is expected to be weak too; see Ch. 5 and 6.

[18] The deepest and densest water however originates from the Antarctic and is very cold. Part of the reason might be that the overflows out of the Med Sea leads to significant mixing (Ch. 3.4 and 5) so the resulting water is not as dense.

1.3.3 *Case study 2: Labrador Sea and Weddell sea*

Given we just talked about the Atlantic and *North Atlantic Deep Water* we first say a few things about the Lab Sea (still going to be lazy). The Lab Sea is just south west of Greenland, and given the high latitude location experiences very cold atmospheric temperatures, which is a prime site for forming cold and dense water. Recall from the text in Ch. 1.2.1 that, from the surface observations of the currents, mass seems to be converging polewards, so for mass conservation reasons the water has to return somehow even though there is no notable evidence for this at the surface. Well if it doesn't return southward at the surface it could do it at depth, and it turns out the Lab Sea is one of the areas that contributes to sinking of waters that subsequently flow back to the south (fueling the *Deep Western Boundary Current* below the Atlantic WBC). A schematic of

the circulation is given in Fig. 1.18.

The Lab Sea contributes to the deep water formation via intense cooling of the water particularly in Northern Hemisphere winter, leading to *deep convection* that transports water over large vertical depths, which is notable since *convection* due to unstable density gradients in the ocean is usually shallow (Ch. 2). One thing which is beyond the discussion here is that the deep convection in the Lab Sea is in combination with interesting effects such as *cabbeling*, arising from the *nonlinear equation of state* (Ch. 2.3), where mixing of two different water parcels actually leads to a water parcel that has a higher density than the simple addition of the individual parcels' densities together[19].

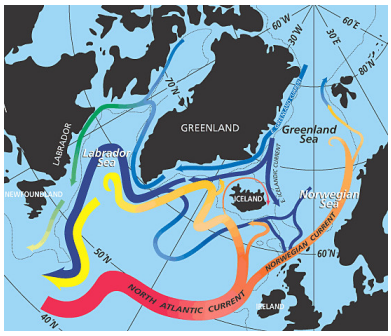


Figure 1.18: Schematic of North Atlantic Circulation. Taken from Wikipedia, image from Jack Cook at WHOI.

[19] cf. Metaphorically, cabbeling leads to $1 + 2 > 3$ in terms of density.

Something analogous happens in the Weddell Sea in the Southern Ocean. Water reaching the Antarctic is cold, but the Antarctic atmosphere is colder still. The intense cooling of the water again leads to formation of cold dense water. In a process similar to the Med Sea, there is a sill holding the water back, which eventually overflows and fills the abyssal ocean with the densest waters in the ocean, the *Antarctic Bottom Water*. Note that of course the water doesn't just carry temperature and salinity around with it, but also chemical tracers and in particular dissolved carbon, so the sinking of these cold waters into the ocean abyss contributes to storage of carbon. Questions then arise as to if and how this abyssal water gets upwelled (some of this in Ch. 5), how this part of the MOC might evolve, and what consequences this could have for the global carbon cycle.

1.3.4 *Case study 3: Black Sea*

Changing gears a bit, we go back to around the Med Sea region to visit the Black Sea. There are several competing theories as to why the Black Sea is 'Black' in the first place. One rationalisation I will take here for the sake of the story is to do with the fact that the Black Sea might have been a somewhat treacherous place to navigate, and shipwrecks used to occur in the region quite a bit before modern day. The particular interest here is that these shipwrecks are well-preserved without decay, covered in black sludges. The deeper parts of the Black Sea seems devoid of life, and the water smells very strongly of sulphur, which in Christianity is suggestive of the description of hell with its fire and brimstone (brimstone is another name for sulphur).

The above observations are now known to be because of **anoxia**, referring to the absence of oxygen content in the water, and that the physics contributes significantly to this phenomenon. The Black Sea

is fed by rivers around it, but also from the Med Sea through the shallow connection at The Bosphorus (or sometimes the Strait of Istanbul). The discrepancy in density between the freshwater from the rivers and the salty water from the Med Sea leads to the salty water sinking and filling up the bottom parts of the Black Sea, while the fresh part stays near the surface, as in the schematic in Fig. 1.19.

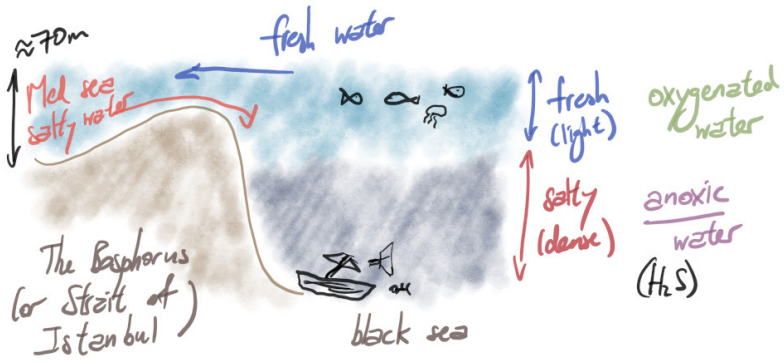


Figure 1.19: Black sea schematics water property schematic.

What this leads to is a strong vertical density gradient that is very difficult to erode, inhibiting mixing and thus vertical transfer of stuff (including oxygen). Any oxygen that was in the deeper parts is used up, which ends up forcing the organisms that use oxygen for *aerobic respiration* to go up the water column (assuming it is not in their nature to want to die by suffocation), consistent with the deep parts of the Black Sea being seemingly devoid of life. However of course that's just visible life. It turns out what remains behind are the micro-organisms that undergo *anaerobic respiration*, such as hydrogen sulphide producing bacteria, which explains the bad smell[20]. The lack of oxygen content means *oxidation* of material is limited, which allows wreckage to remain somewhat intact over longer periods of time[21].

[20] When eggs rot some of the proteins break down into hydrogen sulphide, associated with that rotten egg smell.

[21] Rescue missions in 2018 AD have uncovered an almost intact Greek merchant vessel dating back to around 400 BC.

1.3.5 *Case study 4: South China Sea*

As a final example, the South China Sea is one of the regions that seems to have everything thrown in. It is the largest marginal sea in the region, with the main connections to the oceans via the Luzon, Taiwan and Midoro strait (to the Pacific ocean) and the Malacca strait (to the Indian ocean); see Fig. 1.20. Because of the geographical set up, the region is one of the busiest areas for shipping activity for transport of goods between Asia and the rest of the world.

The region includes shelf regions, as well as deep regions (going down to around 4000 m), so coastal, shelf and ocean dynamics

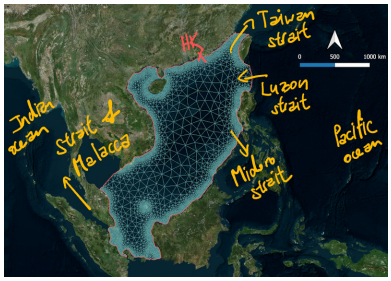


Figure 1.20: Marked up version of the South China Sea, with a triangulated mesh over it (for use with Finite Element models). Mesh and original diagram from Chinmayee Mallick.

interact and feedback onto each other. The region, though relatively small in terms of surface, has been claimed to account for around a third of the world's biodiversity. Fisheries thrive in the area but there are concerns that human activity are stressing the marine ecology in the region (such as nutrient loading, shipping activity, pollution, and over-fishing), leading to various action in the surrounding countries, which contributes to the political tension in the region.

On a regional circulation point of view the Luzon strait connection is probably the dominant factor since it has a sill depth of around 1000 m, while the other straits are shallow in comparison. There is thus the possibility of interaction between the ocean and regional circulation. In additional to the oceanic connections, the region in addition is strongly affected by Asian monsoon winds (Ch. 3.3), which leads to a seasonally varying driving by the regional circulation.

One interesting aspect that is being revisited recently is the regional circulation patterns seems to display a 'sandwich' pattern in the vertical, i.e. a anti-clockwise–clockwise–anti-clockwise flow (or cyclonic–anti-cyclonic-cyclonic pattern, see Ch. 3.2). The understanding is that the surface layer is wind-driven (it can't really be anything else...), while the bottom layer is probably from Pacific flow intrusion. The middle layer on the other hand is a bit mysterious because, dynamically speaking, such a vertical configuration might be expected to be *baroclinically unstable* (Ch. 6), which would erode the middle layer, but it seems to be persistent. It is an ongoing question on what controls the South China Sea circulation, how it interacts with the other ocean and/or coastal components, and what consequences this can have in a changing climate.

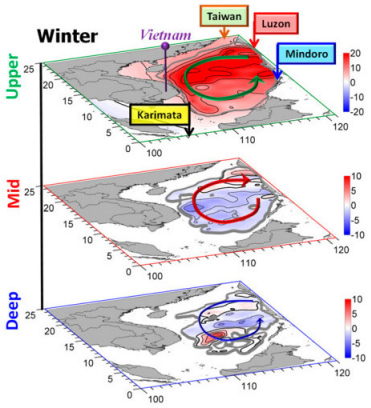


Figure 1.21: Schematic of the sandwich circulation in the South China Sea using numerical simulation data. Figure modified from Fig. 1 of Gan et al. [2016].

1.4 *Concepts in Newtonian mechanics, and governing equations for ocean dynamics*

Hopefully the above narrative has made a case for studying the physical dynamics present in the ocean. Ocean dynamics is motion on length-scales that **Newtonian mechanics**[22] works well. Here a brief overview is provided, introducing the concepts and form of the equations that is used for describing ocean dynamics. While we won't do very much with the actual equation itself, it is there for reference as we will refer to them particularly in Chap. 2 and 3.

[22] After English scientist Issac Newton (1642-1726?), widely recognised as one the most influential scientists of all time; his achievements are too many to fit into this margin (fake internet glory points for those who get this obscure reference). His formulation of mechanics was only recently superseded by Einstein's theory of relativity; we are not going to use relativity to formulate ocean dynamics, that's just overkill...

1.4.1 *Forces and Newton's laws*

Lets do the book keeping definitions first. **Newton's laws of motion** are given by:

1. a body at rest or in *steady* motion in a straight line remains at rest

or in steady motion unless there is a *net* force acting on it

2. rate of change of *momentum* of a body over time is directly proportional to the *net* force applied, and in the same direction as the *net* force

3. for every action there is an equal and opposite reaction

Note: we should be dealing with non-accelerating *inertial frames of reference* really, because if not we have to add in *fictitious forces*. We are actually going to live with the fictitious force in the form of the *Coriolis effect*, but that's for Ch. 3.2.

Lets put this into pictorially before going into a bit of maths. Intuitively, things move when **forces** act on a body. For no particular good reason we will take to the body of interest to be a pig[23].

[23] Spherical or point mass pig if you like. It's a pig because it's easy to draw.

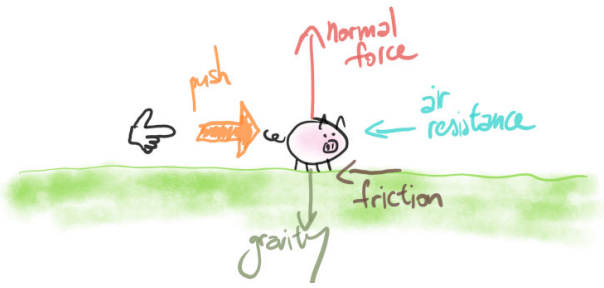


Figure 1.22: Schematic of forces acting on a pig. While it is customary in classical physics textbooks to assume we are in a vaccum, this would of course be against animal rights, so we do have air resistance if the pig is moving.

Assuming the pig is moving in steady state (i.e. moving but not accelerating, so in the regime of 1st law) to the right, there are various forces acting on this pig, namely:

- Earth's gravitational force pulling the pig down, balanced by a *normal force*[24] arising from the ground pushing back (3rd law), otherwise you might expect the pig to sink into the ground (1st + 2nd law)

- a phantom pointy finger pushing the pig to the right, balanced by *friction* from the ground and *air resistance* from the air resisting the motion of the pig (1st + 2nd law)

[24] Normal here refers to being normal or perpendicular (at right angles) to the surface.

Overall there are no net forces in the horizontal *and* the vertical, and the pig moves to the right along its merry way (or maybe not because it is being pushed by a phantom pointy finger). One thing to note in relation to the 3rd law is that, while from the point of view of the pig the ground and atmosphere is resisting its motion, an equivalent view to take is the ground and atmosphere's point of view, where the pig is pushing against them. There are some subtleties involving points of view here, which we will revisit in Ch. 3.1.1, 3.2 and 3.4.

We note first that the magnitude of friction and air resistance depends on the speed of the pig[25]. Now, instantaneously the phantom finger decides to increase/decrease the force magnitude accordingly (but still pointing to the right), friction and air resistance is still the

[25] The magnitude and direction depends on the *velocity*, but more on that later.

same (since the speed hasn't changed), so there is an imbalance of the forces, and so there will be a net change in *momentum*. Assuming the pig has mass m, (linear) **momentum** is defined as

$$\boldsymbol{p} = m\boldsymbol{u}, \tag{1.1}$$

where $\boldsymbol{u}$ is the **velocity**. Both $\boldsymbol{p}$ and $\boldsymbol{u}$ are *vectors*, i.e. something with a direction and a magnitude, to indicate where the pig is going. Since m is fixed in this case, change in momentum really means change in the velocity, so what we have is an acceleration/deceleration (2nd law). The acceleration $\boldsymbol{a}$ is given by the famous equation

$$\boldsymbol{F} = m\boldsymbol{a} = m\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}. \tag{1.2}$$

Note that acceleration and forces are also vectors (hence the boldface), because they all have a magnitude and a direction.

Since the velocity and speed is going to change, friction and air resistance will change, leading to further changes in the acceleration/deceleration, until the forces balance again, after which the pig will be moving on its steady merry way (1st law).

To close this section, note that, in SI units,

- mass usually has units kg while velocity has units m s^{-1}
- acceleration has units m s^{-2} (think of the $\mathrm{d}/\mathrm{d}t$ as bringing down a factor of s^{-1})
- the unit of force is called a *Newton*, and since $\boldsymbol{F} = m\boldsymbol{a}$, 1 N = 1 kg m s^{-2}

1.4.2 *Forces acting on the ocean*

Newtonian mechanics basically involves considering all the forces acting on a body or, in the ocean's case, some fluid, at some time t, and calculate the net force. The fluid will then move, being at some new place, and we repeat the process. Thus we have to consider what kind of forces are acting on the ocean and see what it does to the fluid.

One distinction that will be made here is **mechanical** and **thermodynamic** forcing, to be visited in detail in Ch. 2 and 3. The former is forcing that affects *momentum* and is the one described by Newtonian mechanics, and the latter affects the thermodynamic variables, which affects the density but in turn affects momentum. A sample of these are shown in Fig. 1.23. The main types of external forcing on the ocean we will be concerned with are:

- heating of the ocean by (incoming) *shortwave radiation* provided by the sun, and usually lumped in with this is (outgoing) *longwave*

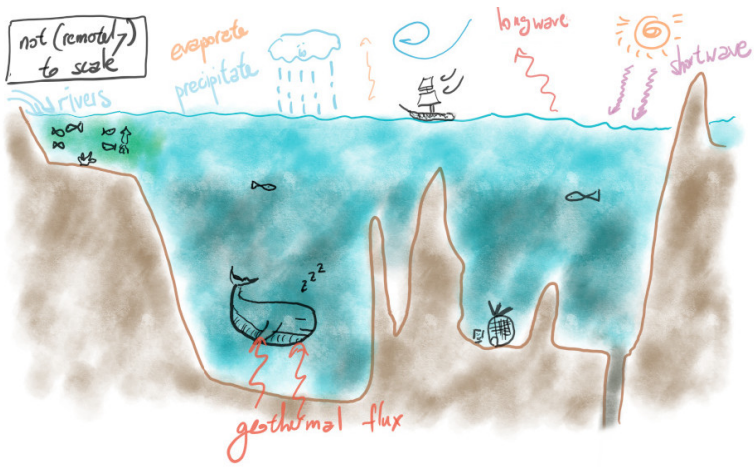


Figure 1.23: Figure based on Figure 2.2 of Pickard and Emery [1990], with some typical forcing drawn on.

radiation by the ocean leading to a cooling, affecting temperature and is thus a thermodynamic forcing (Ch. 2.2.1)

- *freshwater forcing*, either as river runoff or *precipitation* (rain, snow, hail, etc.), leading to a decrease in salinity, and *evaporation* of sea water (via heating from e.g. shortwave radiation), leading to an increase in salinity, and thus are thermodynamic forcings (Ch. 2.2.2)

- gravity as a mechanical force affecting the fluids momentum, principally arising from the Earth's attraction of the fluid leading to buoyancy forces, but noting buoyancy depends on the thermodynamic variables (Ch. 3.1.1)

- *wind* forcing involving the atmosphere, as a mechanical forcing, transferring *momentum* from the atmosphere into the ocean (Ch. 3.3)

Note I've deliberately omitted the Coriolis effect because is not a 'true' force (not that it doesn't mean it is unimportant), so be justified in Ch. 3.2. There are other types of forcing that are usually regarded as weak (e.g. *geothermal flux* from the solid Earth, leading to an increase in temperature of deep water, *stirring* by fish moving around in the ocean) and/or rare (e.g. *tectonic movements* with the solid inputting momentum into the ocean), which may be important in certain moments in space and/or time. We won't talk about those in detail here.

1.4.3 *Equations of motion*

The evolution of momentum and the thermodynamic variables could be described in words or, more concisely and in a less cumbersome way, by equations. Taking u, v, w as the **zonal** (east-west), **meridional** and vertical velocities respectively, we denot $\boldsymbol{u} = (u, v)$ as the horizontal velocity and $\boldsymbol{u}_3 = (u, v, w)$ as the full velocity. One form of the equation that describes ocean dynamics is given by[26]

$$\rho_0 \left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + 2\boldsymbol{\Omega} \times \boldsymbol{u} \right) = -\nabla p + \boldsymbol{F}_{\boldsymbol{u}} + \boldsymbol{D}_{\boldsymbol{u}}, \tag{1.3a}$$

$$\frac{\partial p}{\partial z} = -\rho g, \tag{1.3b}$$

$$\nabla \cdot \boldsymbol{u}_3 = 0, \tag{1.3c}$$

$$\left(\frac{\partial T}{\partial t} + \boldsymbol{u}_3 \cdot \nabla T \right) = F_T + D_T, \tag{1.3d}$$

$$\left(\frac{\partial S}{\partial t} + \boldsymbol{u}_3 \cdot \nabla S \right) = F_S + D_S, \tag{1.3e}$$

$$\rho = \rho(T, S, p), \tag{1.3f}$$

where F and D (in bold and normal format) denote the forcing and dissipations respectively, and the precise meaning of the symbols will be elaborated on in Ch. 2 to 3. The equations are, respectively:

(a) the *horizontal* momentum equation within a rotating frame of reference, to include the *Coriolis effect* $2\boldsymbol{\Omega} \times \boldsymbol{u}$ (see Ch. 3.2), with mechanical forcing and dissipation

(b) the *vertical* momentum equation, which actually reduces to *hydrostatic balance* as shown here from the approximations

(c) the *continuity* equation, corresponding to *mass conservation*

(d) an equation for *temperature* with thermodynamic forcing and dissipation

(e) an equation for *salinity* with thermodynamic forcing and dissipation

(f) the *equation of state* to get the density ρ (Ch. 2.3)

The thermodynamical and mechanical aspects are intrinsically linked. The fluid moves around because it is mechanically forced, as described by the momentum equation. The movement of the fluid transports temperature and salinity around. Additionally, the thermodynamic forcing affects temperature and salinity. However, changes in the temperature and salinity affect the density via the

[26] The *Boussinesq* and *hydrostatic* approximation has been used but we are not going to elaborate on those really. See Vallis [2006] for further reference.

Note Eq. (1.3a) is Eq. (1.2) in disguise as $m\boldsymbol{a} = \boldsymbol{F}$. I didn't put the Coriolis effect in with the forces, even though sometimes it is referred to as the Coriolis force. This is related to the margin note at the beginning of Ch. 1.4.1.

equation of state, and through hydrostatic balance affects the pressure, which in turn leads to pressure gradients driving the flow, and so on.

While this all sounds a very complicated, the goal here is to try and delineate this seemingly tangled ball of mess and make some sense of it. While we will focus on simplifications of the larger problem in order to gain some understanding, the thing one should always bear in mind is that everything is intrinsically linked. While the obvious (and to me somewhat cheap) comment/dig/criticism is that the simpler pictures do not represent the real world, in some ways that was never the intention and it is missing the point. The value is in what you learn from them, not from whether they describe the world down to the finest detail (great if they do of course).

If you are familiar the vector calculus and understand enough about what the equations are showing above, then you probably don't need the following section. For those who are not familiar, want to see a bit more detail, or want to see some propaganda, do continue onto the next section.

1.5 *A hand wavy introduction to vector calculus*

While there are basically no calculations involved in this document beyond working out orders of magnitude and the sign of something, vector calculus concepts will be used, simply because in my opinion it is the natural language to express physics related details in a concise format. The following is a very hand wavy tour of the relevant concepts in vector calculus, focusing on the geometric meaning of the symbols. Looking forward, we will use vectors and scalars generally, dot products to talk a bit about wave propagation (Ch. 6.1), cross product to talk about the Coriolis effect (Ch. 3.2.2), vector calculus generally, divergence when talking about Ekman pumping/suction (Ch. 3.3.3), and curl when talking about wind stress curl and vorticity (Ch. 3.3.3).

1.5.1 *Vectors and scalars*

As we have already seen in the discussion of Newton's laws, usually we are dealing with something that has a magnitude as well as a direction, i.e. **vectors**. **Scalars** on the other hand are just numbers. In this text vectors will be boldfaced[27], and scalars are always undecorated. Here are some examples:

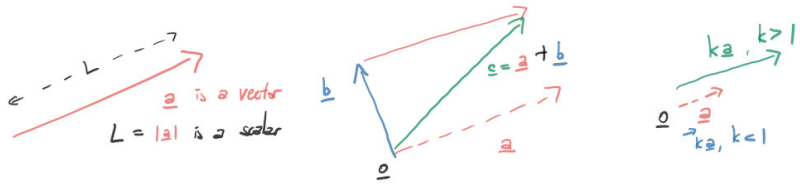
- gravitational acceleration $\boldsymbol{g}$ towards center of mass (and perpendicular to the *geoid*; see Ch. 3.1.1) is a vector, with magnitude $g = |\boldsymbol{g}|$, a scalar

The antagonist in the first *Despicable Me* film is called *Vector* (was Victor, the son of the banker), because he is "committing crimes with both direction and magnitude".

[27] Sometimes you see vectors with underlines, or arrows on top of them, however it is also not uncommon in theoretical physics or maths literature to see them not decorated at all.

- a fluid parcel travels with some velocity $\boldsymbol{u}$ (a vector), which has speed $|\boldsymbol{u}|$ (a scalar) in some direction
- pressure $p = p(x, y, z, t)$ is a *scalar field* but $\nabla p(x, y, z, t)$ is a *vector field*[28] (the gradient operator ∇ will be talked about soon)

While you can do the usual elementary operations with scalars, for vectors it is more limited:



[28] A scalar/vector *field* in this context is just a function that associated scalars/vectors to different points in space. A 1d scalar field (e.g. $y = f(x)$) draws a *curve*, a 2d scalar field draws a *surface*, and so on.

Figure 1.24: Schematic of elementary vector operations.

- add/subtract vectors to give another vector, $\boldsymbol{a} + \boldsymbol{b} = \boldsymbol{c}$, equivalent to joining vectors end to end
- multiply/divide vectors by scalars to get a vector, $k\boldsymbol{a}$, equivalent to stretching/squeezing a vector

!!! YOU DO NOT MULTIPLY/DIVIDE A VECTOR BY A VECTOR!!!

There are two other operations to do with vectors we will touch on, but first we note that we can represent a vector in terms of a *basis*[29]. The **standard basis** in 3d is simply

$$\boldsymbol{e}_x = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_y = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{e}_z = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

An example of how to use this is as follows for a 2d example is given in Fig. 1.25. We will only use the standard basis here.

The **modulus** of a vector is then $|a|$ of $\boldsymbol{a} = (a_1, a_2, a_3)$ is then given by $|\boldsymbol{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}$ (just Pythagoras' theorem). This is equivalent to taking the *length* of a vector, which takes a vector and returns a scalar, since length itself is a magnitude and has no direction.

The **dot product** or the **scalar product** between $\boldsymbol{a}$ and $\boldsymbol{b} = (b_1, b_2, b_3)$ is defined as

$$\boldsymbol{a} \cdot \boldsymbol{b} = a_1 b_1 + a_2 b_2 + a_3 b_3, \tag{1.4}$$

which takes two vectors and returns a scalar. We claim without proof that

$$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}| \cos\theta, \tag{1.5}$$

[29] Think of the basis as the lego bricks and any vector can be built by some choice of lego blocks. Just like lego blocks you can build the same vector using different kinds of blocks, i.e. the basis is not unique, but the standard basis is one very convenient choice that we will stick with. Here vectors are put in columns or rows and just chosen depending on whichever form is convenient for writing (we do not distinguish vectors and *co-vectors* or 1-*forms*).

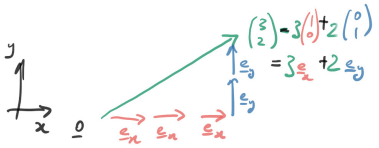


Figure 1.25: Example in 2d representing a vector in the standard basis.

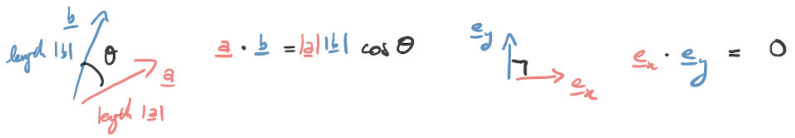


Figure 1.26: Schematic of dot product.

where θ is the angle it takes to take $\boldsymbol{a}$ to overlap $\boldsymbol{b}$, that anti-clockwise orientation as positive as per convention. Thus if $\boldsymbol{a}$ and $\boldsymbol{b}$ are **perpendicular**, then $\boldsymbol{a} \cdot \boldsymbol{b} = 0$, and vice-versa.

Finally, the **cross product** or the **vector product** between $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as

$$\boldsymbol{a} \times \boldsymbol{b} = \boldsymbol{c} = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix}, \tag{1.6}$$

so taking two vectors and returning a vector. By itself equation (1.6) might look a bit meaningless, but geometrically what this is doing is returning a vector $\boldsymbol{c}$ that is perpendicular to both $\boldsymbol{a}$ and $\boldsymbol{b}$, i.e. $\boldsymbol{a} \cdot \boldsymbol{c} = \boldsymbol{b} \cdot \boldsymbol{c} = 0$. This is done by the *right hand screw* convention as demonstrated in Fig. 1.27. In that example, you want to "twist" $\boldsymbol{e}_x$ into $\boldsymbol{e}_y$, which requires your right hand to be pointing *up* when you twist, i.e. in the direction of $\boldsymbol{e}_z$.

If on the other hand you want to find $\boldsymbol{e}_y$ into $\boldsymbol{e}_x$, then you follow the same logic and conclude that your right hand needs to be pointing down, i.e., in the direction of $-\boldsymbol{e}_z$. This observation is consistent with the *anti-symmetric* property of the cross product, i.e.

$$\boldsymbol{a} \times \boldsymbol{b} = -\boldsymbol{b} \times \boldsymbol{a}.$$

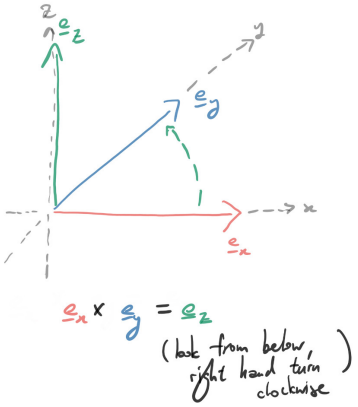


Figure 1.27: Schematic of the cross product, following the right hand screw convention.

1.5.2 *Calculus: derivatives and integrals*

For whatever reason calculus seems to freak a lot of people out (because there are so many rules and tricks to remember and it feels like mathematical gymnastics for the sake of doing it?) At the intuition level and for our purposes it is just **gradients** (i.e. *rate of change*) and **integrals** (i.e. *sums*). If we ever want to talk about how things *change* in, for example:

- how the temperature changes in some material as heat is applied,
- how the rate of reaction changes depending on the concentration of chemicals,
- how the value of Dogecoin (Fig. 1.28) has changed over time on the market,
- how the ocean/atmosphere moves around as forces are applied,

Figure 1.28: Such market, how trade, to the moon, wow.

then calculus is a tool (probably the tool) we probably want to use to describe the statements in a concise and quantitative way.

Lets start with **gradients**, which for **linear** (straight) things is defined as

$$m = \frac{\Delta y}{\Delta x},$$

i.e., how much do you go up/down (say) when you move along horizontally; see Fig. 1.29. If $m > 0$ then we have positive slope, and we go up as we move towards positive x. If $m < 0$ then we have negative slope and we go down as we move towards positive x.

That's all well and good, but generally functions are not linear (as in Fig. 1.29) then what? The beauty of calculus is that it is generally (!) mathematically sensible and well-defined *if we just treat the nonlinear curves as if it is linear by zooming in sufficiently*. This is illustrated in Fig. 1.30. Initially Δx is large and the approximation of the gradient at a point on the curve by a straight line is terrible. So we decrease Δx (i.e. the more we zoom in), and the approximation gets better and better. So why don't we take Δx going to zero? Calculus tells you this limit can be well-defined and controlled accordingly, and it is generally meaningful to talk about

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{y(x + \Delta x) - y(x)}{\Delta x} \to \frac{\mathrm{d}y}{\mathrm{d}x}.$$

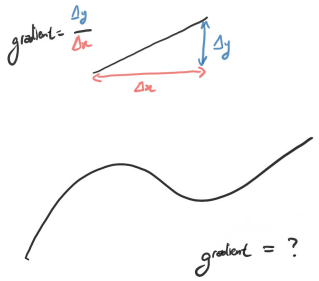


Figure 1.29: Gradient of a straight line is well-defined. But how to define gradient of a nonlinear curve?

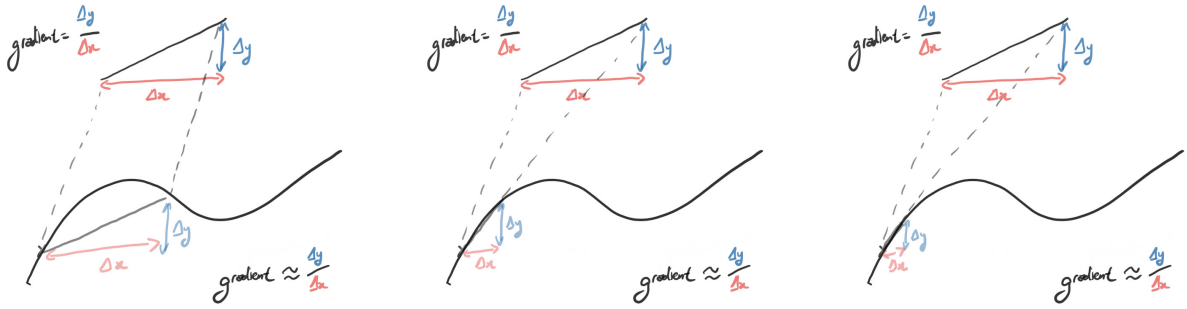


Figure 1.30: Idea behind the derivative, approximating the gradient by taking increasingly smaller increments in Δx such that the linear approximation works for sufficiently small Δx.

And that's it really (at least from a practical point of view)! Derivatives are basically gradients, nothing more. Some examples:

- assuming $p = p(z)$ only, $\mathrm{d}p/\mathrm{d}z$ would be the change of pressure with respect to depth
- assuming $[CO_2]$ is a function of temperature T only for whatever reason, then $\mathrm{d}[CO_2]/\mathrm{d}T$ would be the change of CO_2 concentration with water temperature T, which might be expected to be negative (why?)

The above assumes functions of one dimension, so no ambiguity in talking about the **total derivative** $\mathrm{d}/\mathrm{d}(\text{stuff})$. For a function with multiple dependencies we might talk about the **partial derivative** $\partial/\partial(\text{stuff})$. Instead of a curve and its gradient, the partial derivative can be thought of in 2d as the gradient of a surface in only one of the directions. Some examples:

- for $p = p(x, y, z)$, $\partial p/\partial z$ would be the change of pressure with respect to depth *regarding x and y as fixed*
- for $[CO_2]$ a function of temperature T and p only for whatever reason, then $\partial[CO_2]/\partial T$ would be the change of CO_2 concentration with water temperature T *regarding pressure p as fixed*

We are not going to do any actual calculations that involve taking derivatives as such, but see Appendix for a collection of rules and examples.

The **integral** $\int$ can be thought of as a sum (hence the stretchy 'S' symbol $\int$) and the opposite of a derivative[30]. Something like

$$\int_{-H}^{0} \rho(z)\ \mathrm{d}z, \qquad \int^{z} \rho(z')\ \mathrm{d}z',$$

means the function (or field) summed (integrated) in the z direction. The first one means sum the function ρ between $z = 0$ and $z = -H$ (a **definite integral**), and you end up with a function that is no longer a function of z (because you summed over it). The latter on the other hands means you are doing a cumulative sum up to some depth z, so the resulting object *is a function of z*; this is sometimes just denoted $\int \rho(z)\ \mathrm{d}z$, which is the **indefinite integral**, the **primitive** or the **anti-derivative**. Further examples:

- for x denoting the longitudinal direction, $L_x^{-1} \int_0^{L_x} f\ \mathrm{d}x$ would mean the **zonal average** of f, and is no longer a function of x
- for V the volume of a region, $V^{-1} \iiint f\ \mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z$ would be the domain-averaged temperature, which would be a function in t possibly

The idea behind integrals again is similar to the derivative, think of it as a generalised way of working out the area under a graph as in Fig. 1.31. If the function is linear this is easy: you end up calculating areas of a trapeziums (triangle plus a rectangle). If the function is not linear (but at least continuous), then you can play the same trick as before, and chop up the graph into increasingly small segments of width Δx, so then again the trapezium approximation gets better and better. Calculus then tells you the $\Delta x \to 0$ can be well-defined and

[30] This one has the grand name of the Fundamental Theorem of Calculus and due to Maria Gaetana Agnesi (1718-1799). Besides being a child prodigy, she was according to the Britannica "*considered to be the first woman in the Western world to have achieved a reputation in mathematics*". She wrote the first textbook discussing both differential and integral calculus and the connection.

There is a dummy variable z' for clarity reasons. Formally when you do $\int \rho\ \mathrm{d}z$ you would integrate out the z-dependence, i.e. the object is no longer a function of z, but then $\int^z \rho(z)\ \mathrm{d}z$ should be thought of as a function of z. To avoid this (hypothetical?) confusion, a dummy variable is used: the resulting object is certainly no longer a function of z'.

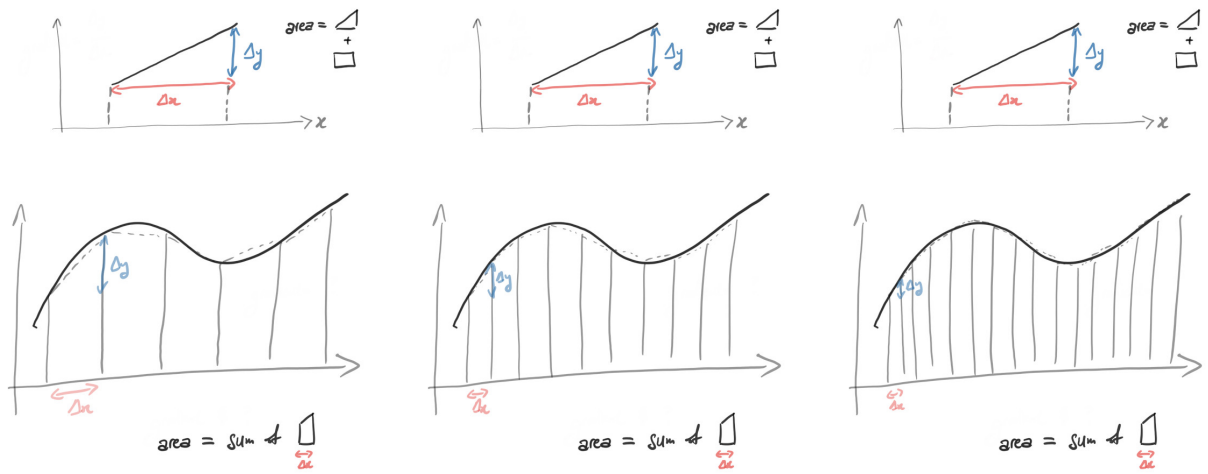


Figure 1.31: Idea behind the integral, approximating the sums by taking increasingly smaller increments in Δx such that the linear approximation works for sufficiently small Δx. The "chopping" is called a *partition*, and in this case we chopped in "x" (the *Riemann integral*). You will also notice you could chop it horizontally, i.e. in "y" = $f(x)$, which turns out to be a more robust definition but needs more complicated machinery; see the *Lebesgue integral*.

actually works, and there are associated rules for doing integrals. Again, since we are not going to do any actual calculations that involve analytically computing an integral, see the Appendix for a collection of rules and examples.

1.5.3 *Vector calculus: grad, div and curl*

The vector calculus of interest here will mostly involve three operators associated with derivatives, and are used quite a bit through the document. There are some integral related ones that do make a showing in geophysical dynamics relatively often (e.g. Green's theorem, divergence theorem and Stokes' theorem) but we won't use them here; see the Appendix for those.

The **gradient** (sometimes just **grad**) operator ∇ ('nabla') takes a scalar field and returns a vector field as

$$\nabla p(x,y,z) = \begin{pmatrix} \partial p/\partial x \\ \partial p/\partial y \\ \partial p/\partial z \end{pmatrix} = \frac{\partial p}{\partial x}\boldsymbol{e}_x + \frac{\partial p}{\partial y}\boldsymbol{e}_y + \frac{\partial p}{\partial z}\boldsymbol{e}_z. \tag{1.7}$$

Again these are just gradients and nothing more (except now it has a direction because it is a vector). The two that we will mostly encounter are $-\nabla p$, the *negative pressure gradient* (i.e. high pressure regions "pushing" stuff into low pressure regions), and $\nabla\rho$, which is the *density gradient*[31].

The **divergence** (sometimes **div**) operator $\nabla\cdot$ takes a vector field and returns a scalar field[32] as

[31] Density is related to $\partial p/\partial z$ via *hydrostatic pressure*, and contribute to *thermal wind* (Ch. 5.1.3)

[32] Recall dot product above.

$$\nabla \cdot \boldsymbol{u}(x,y,z) = \nabla \cdot \begin{pmatrix} u(x,y,z) \\ v(x,y,z) \\ w(x,y,z) \end{pmatrix} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}. \tag{1.8}$$

Geometrically what div represents is *convergence* or *divergence* of a vector field. Fig. 1.32 shows a 2d schematic, with the dashed lines showing a before area and solid lines showing an after area. For converging flow, things are piling in and the area shrinks, with $\nabla \cdot \boldsymbol{u} < 0$. For diverging flow, things are moving out and area expands, with $\nabla \cdot \boldsymbol{u} > 0$. For pure rotations or uniform translations, the area simply gets moved around with no change in volume, $\nabla \cdot \boldsymbol{u} = 0$. This is used later particularly in relation to *Ekman pumping/suction* (Ch. 3.3.3); since we have mass should be conserved, if there divergence at the surface, there has to an associated *upwelling* to replenish the mass that are moving out, and vice versa for convergence.

The last one of interest here is the **curl** operator, which returns a vector field from a vector field as[33]

$$\nabla \times \boldsymbol{u}(x,y,z) = \begin{pmatrix} \partial w/\partial y - \partial v/\partial z \\ \partial u/\partial z - \partial w/\partial x \\ \partial v/\partial y - \partial u/\partial x \end{pmatrix}. \tag{1.9}$$

Note that for almost the entire document we will be focusing on the z-component of the curl ζ ('zeta'), i.e.

$$\zeta = \boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{u}) = \frac{\partial v}{\partial y} - \frac{\partial u}{\partial x}. \tag{1.10}$$

These will show up as either the (relative) *vorticity* (the curl of the velocity) or the *wind stress curl* (Ch. 3.3.3).

One way to think of the curl is how much the vector field spins around (how the field "curls" around I suppose); see Fig. 1.33 for a schematic. A uniformly converging, diverging and translating flow has no curl and is *irrotational* (i.e. $\zeta = 0$ here), as the fluid area does not change its orientation. For a rotation and a shear however it does. By convention anti-clockwise rotation is *positive* curl (because mathematicians measure angles in an anti-clockwise manner), so the examples shown in the figure has negative curl because the fluid area rotates clockwise.

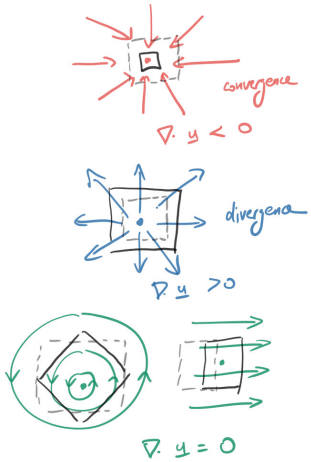


Figure 1.32: 2d schematic for the divergence of a vector field. Dashed lines denote the before area while the solid lines denote the after area.

[33] Recall cross product above, and compare the formula with Eq. (1.6).

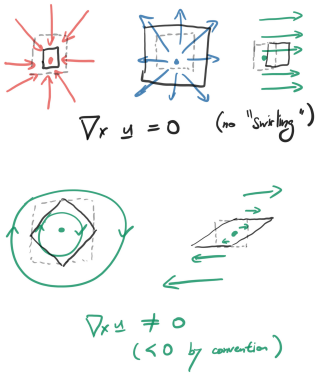


Figure 1.33: 2d schematic for the curl of a vector field. Dashed lines denote the before area are while the solid lines denote the after area.

1.6 *Conventions used here*

- x, y, z you can/should think of as east-west, north-south and up-down, respectively called the *zonal*, *meridional* and *vertical* direction
- $x, y, z > 0$ are East, North and up, so $\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z$ points east, north and up

- $\boldsymbol{u}_3 = (u, v, w)$ is a vector field, and u, v, w are scalar fields and the zonal, meridional and vertical velocity, with $u, v, w > 0$ the east, north and upward velocity
- vectors will be **bold** whilst scalars are un-decorated
- anti-clockwise is positive angles and curl regardless of hemisphere

Summary and further reading

The focus of this book is going to be primarily on *dynamics*, so the reader is referred to the books such as Talley et al. [2011] or Pickard and Emery [1990] for a broader overview of the ocean. For more rigourous treatments of the dynamics, please refer for example to Vallis [2006], Williams and Follows [2011] or Wunsch [2015]. There are many other books out there, but these are the ones I have personally used the most.

Chapter exercises

1. Provide an estimate of the aspect ratio for the ocean. What about for the atmosphere (take the stratosphere as the vertical limit for the atmosphere if you like)?

2. I have left out many currents and seas around the world. Find maybe two or three of these yourselves and see what features they have (e.g. geographical location, depth, extent, circulation properties etc.).

3. The *Leeuwin current* is a current on the west coast of Australia. Look up maybe the Wikipedia page or otherwise, find a few things on how this current is similar and/or different to conventional Western Boundary Currents.

4. Look up the dimensions and some current properties associated with the *Bering strait*, and from that give an estimate of the water transport. From this, comment/speculate on how 'connected' the Arctic and Pacific actually is.

5. In prehistoric times it is known that the Drake passage and the Tasman gateway (between Australia and Antarctica) might not have been completely open. What kind of effects might this have? Speculate a few of these, but try and justify your answers (Harder question: can you still get an ACC if e.g. the Tasman gateway was partially closed and there are no open latitudes? Justify your answer.)

6. Show that for $\boldsymbol{c} = \boldsymbol{a} \times \boldsymbol{b}$, $\boldsymbol{c}$ really is perpendicular to both $\boldsymbol{a}$ and $\boldsymbol{b}$ by explicitly computing $\boldsymbol{a} \cdot \boldsymbol{c}$ and $\boldsymbol{b} \cdot \boldsymbol{c}$ using equation (1.6).

7. Fig. 1.34 shows a typical road sign for steep slopes. The sign means you go down 1 units for every 8 units you move horizontally (roughly speaking). With this, compute the magnitude of the gradient, and work out the slope angle relative to the horizontal (give it in degrees). For those mathematically inclined, estimate the angle but don't use a calculator (gradient of 1/10 might be neater to do).

8. If a continental slope has height 3000 m and extends horizontally over 30 km, find the magnitude of gradient.

9. Repeat the above by but give the magnitude and sign of gradient if we are measuring from the coast towards the ocean, and we have a continental slope configuration as in Fig. 1.13.

10. Look up the value of Dogecoin and describe its 2021 price changes in terms of gradients.

Figure 1.34: Image from HK transport department (www.td.gov.hk).

11. Give a non-zero vector field that has zero curl and div (drawing or mathematical representation is fine).

12. Is it possible to construct a case where the horizontal winds on Earth are everywhere non-zero? (In vector calculus speak, is it possible to have a 2d vector field on the surface of the sphere that is everywhere non-zero at any instance?)[34] If you can, draw it. If not, convince yourself why not, and see what is the minimum number of zero points you must have on this surface.

[34] Look up the *hairy ball* theorem if you want a hint.

2 *Seawater properties and thermodynamic forcing*

Key takeaway of this chapter: FROM A DYNAMICAL POINT OF VIEW, IT'S ALMOST NEVER IN-SITU DENSITY YOU CARE ABOUT!

We go on to justify the above claim. In Ch. 1 we made a distinction between the *thermodynamic* and *mechnical* contribution to the fluid dynamics, the former to mean anything that directly affects the *density* of the fluid, and the later to mean anything that drectly affects the *momentum* of the fluid. Of course the two are intimately linked, but we will start first with a discussion of the *thermodynamic* aspects because it is perhaps more intuitive to talk about[1]. Referring to the equations given in Eq. (1.3), the equations of interest are

$$\left(\frac{\partial T}{\partial t} + \boldsymbol{u}_3 \cdot \nabla T\right) = F_T + D_T, \tag{2.1}$$

$$\left(\frac{\partial S}{\partial t} + \boldsymbol{u}_3 \cdot \nabla S\right) = F_S + D_S, \tag{2.2}$$

$$\rho = \rho(T, S, p). \tag{2.3}$$

Here, the $\boldsymbol{u}_3 \cdot \nabla T$ term represents the *advection* term, i.e. how water moving around carries in this case temperature T around, and we have the forcing as F, while D represents the dissipation and *diffusion* terms (see Ch. 3.4 for more detail about diffusion).

The key quantity that we are aiming to get to is the fluid **density**[2] ρ (with units of kg m^{-3}) or the **buoyancy** $b = -(\delta\rho/\rho_0)g$, where ρ_0 is a reference density and g is the gravitational acceleration (in units of m s^{-2}; see Ch. 3.1.1). The density measures how much 'stuff' there is per volume, and if two blobs of water have identical volume but one is more dense, then the denser one is heavier. An equivalent measure would be by the buoyancy: the heavier blob is less buoyant so has smaller buoyancy. You actually know some of this from intuition already, for example via the thought experiment in Fig. 2.1. On the one hand we have warm water over cold water, which we know will be stable, but on the other hand if we have cold water over warm

[1] I hesitate to use the adjective 'easy' because it can get a little mind bending and slightly confusing.

[2] For now I am referring to *in-situ density*, but we will get to the subtleties later.

water we expect the water to overturn in the vertical, because cold water is heavier, denser or less buoyant than the warm water. While temperature is used here, for seawater salinity is also important, though ultimately it is actually the density that matters.

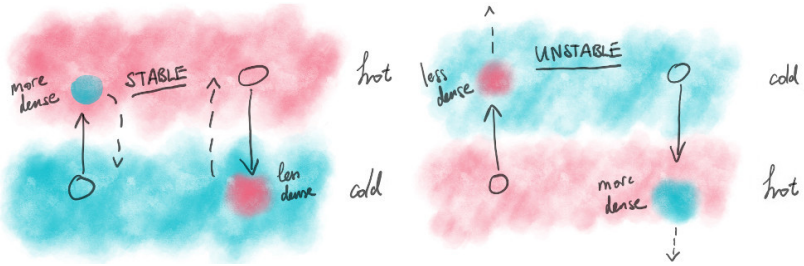


Figure 2.1: Parcel argument for static instability, red and blue denoting warm and cool water respectively. More in Ch. 6.

As can be seen from the thought experiment in Fig. 2.1, density is expected to play an important role in the vertical movement of seawater. It turns out that minor differences in density are also one of the primary drivers for horizontal dynamics in the ocean (through a combination of *hydrostatic* and *geostrophic balance*), which we will visit again in Ch. 3 and 5.

2.1 *Seawater properties*

Since density is argued to be important, we want to know what contributes to density of seawater. Before we do that, just a quick clarification on two physical concepts. **Mass** (usually units of kg) is about how much 'stuff' the body is made of, and is a scalar. **Weight** is force and so a vector, and is the mass multiplied by the magnitude of gravitational acceleration in the direction of the gravitational acceleration, i.e. $\boldsymbol{F} = m\boldsymbol{g}$. Because we are talking about blobs of water, instead of mass we use density, and it is the changes in the density that leads to a difference in the force, and thus an acceleration by Newton's laws.

Temperature T of the water is a measure of how warm or cool something is, and is often measured in degrees Celcius (° C) of Kelvin (K, but no degrees)[3]. One can think of temperature of a blob to be related to the amount of thermal *energy* in the blob, so sometimes we talk about temperature and energy content interchangeably, related to the **heat capacity** (in units of J K^{-1}), which is the energy required to raise the temperature by a certain amount. For practical purposes, we take the heat capacity of seawater to be a constant, and energy Q (in units of Joules, J), mass m, the **specific heat capacity** c (in units of J kg^{-1} K^{-1}, note the per mass bit) and the change in

[3] The 'spacing' of degrees Celcius and Kelvin are exactly the same, except there is a shift so that 0° C = 273.15 K. 0 K is known as *absolute zero*.

temperature of a material ΔT is given by

$$Q = mc\Delta T, \tag{2.4}$$

where for seawater[4], $c \approx 3850$ J kg^{-1} K^{-1}. This formula is used when we are talking about *heat content* in the ocean, which is roughly the thermal energy in the ocean (cf. Fig. 1.3). To measure temperature, the obvious thing is to use a *thermometer*. Even though a thermometer would provide a gold standard of sorts, we generally can't do that easily if we want a large spatio-temporal coverage of the ocean. There have been suggestions to use for example the *speed of sound* in seawater as a way to get at the temperature: speed of sound depends on the density of water, and differences in density lead to wave *refractions* (Ch. 6.1) therefore measurable changes in the travel time, from which the density and temperature could be inferred for.

[4] The specific heat capacity for air is much smaller, hence why it takes so much longer to heat up the kettle for that cup of tea than warming the surrounding air, but also the cup of tea keeps its heat much longer.

With regards to density, we expect that increasing T decreases ρ in pure water. However the rate of increase is not necessarily linear above 4° C, and also the density of water actually *decreases* when it is cooled below around 4° C where water is densest[5], and this essentially points to a *nonlinear* dependence of density in water as well as seawater to temperature; see top panel of Fig. 2.2 and note the turning point (the peak of the graph). We say a bit more about this in Ch. 2.3.

[5] Hence why ice floats.

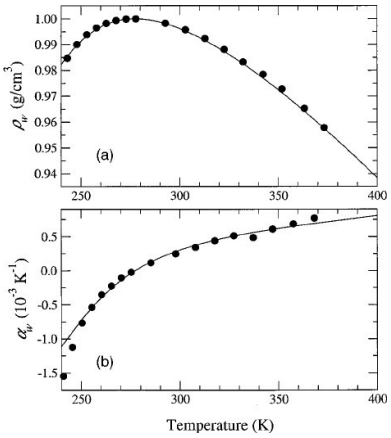


Figure 2.2: (top) density $\rho = \rho(T)$ for pure water, (bot) the coefficient of thermal expansion $\alpha = \alpha(T)$. From Ashbaugh et al. [2002].

As already mentioned, one distinguishing feature of seawater is that it is salty[6]. The saltiness derives from the fact there is sodium chloride (NaCl) dissolved in seawater, and salt contributes to the mass and density of the fluid. We denote the **salinity** of seawater by S, which measures how salty or fresh something. The measure of salinity is a concentration and sometimes given by g kg^{-1} (grams of NaCl dissolve in 1 kg of seawater), and note that this is non-dimensional since it is a mass divided by another mass. However you sometimes see salinity given in 'units' of PSU (*practical salinity unit*), because some people think it is weird to not have units tagged on with a measure; the use of PSU is strongly discouraged nowadays [IOC et al., 2010].

[6] 'Salt' in chemistry means something slightly more general.

The chemical measure of salinity is by measuring the *chlorinity*, i.e. the concentration of chlorine atoms in the water sample. This requires analysing the water sample by, e.g., *titration* against silver nitrate solution to precipitate the chlorine. The resulting processed salinity is known as **absolute salinity** S_A, defined as

$$S_A = 1.80655 \times \text{Chlorinity}. \tag{2.5}$$

As you can probably tell, this is impractical to do on a regular basis. An alternative is to note that seawater conducts electricity, and so the **conductivity** depends on the concentration of salt, so we could

potentially conductivity to obtain salinity[7]. Conductivity is relatively speaking much easier to measure, although it is less accurate as strictly speaking the conductivity depends on temperature and pressure as well. Salinity measured this way is usually called **practical salinity** S_p (or just S).

For reference, freshwater tends to have low salinity, on the order of 0.01 g kg^{-1}. Seawater tends to have salinity around 35 g kg^{-1}, and doesn't vary much (from 33 to 37 or so), though the Med Sea can have salinity up to 40 g kg^{-1}. Estuary regions are influenced by rivers as well as oceans so salinities there can vary from 0.5 up to 35 g kg^{-1}. For completeness, the Dead Sea has salinities of around 200 g kg^{-1}, and that's why they suggest you don't spend too long in there, and definite wash off afterwards, because the saline water dehydrates you quite substantially[8].

There are other things dissolved in water (e.g. chemical *tracers* such as oxygen and dissolved inorganic carbon), but these don't contribute much to the physics so we won't really talk about it (the physics on the other hand is very important for the content). Before we go on, the last major property we will talk about here is **opacity**. Something is *opaque* if light (more accurately, *radiation*; more in a bit) does not travel through the material very well[9]. Air for example is not opaque, because otherwise you would not even be able to see this set of notes. Clouds are opaque because it absorbs/disperses light. Seawater opacity depends on the water condition (e.g. the *turbidity*) but usually it would be regarded as opaque: you usually can't see that deep into the ocean (e.g., Fig. 2.3). This is mostly the reason why you only find *phytoplanktons* above around 200 m depth in the *euphotic zone*, because they need sunlight for *photosynthesis*, and there is not enough light below a certain point.

[7] Pure water is a very bad electrical conductor because not that many H_2O molecules disassociate into H^+ and OH^-, and free charges are needed to conduct an electric current. In seawater, the dissolved NaCl has their ionic bonds broken by water (because water has a net dipole), resulting in Na^+ and Cl^- charged particles, and therefore a possibility to conduct an electric current.

[8] But is good for cleansing purposes I guess, and it means the muds there are full of minerals, hence it is highly prized from a cosmetics point of view.

[9] Describing things (e.g. a book or someone's speech) say as opaque would then be that something is hard to understand, confusing, or impenetrable, possibly like this set of notes.

Figure 2.3: Picture of the sea. You can't see through it that well so seawater is regarded as opaque. CCo Public Domain, taken from `phys.org`.

2.2 *Observations and forcing*

The fact that seawater is opaque has important consequences for the ocean. Naturally, a main source of energy on Earth is from the Sun, which is responsible for heating the Earth system and, in turn, driving the winds in the atmosphere that forces the ocean mechanically. However, given sunlight (more precisely, *solar radiation*) doesn't penetrate much into the deeper parts of the ocean we have the set up in Fig. 2.4.

In the atmosphere, because the air allows solar radiation through, the atmosphere may be heated below. Much like boiling a pan of water from below, the fluid at the bottom becomes warmer, less dense than the water above, and overturns, leading to *convection*. Clouds exist because there is convection leading to upward motion of air

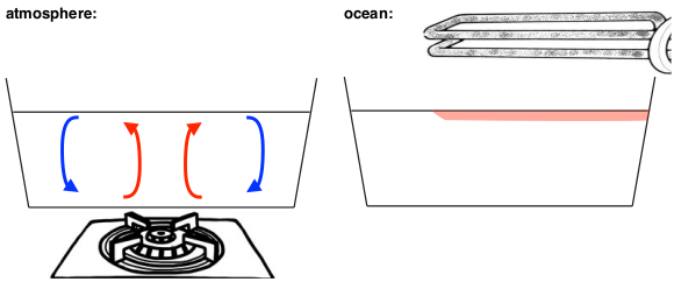


Figure 2.4: Schematic of principal sources of thermal forcing between atmosphere and ocean. Courtesy of David Marshall (Oxford).

carrying moisture up high, so that when the air cools condensation occurs and forms clouds (mostly happening in the *troposphere*). In the ocean, however, since water is opaque, solar radiation gets dispersed easily and the scenario is really like a case where you trying to heat up a pan of water, but from above, which is you may guess is almost impossible[10]! Without some sort of fluid transport the heat basically takes ages to *diffuse* into the interior of the ocean (we are talking about at least thousand year time-scales easily; see Ch. 3.4), and there cannot be significant motion in this kind of set up[11]! Of course despite this inefficient setup, there is in fact a large-scale overturning circulation which can move heat around in the ocean, and we explore some of the reasons for why this is in Ch. 5.

[10] I once tried this when visiting my now wife to boil some dumplings back when she was in her student halls. After 30 mins we gave up and got take-away in the end.

[11] This is related to something now called *Sandstroöm's theorem*, formalised in Paparella and Young [2002] and Nycander [2010] for example.

2.2.1 *Temperature*

To talk about seawater temperature distribution and so on we first make a digression to talk about some physical aspects of **solar radiation**. Energy can generally be transferred in three ways: **conduction**, where a particle hits another particle physically; **convection**, where energy is moved around by a collection of particles; and **radiation**, where energy is transferred by **electromagnetic waves**[12]. The classic examples demonstrating the three is a heater. You could get warmed by the radiator through touching it (conduction, but maybe don't actually do this), by being above the heater and getting the warm air (convection, as heat is transferred through movement of air), or just by sitting slightly away from it (radiation, as **infra-red radiation**). The first two need matter to be present, while radiation does not (hence radiation can travel through space).

[12] These are *photons*, and visible light is a special case. Photons are a bit weird because they are wave-like and particle-like at the same time. See *wave-particle duality* in most *quantum mechanics* books.

The Sun emits all sorts of radiation and these are normally classified depending on their *frequency*/*wavelength*[13]. The types of electromagnetic waves are given in Fig. 2.5. Note that there are many familiar names here, from long radio waves that we use in communication, to microwaves for cooking, infra-red as heat, visible light between 380 to 700 nm (nano-meters, 10^{-9} m), to short and high frequency waves

[13] How often the waves oscillate and how wide two wave crests are apart, which are of course inversely correlated with each other. Measured in Hertz $\text{Hz} = \text{s}^{-1}$ and m respectively. More in Ch. 6.1.

such as ultra-violet (UV) and X-rays.

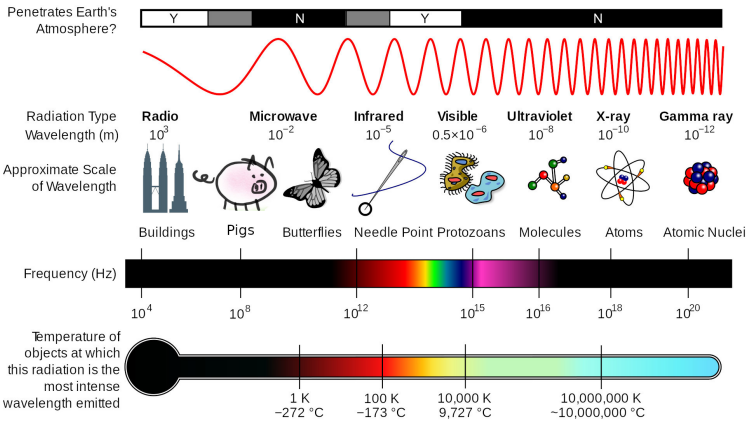


Figure 2.5: The electromagnetic spectrum by wavelength and frequency. Image from Wikipedia, adapted from an image form NASA.

This bit is going to be a bit hand-wavy, but the key take-away is that incoming solar radiation is largely in the form of **shortwave radiation**, energy loss by the ocean is via outgoing **longwave radiation**. If you are eager to get back to explicit talk about the ocean then you can skip the paragraph without losing that much I suppose.

Now, one can imagine that it takes more energy to make the wave oscillate more, so there is more energy in the high frequency waves[14], i.e. the short waves. A body in thermal equilibrium that is hotter emits more radiation, and is able to emit higher frequency radiation; if a cooler body emitted the same type of radiation it would cool down too fast, and thus not be in equilibrium. In fact, with an approximation[15], the kind of radiation and magnitude of radiation a body in thermal equilibrium can emit is uniquely determined by the temperature of the body[16]. For the current day Sun, with a surface temperature of around 6000 K, a good chunk of energy is radiation and thus energy is emitted at the visible and the lower end of the UV spectrum. However, because of the composition of the atmosphere some of the radiation gets reflected or absorbed and then emitted back into space[17], and the main component of solar radiation at the surface of the Earth is *shortwave radiation*. By a similar argument, since the Earth has a temperature, it still emits radiation, but because it is much cooler than the Sun, it emits radiation at a much lower frequency, thus mostly in the form of *longwave radiation*, such as infrared[18].

With that digression, Fig. 2.6 shows a typical profile of **SST** (sea surface temperature) and the incoming shortwave radiation Q_{sr} (in units of W m^{-2}), showing the horizontal distribution overlaid on a map. We also show the *zonally averaged* profile over latitude, defined

[14] This is the *Planck relation*.

[15] The *black body approximation*

[16] This is *Planck's law*, after the German physicist Max Planck (1858-1947).

[17] Look up *absorption spectrum*, and it is mainly ozone (O_3) absorbing the high frequency UV bands, and water vapour absorbing a lot of the remaining incoming solar radiation.

[18] *Greenhouse gases* are then the gases that permit shortwave radiation to pass into Earth, but traps long-wave radiation by absorbing it and re-emitting it back to Earth. 'Fun' trivia: while carbon dioxide CO_2 is talked about so much, the biggest contributor to greenhouse effect on Earth is actually water vapour H_2O. CO_2 is not even that potent a greenhouse gas, but it is a problem because there is so much of it...

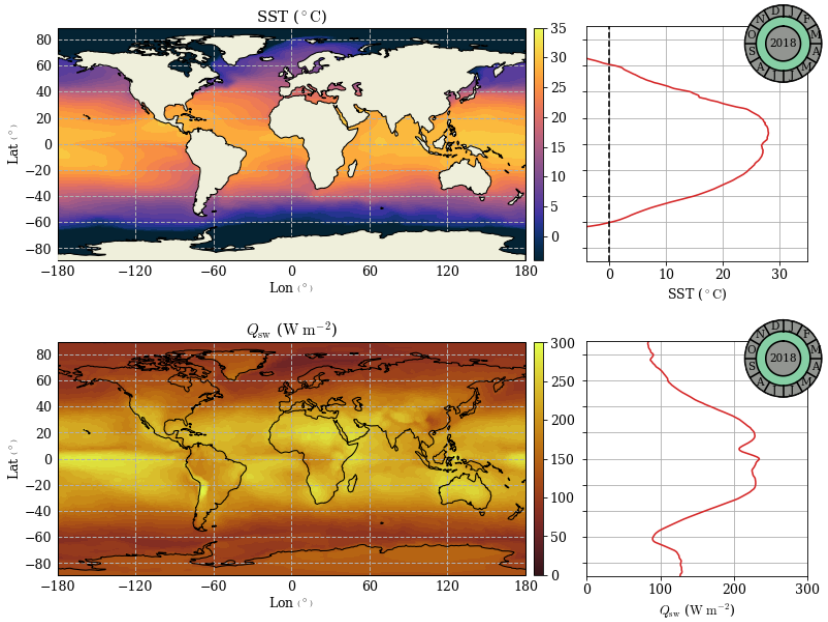


Figure 2.6: Year-averaged (left) and year and zonally averaged (right) sea surface temperature (top) and incoming shortwave radiation (bottom), from the JRA55 dataset (Kobayashi et al. [2015]). See `plot_jra55_sample.ipynb` for code, and `sst_day_avg_2018.mp4` and `rsds_day_avg_2018.mp4` for movies of the analogous daily averaged data through a particular year.

as

$$\overline{f}(y) = \frac{1}{L_x} \int_{0^\circ}^{360^\circ} f(x, y)\, \mathrm{d}x, \tag{2.6}$$

where (x, y) denotes the longitude and latitude, L_x is the circumpolar length, and x here is given in *degrees*[19], i.e. we average stuff in longitude. The data is additionally averaged over a year to show the typical profile. First, note that the SST is largest at the equators at around 30° C and smallest at the high latitudes, dipping below 0°C, which is what you would expect since the Equator receives the most heat[20]. Indeed, the Q_{sr} averaged over the year is also largest around the equator and smallest at the poles. In the supplementary movie files `sst_day_avg_2018.mp4` and `rsds_day_avg_2018.mp4`, which shows the daily-averaged data over the year, you can see the **seasonal cycle**, where the peak Q_{sr} starts below the equator, since January is **Austral summer**, i.e. where Southern Hemisphere receives more heating), moves north as the months progress, and is in the Northern Hemisphere by the time we get to **Boreal summer**, i.e. Northern Hemisphere summer, before venturing south again as the months progress. The SST signal follows this behaviour somewhat but the variations are much smaller compared to Q_{sr}, and this is again because of seawater has a higher heat capacity (because of the amount of energy that needs to be gained or lost to change the water temperature). The outgoing long wave radiation Q_{lr} is not shown, but is correlated well with the SST evolution (why?)

What about changes of temperature with depth? Fig. 2.7 shows

[19] We should really be using *radians* (180° = 1 rad) for calculations, but we don't do calculations here so stick with degrees.

[20] Because on average the Equator is closest to the Sun, even if the Earth has a tilt leading to seasonal changes in the forcing.

the changes of temperature[21] at fixed longitude and latitude as a function of depth, and red and blue lines denote the (Boreal) summer and winter profiles respectively. There are several features of note:

- the temperature profile tends to be 'vertical' near the surface particularly in the winter, i.e. the temperature gradient with depth is close to zero, since temperature is barely changing
- in general, there is a region where the where the temperature gradient is large, in a layer between the surface and the deeper parts
- in general, as we get deeper, the gradient decreases again

[21] A cheat and a note here: this is actually *conservative* temperature; more on this later.

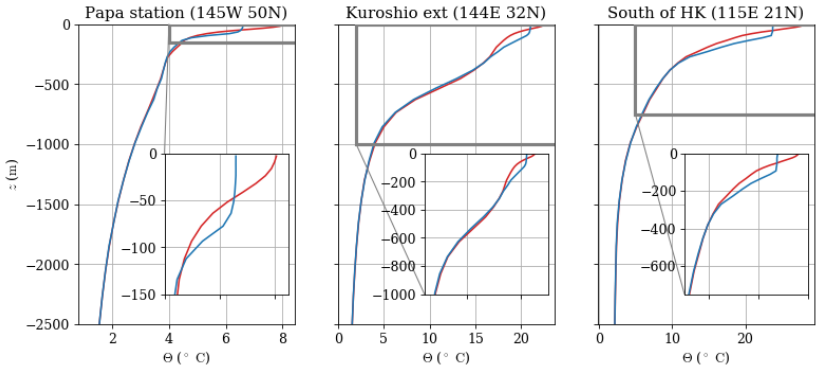


Figure 2.7: Vertical variation of conservative temperature at some designated locations, based on WOA13 data. Red and blue line denote summer and winter climatology. See `plot_WOA13_sample.ipynb`

The exact depths and extent of where the noted features occur at depends on the geographical location[22]. The first observation is of the **mixed layer**, with the winter signal being particularly easy to see. The definitions and measures of the mixed layer varies, but is generally characterised by where the stratification is weak, and is usually located within the top 100 m or so (except in high latitudes during the respective winters). The weak gradients are indicative of strong convective activity (see Ch. 6.2), mixing everything up and eroding the *stratification*[23]. The second roughly denotes the **thermocline**, which is the region where the temperature gradient is the largest in the vertical, i.e. the transition is the fastest, and is located roughly between 200 to 1000 m or so. Below the thermocline is usually regarded as the deeper parts of the ocean, where the temperature gradients are relatively small.

To anticipate the discussion about *watermass properties* when we want to talk about the MOC in Ch. 5.2, Fig. 2.8 shows a **meridional section**[24] of yearly-averaged temperature, in both the Atlantic and Pacific. Temperature is largest at the surface and near the equator, as

[22] The middle graph shows the Kuroshio, where there is a large gradient in temperature that goes down to around a 1000 m, and this signal is consistent with *thermal wind shear relation* and that the Kuroshio is a WBC (Ch. 5.1.3.)

[23] *Stratification* refers to 'layers'. Although you can think of these as gradients, we usually use stratification when talking about density.

[24] Fix a longitude and show data in latitude and depth. *Longitudinal sections* are analogously defined.

expected. There are several features of note, which we will revisit in more detail later:

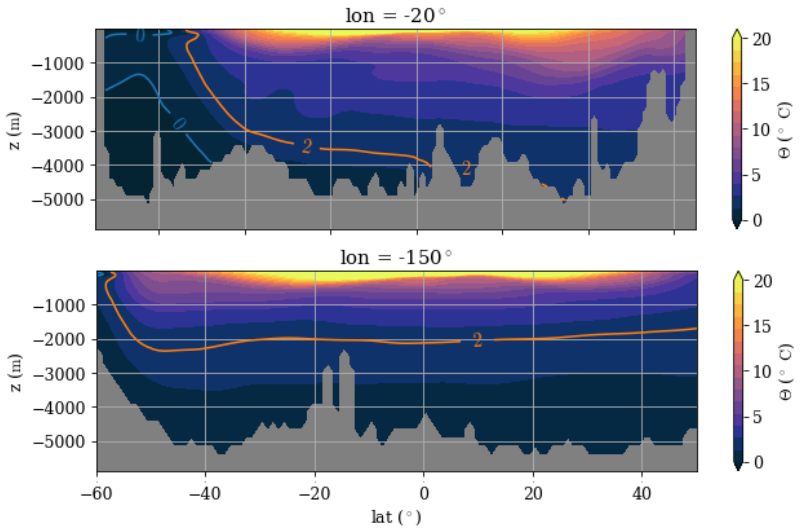


Figure 2.8: Meridional section of yearly-averaged conservative temperature in the Atlantic (top) and Pacific (bot), based on World Ocean Atlas 2013 data. Meridional range chosen to roughly correspond to Talley et al. [2011] Fig. 4.11 and 4.12. See `plot_WOA13_sample.ipynb`

- in the Atlantic, there is a hint that the warmer water seems to intrude deeper in the North
- in the Atlantic, the cold water particularly marked by the 2° C and particularly 0° C **isotherm** (lines of constant temperature) at the deep seems to intrude from the South into the basin
- in the Pacific, the aforementioned intrusion of deep cold water seems to be absent (no 0° C isotherm), and the 2° C is much higher in the water column

These are to do with the global MOC and are related to the NADW and AABW watermasses, as well as the pattern of the global MOC; see more in Ch. 5.2. This rather brief discussion is mostly to demonstrate that water formed from various locations could have a signature in observations (Ch. 7), providing a way to identify their origin and how they might change as time goes on.

2.2.2 *Salinity*

Salinity forcing is a bit simpler to talk about, although a caveat here is the total mass of salts in the ocean are effectively conserved, so *salinity forcing* really refers an increase or decrease of salt concentration via *changes in the amount of freshwater content*. Other than that subtlety, incoming shortwave radiation Q_{sr} heats up seawater and causes **evaporation**, but because water vapour itself cannot

carry salt[25], evaporation leads to an *increase* in salinity. By contrast, **precipitation** such as rain, snow, hail etc. add freshwater into the ocean, thus diluting the solution, and leads to a *decrease* in salinity. The combination is sometimes referred to as **EmP** (evaporation minus precipitation). Besides EmP, **river runoff** and **ice melt** leads to a *decrease* in salinity; correspondingly, growth of *sea ice* actually leads to whats called *brine rejection*, for the same reasons as the note above. It perhaps of interest to note that ocean averaged salinity can change quite drastically on very long time-scales, to do with climate transitions as we go in and out of natural *ice ages*, precisely because ice melt leads to decreases in salinity. Changes in salinity affect the density (and really, density gradients) in the ocean, which has important consequences for the MOC[26].

[25] Because a solution is needed to break the ionic bonds in salt. No liquid, no keeping ions apart, and salt crystals form again and are left behind. You could try this yourself with saltwater in a pot (but be very careful not to over heat the pot)!

[26] This is a topic in *paleoclimate*, which we won't touch on here.

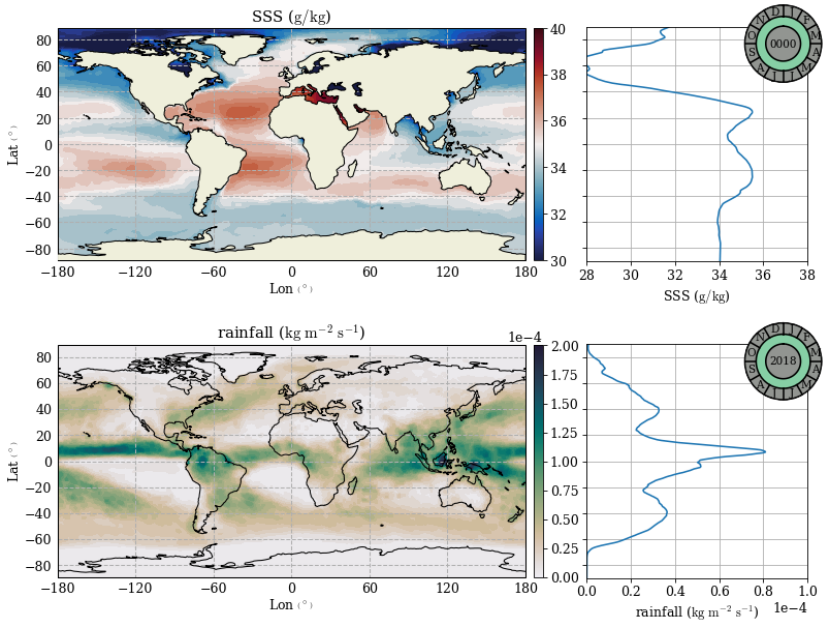


Figure 2.9: Year-averaged (left) and year and zonally averaged (right) sea surface salinity (top) and rainfall (bottom), from the JRA55 dataset Kobayashi et al. [2015]. See `plot_jra55_sample.ipynb` for code, and `sss_day_avg_2018.mp4` and `prra_day_avg_2018.mp4` for movies of the analogous daily averaged data through a particular year.

There are many fields we can show in relation to salinity but, for brevity, Fig. 2.9 shows the **SSS** (sea surface salinity) and average rainfall (in units of kg m^{-2} s^{-1}), again averaged over the year, and also showing zonal averages. The observation to note here is that SSS is typically large in the surrounding regions around the equator, though not as large generally as the Med Sea. Note however rainfall is also largest at the equatorial regions. This may seem contrary to what we have been talking about but remember it is EmP that ultimately matters. While there the heating is largest at the Equator, leading to large evaporation, large evaporation is also conducive to cloud formation, so precipitation is also largest in the region. For the

actual SSS signature, it is EmP that matters, and it turns out EmP is largest in the subtropical regions[27]. One interesting feature to note is that the Atlantic is noticeably more salty than the Pacific, and is largely to do with the MOC[28]; see more in Ch. 5.2.

In the supplementary movie files `sss_day_avg_2018.mp4` and `prra_day_avg_2018.mp4`, the latter showing the daily-averaged data over the year. You can see the seasonal cycle in the rainfall, but not so much in the SSS. The rainfall not so surprisingly follows Q_{sr} (why?) but the SSS displays smaller variations in the Equator, but is higher in the Arctic, because SSS is influenced by more than just EmP[29] The Atlantic is still saltier than the Pacific in general.

In Fig. 2.10 we show the vertical profile of salinity at some specific locations (not the same ones as Fig. 2.7), with red and blue lines denoting the (appropriate) summer and winter profiles. Note that we have similar features as in Fig. 2.7, although the winter signal is not so different to the summer signal. The equivalent of the thermocline for salinity is called the **halocline**. The deeper parts of the ocean again have weaker gradients, except in the Pacific (left column), where salinity is actually higher in the deep than in the intermediate layers. Unlike temperature this is ok: density increases with salinity, so the increase in salinity with depth denotes a stable stratification. The increase salinity as we go up the water column does not denote an unstable stratification however (why?)

[27] This is to do with the Hadley cell (mentioned in Ch. 1.3.2) suppressing precipitation but leaving evaporation sizeable. More in Ch. 3.3.

[28] We talked a bit about this in Ch. 1.3.2 and 1.2.3 when talking about Med Sea water overflows and the Aghulas leakage respectively.

[29] Also I cheated by showing monthly climatology instead of daily-averaged, which is what I had at hand.

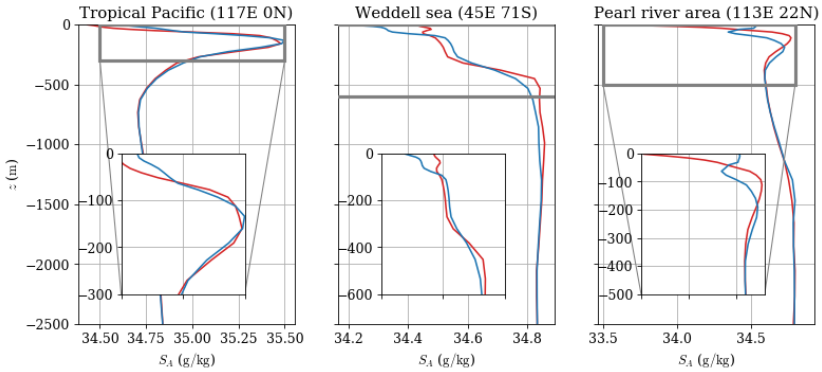


Figure 2.10: Vertical variation of absolute salinity at some designated locations, based on WOA13 data. Red and blue line denote summer and winter climatology. See `plot_WOA13_sample.ipynb`

Fig. 2.11 shows meridional sections of salinity, showing in particular the 35 and 36.4 **isohalines** (lines of constant salinity). As in the discussion relating to Fig. 2.8, we show this here in anticipation of the discussion on watermass properties when we talk about the MOC in Ch. 5.2. Previously we highlighted how there seems to be a warm intrusion at the North of the Atlantic. We see here that this is a warm *and* salty signal, and the North Atlantic water intrusion is much more obvious with the large blob of high-ish salinity at around

1000 m depth, as well as showing the intrusion in the deep into the Southern Ocean between 2000 and 4000 m depth. By comparison, there is a relatively fresher Southern Ocean intrusion into the Atlantic at intermediate depths (the blue tongue just above 1000 m) and also in the abyss. These signals correspond respectively to the NADW, AAIW and AABW watermasses respectively; see Ch. 5.2. In the Pacific there is also a tongue of freshwer seawater, again relating to the AAIW, and in this diagram we note that that salinity increases with depth in the Pacific from around a few hundred meters depth, which is consistent with the vertical profiles in Fig. 2.10(*a*).

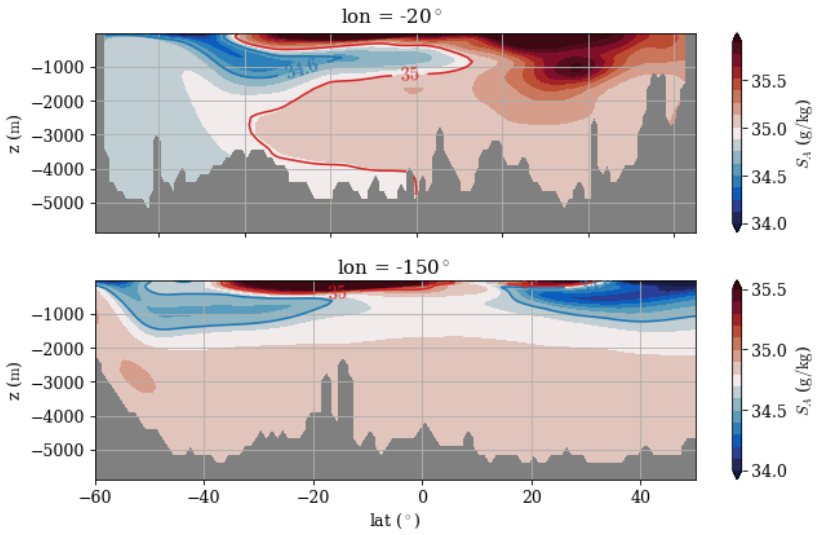


Figure 2.11: Meridional section of yearly-averaged absolute salinity in the Atlantic (top) and Pacific (bot), based on World Ocean Atlas 2013 data. Meridional range chosen to roughly correspond to Talley *et al.* (2011) Fig. 4.11 and 4.12. See `plot_WOA13_sample.ipynb`

2.3 *Density and equation of state (EOS)*

To recap, the density ρ of seawater depends on temperature T, salinity S, and technically on seawater pressure[30] p as well. The exact functional relation between ρ, T, S and p is called the **equation of state** (EOS). Before we go into details relating to the EOS, we motivate the discussion a bit more by talking about some observational details.

Fig. 2.12 shows the yearly-averaged sea surface density (actually showing $\sigma = \rho - 1000$ kg m^{-3} to save on writing so many digits), as well as the yearly and zonally-averaged SST and SSS. A fact to observe is that, while it looks like there is substantial variation, numerically it turns out over most of the ocean, the density[31] varies by no more than around 2% from a reference value of $\rho_0 = 1026$ kg m^{-3}. Of course numerically small doesn't mean it is dynamically small, and these minor differences have significant influences on the dynamics.

[30] But from a dynamical point of view we will want to get rid of the pressure dependence.

[31] I am deliberately being vague about the type of density being used (related to above note). The statement really refers to potential and *not* in-situ density.

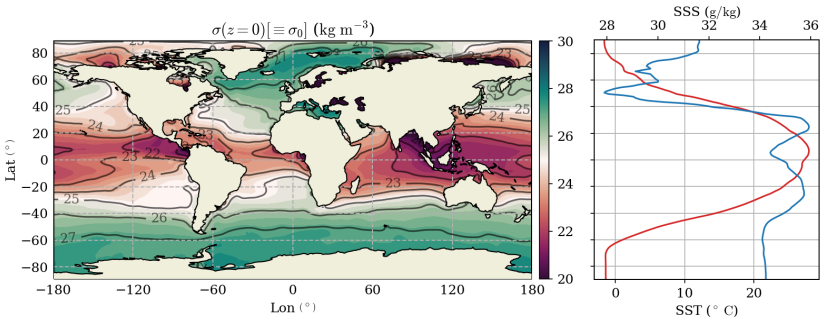


Figure 2.12: (left) Density at the surface (in-situ or referenced to sea surface) and (right) zonal averaged SST (red) and SSS (blue). Year-averaged data based on World Ocean Atlas 2013. See `plot_eos.ipynb`

Thus one point we note is that you want an accurate EOS, otherwise the dynamics and its resulting consequences could be substantially different. The second is that there is competition between T and S contribution to density, as seen in the correlations between the density and the SST and SSS graphs. Broadly speaking increase in T decreases density, whilst increase in S increases density. The more subtle question to ask is the magnitude of control of T and S on density, and where in the globe is one contribution more important than the other[32]

[32] Spoiler: over most the ocean it is temperature that is important contributor to density.

Just some more pieces of important terminology that will be frequently used in this document (more so than the ones relating to temperature and salinity actually). **Isopycnals** refer to the lines of constant density (cf. isotherms, isohalines, isobars in Ch. 3.1.2). The **pycnocline** is the equivalent of thermocline and halocline for density, i.e. where the change in density is largest, and again roughly delineates the upper and lower parts of the ocean (look ahead to Fig. 2.15 if you would like an example now). Two terms that will be increasingly used from Ch. 3.4 onwards is **along-isopycnal** and **diapycnal** (across-isopycnal) motion and mixing. It is the along-isopycnal and diapycnal components that are relevant for dynamics, rather than horizontal and vertical. Roughly, this is because motion across isopycnals requires moving the isopycnals, and this is doing *work* against buoyancy, while you don't need to do that if you go along isopyncals[33]. For example, while the ocean basin interior has relatively flat isopycnals so along-isopycnal and diapyncal happens to be horizontal and vertical, this is not true in the case of the Southern Ocean where the isopycnals are tilted (see Ch. 5.1).

[33] Think walking up stairs compared to walking on flat land. It is very tiring working against gravity! Blame Newton for this.

2.3.1 *Linear EOS*

For the EOS, we note that, since ρ is positively and negatively correlated with T and S respectively over a large part of the ocean, we

could try

$$\rho \sim \beta S, \qquad \rho \sim -\alpha T,$$

where we take $\alpha, \beta \geq 0$. The simplest thing to do with would be $\rho \sim -\alpha T + \beta S$, but this can go negative, so we want to add a few more things in. The standard way of doing it is actually

$$\rho = \rho_0[1 - \alpha(T - T_0) + \beta(S - S_0)], \qquad \alpha, \beta \geq 0, \tag{2.7}$$

where ρ_0, T_0 and S_0 are references that you choose depending on what you care about (see next paragraph). This is known as a **linear EOS**, because ρ depends on the arguments T and S linearly (notice there is no p dependence here). The coefficients α and β ('alpha' and 'beta') are respectively known as the **coefficient of thermal expansion** (in units of $°C^{-1}$) and **coefficient of haline contraction** (in units of g^{-1} kg)[34]. The way you use the linear EOS (2.7) is to choose the reference and parameters values, chose a T and S, put the numbers in, and you get a number back that tells you the density. That it! You can do this in a calculator or Excel.

[34] The name is because in material sciences the normal talk is thermal expansion (decrease in density), so for consistency we have haline contraction for talking about *decrease* in density.

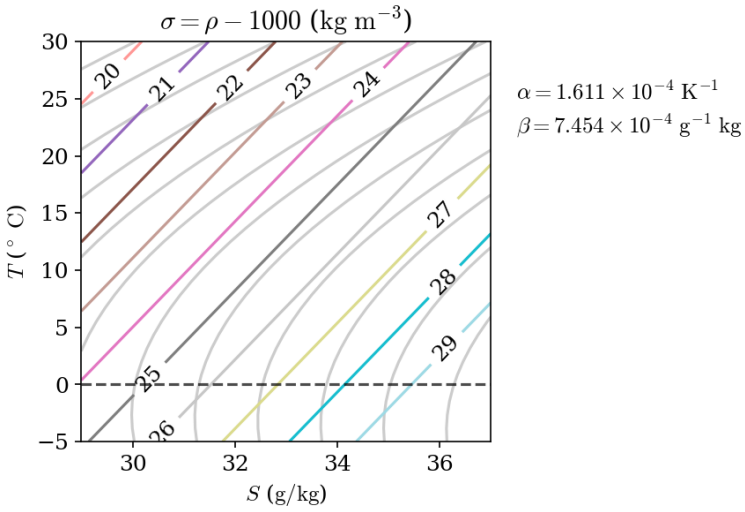


Figure 2.13: Linear EOS in TS space with TEOS-10 as gray contours (same contour levels). The reference values used are $T_0 = 10$ and $S_0 = 35$ in the usual units, as used in the NEMO ocean model [Madec, 2008]. See `plot_eos.ipynb`

To fully explore how ρ depends on T and S, I decided I really don't want to use Excel or do calculations one by one on a calculator (people used to!), because ain't nobody got time for learning Excel syntax. A few incantations in your favourite programming language later (I used Python here), the linear EOS over **TS space** is shown in Fig. 2.13; this is plotted on top of the "real" EOS of the ocean[35]. You can convince yourself that if you increase T or S the density is going the right way, so qualitatively speaking the right thing is happening. However, note that the agreement in terms of the sensitivities (the

[35] Of course the world is not nice enough to give us a linear EOS.... By the way, the "real" EOS used here is actually the *TEOS-10* standard, which is a model fitted from data (more later); we actually don't have an analytical form for the real EOS of the ocean, and not for a lack of trying.

contour slopes) is only reasonable at isolated locations, for example around the reference values $T_0 = 10$ and $S_0 = 35$ in the usual units[36], but is pretty bad in the upper left part of TS space. The linear EOS could be regarded as a leading order expansion of the "real" EOS about the reference, so it is maybe a reasonable approximation around reference, but there is no reason it is good away from the reference values (and indeed it isn't). However, if you don't think dynamical phenomena associated with nonlinear EOS is particularly important (e.g. away from regions where *deep convection* occurs, or near places where phase transitions happen, such as ice regions), then linear EOS is probably ok to use.

Another thing before we move on is that when we choose a reference (T_0, S_0), we also need to choose appropriate values for α and β so the resulting slopes in TS space does not deviate too much from the "real" EOS; this highlights another complication that the "real" α and β themselves are dependent on T and S, and indeed they do! This is perhaps not unexpected: recall that water is densest at around 4° C from the discussion near the beginning of this chapter, and is also seen in the bottom panel of Fig. 2.2.

[36] I'll probably start dropping units unless it is necessary for the prose or the discussion.

2.3.2 *Beyond linear EOS*

Although we know that we should not take the α and β parameters to be constant, a possible exercise to consider is whether we could get something reasonably simple but that is still a reasonable approximation to the "real" EOS. Roquet et al. [2015a] for example suggests something like (cf. their Eq.15 but simplified)

$$\rho = \rho_0 \left[1 - \alpha \left(T_a + \frac{\lambda_1}{2} T_a^2 \right) + \beta \left(S_a - \frac{\lambda_2}{2} S_a^2 \right) - \nu T_a S_a \right], \qquad (2.8)$$

where for cleaniness $T_a = T - T_0$ and $S_a = S - S_0$ are the temperature and salinity *anomalies*. The red terms in (2.8) now highlight the *quadratic nonlinearities*, but otherwise the coefficients α, β, $\lambda_{1,2}$ and ν are constant coefficients[37]. Fig. 2.14 shows how this nonlinear EOS (2.8) compares with the "real" EOS in TS space. With the appropriately chosen coefficients, the general agreement is actually pretty reasonable, except when the temperature is dipping below around 0° C, which is around when ice might start to form.

As mentioned in passing in a note above, we don't actually have a form of the "real" EOS derived from first principles. What we do have are very good approximations constrained by many seawater samples, which allows a *regression* (fitting) of a model to the data[38]. The "real" EOS is so important that there are UNESCO working groups dealing withe the relevant standards, because the density of

[37] The new coefficients are known as the *cabbeling coefficients*. I am skipped the *thermobaric coefficients*, which has an extra z term denoting depth dependence. Skipping the units here because this is just a side digression.

[38] Not dissimilar to the *data driven methods* that is the trend these days.

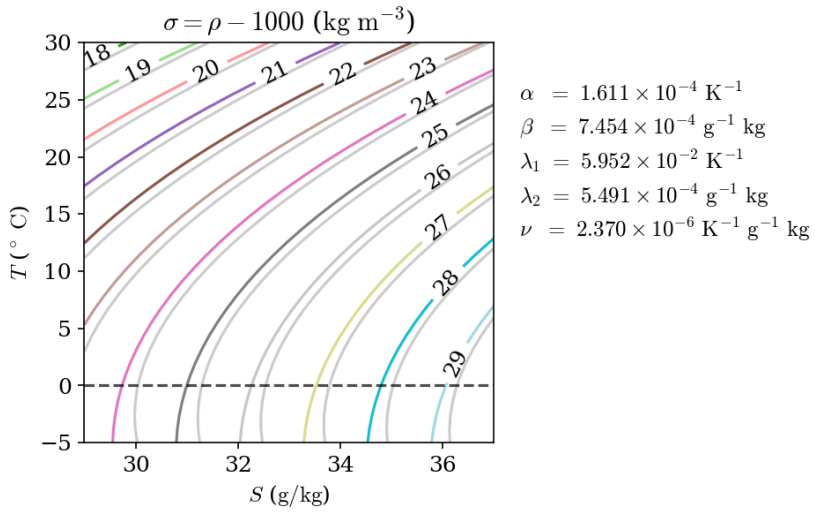


Figure 2.14: Toy nonlinear EOS (no *thermobaric effect*) in TS space with TEOS-10 as gray contours (same contour levels). See `plot_eos.ipynb`

fluids around us (air, water and seawater) leads to so many phenomena that directly affects our livelihood. The "real" EOS has evolved over the years, from the EOS-80 standard [Fofonoff and Millard Jr., 1983, Bryden, 1973] which for seawater uses practical salinity and *potential temperature* inputs, to the current TEOS-10 (Thermodynamic Equation of SeaWater 2010) standard, which for seawater uses absolute salinity and *conservative temperature* as inputs. One important difference is that in the older EOS-80, the thermodynamic quantities are not entirely consistent with each other, while this is fixed in TEOS-10 through taking a Gibbs function[39] approach, such that thermodynamic quantities of interest may be calculated all from the Gibbs function. It is also more sensible to link up conservative temperature needed in TEOS-10 with heat content (cf. Eq. (2.4) and Fig. 1.3), related to subtleties that is beyond the scope here. More on potential and conservative temperature in the next section. For completeness, the TEOS-10 EOS formula used to in Fig. 2.13 and 2.14 is the 75 term polynomial approximation given in Roquet et al. [2015b], which is computationally cheaper than evaluating directly from the TEOS-10 Gibbs' free energy (which makes a difference when it is used in a high resolution numerical ocean model).

[39] Or, more formally, the *specific Gibbs free energy*, a fundamental quantity in *statistical mechanics*. Named after the American physical chemist Josiah Willard Gibbs (1839-1903), who made fundamental contributions to thermodynamics, statistical mechanics (he coined this term), and developed modern day vector calculus (independently of Oliver Heaviside).

2.3.3 *The problem with in-situ density*

Right, finally, lets talk about why we mostly don't care about in-situ density if we are considering dynamics, as advertised at the beginning of this chapter, and alluded throughout this chapter where

there are repeated undefined references to *potential / conservative* temperature and *potential* density. This following discussion was never that intuitive to me (I don't have a very good affinity with *thermodynamics*), but lets give this a go.

Maybe lets start with *how* in-situ density 'fails' in an example before we go onto *why* it 'fails'. Fig. 2.15 shows a meridional section and a vertical profile in the Atlantic for sake of choosing something, and the top panel and the blue line shows **in-situ density** in the form $\sigma = \rho - 1000$, which is the density a water parcel would have if you measured its density *at* the depth where you collect the water sample, with $\rho = \rho(T, S, p)$. The first thing we note that is density is increasing dramatically with depth. From the in-situ density data, we have to conclude there is basically no net up and down transport of water, because the ocean is so stably stratified, so there is a negligible MOC. It would also imply there is no strong transport in the Southern Ocean (from *thermal wind shear relation*, see Ch. 5.1.3).

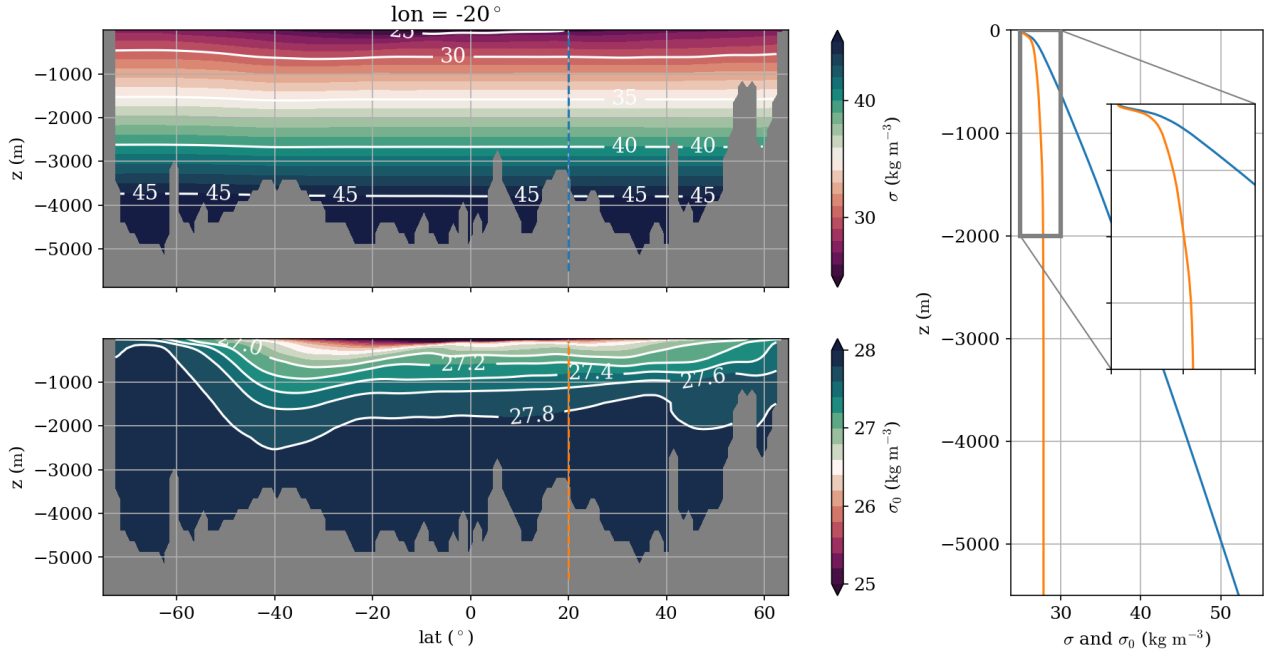


Figure 2.15: Meridional section in the Atlantic of (top left) in-situ density and (bot left) potential density referenced to sea level, with the corresponding vertical profiles plotted (right). See `plot_eos.ipynb`

But we know that is wrong! We know there is up and down transport of water in the ocean, for example in Fig. 2.11 in the salinity signature where there is clear signal of water being dragged down from the surface into the interior in the North and South Atlantic. We know there is a strong ACC taking around 130 Sv of water around the globe, which means we should see tilting isopycnals. Another thing to observe is that below the thermocline and haloclines in

Fig. 2.7 and 2.10, the temperature and salinity is relatively constant, but there is a fairly substantial increase in the in-situ density, indicating something else is influencing the density. We do have a MOC, and there are additional signatures in other chemical tracers that serve as a *proxy* for age of watermass last in contact with the atmosphere (e.g. radioactive carbon), but I think this is enough evidence to point to in-situ density being an unsuitable measure for density from a circulation point of view, certainly when the deep ocean is involved. By contrast, the bottom panel and orange line in Fig. 2.15 shows the *potential density* (referenced to the sea surface), and lo and behold! Everything seems to be consistent again, where we have a pycnocline, the deeper parts where the ocean stratification decreases to a constant, tilting of isopycnals in the Southern Ocean, and so forth. So what is this potential density, and what is going on?

You might have guessed, the culprit is pressure p, and potential density is a measure that removes *some* (not all) contributions from pressure to the density. So the complication with seawater, while not as 'squishy' as air, is still *compressible*. Seawater remember is very heavy ($\rho_0 = 1026$ kg m^{-3}), so if you imagine a parcel of water below the sea surface, the pressure pressing down on the parcel would be approximately related the amount of water above the parcel[40], so naturally there is a squashing of the water parcel below the ocean surface.

One concept we need to introduce is that of **work done**. Suppose we have a water parcel at some depth such as in Fig. 2.16. The inward and outward pointing forces balance (so there is no net motion), and the equilibrium position is given by the black dashed line. The inward pointing forces would be external pressure, and outward pointing force would be some internal pressure of the fluid which is trying to resist being squashed. Then, to squash the parcel further, we need to increase the external forces (assuming here the outward force is just whatever the parcel needs to do to resist being squashed), but to do this, we need to put **energy** into the system, i.e. by doing work. If we did this, we *added* to the internal energy of the parcel, because we put work into the system, and we also decreased the volume. If we regard temperature and energy as related (e.g. Eq. 2.4), then the parcel should also experience an increase in temperature.

Now if this parcel is moved up and down as in Fig. 2.17 and, in particular, it is moved in such a way that there is no exchange of mass or heat with the surroundings, then the only change in the internal energy of the parcel is due to work done (pressure volume work). If we assume there is no dissipation, then this kind of process is *reversible*, in the sense that you can in principle recover all the

[40] This is the *hydrostatic approximation*; see Ch. 3.1.2.

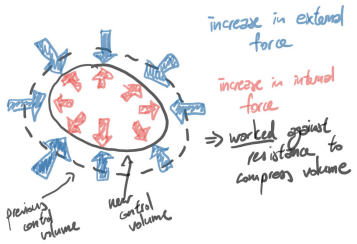


Figure 2.16: Working to compress a volume.

All of this paragraph concerns the 1st *of thermodynamics* $\Delta U = Q - W$ where ΔU is the change in internal energy (related to temperature), Q is the added heat, and W is work done by system to its surroundings. For adiabatic processes, $Q = 0$ (no exchange of mass and heat), and with increased pressure on the fluid parcel, $W < 0$ (work is being done to the fluid), so $\Delta U > 0$ and temperature increases. Without dissipation, the process is *isentropic*, i.e. $\Delta S = 0$ where S is the *entropy*, and process is reversible in this sense (cf. 2nd *of thermodynamics*).

energy that has been put into the system as work done. However, when we are concerned we dynamics we are usually interested in *irreversible* processes, and in this sense we are not interested in the change in internal energy due solely from pressure effects.

This effect is particularly noticeable in the in-situ temperature signal when we go deep into the ocean. Fig. 2.18 shows the in-situ temperature T and potential temperature θ profile with depth in the Mariana Trench in the Pacific, which goes beyond 10,000 m depth. Note that the T profile after a certain point *increases* with depth, denoting unstable stratification, and we might expect to see overturns that erodes these unstable gradients, so we shouldn't even be seeing these things in the first place! This kind of configuration surely cannot exist, unless there is an accompanying increasing in salinity (and there isn't). By contrast, the use of θ gets rid of this effect, and indicates there is a weak stratification at depth at least in terms of temperature.

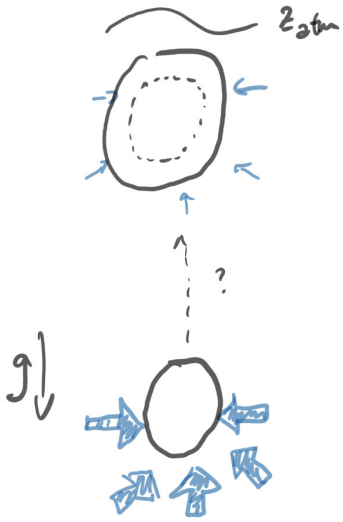


Figure 2.17: Parcel moving around adiabatically. Assuming there is no dissipation, then this process is *isentropic* (*entropy* doesn't change) and *reversible*.

2.3.4 *Potential density and neutral density*

When we are talking about quantities that are potential we need to define a *reference depth* (and therefore a *reference pressure*). **Potential temperature** θ of a water parcel is the temperature this parcel would have if you went to the depth of that water parcel, put it in a plastic bag, seal it and move it the reference depth/pressure (as in the *adiabatic process* described above), and measure the temperature of that bagged up water parcel has at that depth/pressure. Usually sea surface is used as the reference depth, but in principle other depths maybe used.

Potential density referenced to the sea surface ρ_θ would be the density computed from the appropriate EOS with the potential temperature θ, and the reference depth and pressure used to compute θ, i.e. if P_{atm} is used to compute θ, then $\rho_\theta = \rho(P_{\text{atm}}, \theta, S)$, in contrast to $\rho = \rho(p, T, S)$. Depending on the application, $\rho_{1,2,3,4}$ are also seen if the reference depths (and the related pressures) of 1/2/3/4000 m are used. Examples are seen in Fig. 2.19. The choice of reference makes a difference depending on the ocean depth of focus, as potential density doesn't remove all pressure contributions. Compare this with the linear EOS as a leading order approximation near the reference: the method is good near the reference but is not as good elsewhere.

The **neutral density** denoted γ^n can be thought of as the continuous analogue of potential density, where the references are all local. Then the idea is that at every point in the ocean over a very small region there is a direction that the parcel can move without experiencing any buoyancy forces, so this is the along-isopyncal direction

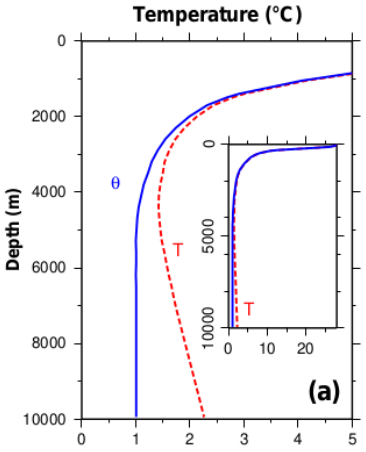


Figure 2.18: Vertical profile of in-situ (red) and potential temperature (blue) in the Mariana Trench to highlight the differences. From Talley et al. [2011], Fig 4.10(*a*).

(or *epineutral* direction). So one would think we can construct these **neutral surfaces** and water parcels have to move along these surfaces. Unfortunately it turns out you can prove these surfaces generally don't exist! Saying that, you can get very good approximations to them, and in practice it is these approximations that are used to compute the 'neutral density'. There are existing algorithms to work these out, and sometimes (approximate forms of) γ^n are used to identify watermasses (see Ch. 5.2.2).

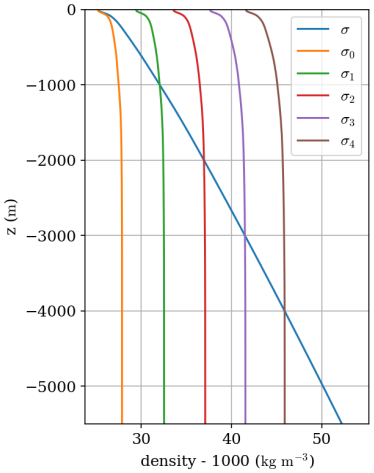


Figure 2.19: Vertical profiles of in-situ and potential density (referenced to various depths) at the same location as in the previous graph. See `plot_eos.ipynb`

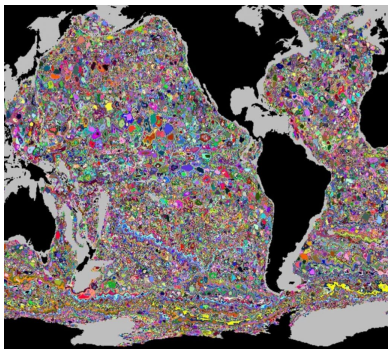


Figure 2.20: Topobaric surfaces, an almost neutral surface. From Stanley [2019], Fig. 4.

Summary and further reading

We close this section by noting again that, FROM A DYNAMICAL POINT OF VIEW, IT'S ALMOST NEVER IN-SITU DENSITY YOU CARE ABOUT unless you are really close to the sea surface, otherwise the irrelevant compressibility effects arising from pressure needs to be removed for the resulting data to tell you relevant things.

Science and scientific standards are continually evolving, as more research is carried out, sometimes meaning what we might have treated as gospel truth is, in fact, not, and we can potentially do a bit better. For example, it turns out that the chemical measure of salinity through chlorinity has some problems, since it makes the assumptions that all dissolved salts have the same ratio everywhere in the ocean, but turns out while the deviation from this assumption is small, the deviations make enough of a difference to matter. For example, for a long time it was international standard to use practical salinity and it was a good idea to append units to something that was non-dimensional (GET RID OF THE PSU!) Now this is no longer true, as practical salinity is being recommended to be phased out and replaced with absolute salinity. This is not to say we stop measuring salinity by conductivity, but that we use a formula to relate the resulting salinities[41]. It turns out also potential temperature is a better variable to use than in-situ temperature as it gets rid of pressure effects, but it is not as good compared to **conservative temperature**, which is a more appropriate temperature-like variable to describe ocean heat content [McDougall, 2003]. It is conservative temperature and absolute salinity that are now adopted as the go-to variables as part of the TEOS-10 standard [IOC et al., 2010]. Neutral density for a long time was touted as the best thing, and increasingly there are alternative ideas and/or better and more efficient approximations coming through (e.g. *topobaric surfaces* from Stanley [2019]; see Fig. 2.20 here).

[41] The better practice really is to record the raw data, from which processed data may be derived, and having a trail leading back to the primary data, which is what is being advocated.

Chapter exercises

1. What are the units of buoyancy $b = -(\delta\rho/\rho_0)g$?

2. Why is the statement "*a pig weighs around 200 kilograms*" not technically correct?

3. Look up the specific heat capacity of air at around room temperature (300 K say) and compare this with the specific heat capacity of water. Is this what you expected? if you put the amount of energy that it takes to warm up 1 kg of seawater by 1 K, how much would 1 kg of air roughly warm up by?

4. For a further challenge, make an estimate of how much volume 1 kg of air occupies, estimate the volume of the atmosphere, work out the amount of energy it takes to raise the whole atmosphere by 1 K (could just do the troposphere for simplicity), and the volume of sea water that could be raised 1 K by this amount of energy. Given this volume and that we know the area the ocean covers, what is the associated depth?

5. Chemistry one: how do you expect conductivity (and hence salinity measurements) to depend on temperature? Justify your answer.

6. Explain why we expect outgoing long-wave radiation Q_{lr} to correlate well with SST patterns.

7. What might happen to global sea level over long time-scales if for whatever reason we blocked off the Strait of Gibraltar so there are no water exchanges possible? Answer this in terms of EmP structure in the region.

8. Harder question: estimate how much global sea level might change in the question above, and over how long (you need to work out some volumes here).

9. In Fig. 2.10 we see that at certain locations salinity increases as we go up from the bottom towards sea surface. Why does this not necessarily imply an unstable density stratification?

10. In the linear EOS used in Eq. 2.7 as written, how should we interpret the density, as in-situ, potential, or some others? Does it matter? What about for the nonlinear EOS in Eq. 2.7 as written? What about in the same nonlinear EOS Eq. 2.7 but with the *thermobaric* terms added in (you might need to look this up in Roquet et al. [2015b] for what that word means)?

11. Make plots of the EOS diagrams yourself and explore the dependencies to parameters accordingly. Use some of the provided code on the GitHub repository as a reference if you like.

12. classification of estuaries

3 Mechanical forcing

The last chapter dealt with things that affect T, S and ρ accordingly. Here in this chapter we focus on the mechanical forcing terms that affect the momentum equation, given by

$$\rho_0 \left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + 2\boldsymbol{\Omega} \times \boldsymbol{u} \right) = -\nabla p + \boldsymbol{F}_{\boldsymbol{u}} + \boldsymbol{D}_{\boldsymbol{u}} \tag{3.1a}$$

$$\frac{\partial p}{\partial z} = -\rho g. \tag{3.1b}$$

We have omitted $\nabla \cdot \boldsymbol{u}_3 = 0$ because that should be thought of as a constraint of mass conservation, i.e. no creation or destruction of volume if the flow is non-divergent with no sources and sinks. Recall that $\boldsymbol{u} \cdot \nabla T$ referred to the advection term for temperature so, similarly, $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$ refers to the self-advection of a fluid parcel and is called the **inertia** term. A major difference here is that $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$ is *nonlinear*, and is a major source of why fluid dynamics generally is so hard/interesting, as this term leads to phenomena associated with *turbulence*. We will not touch on this term too much in this document, as most things we talk about invariably tries to get rid of or approximate this term in some way.

3.1 Gravity and pressure

Given we just talked about the thermodynamic variables, the first thing we talk about is

$$\frac{\partial p}{\partial z} = -\rho g.$$

This actually comes from the vertical component of the approximated momentum equation. The notable term here is g, which leads us to talk about *gravity*. Gravity is important because buoyancy $b = -g\delta\rho/\rho_0$ needs gravity, and buoyancy as argued in Ch. 2 places a strong constraint on the dynamics (because working against gravity is unfavourable in terms of *work*). Gravity is also important for *static instabilities* (Ch. 6.2) and *tides* (Ch. 6.3), so we spend a little bit more time on outlining some related concepts.

3.1.1 *Gravity*

Recall from Ch. 2 that mass is a scalar and is how much 'stuff' a body possesses. Weight on the other hand is a vector because it is a force, related to gravity and mass.

One of the biggest successes of Newton's formulation of mechanics was for explaining how the heavenly bodies moved relative to each other, via **gravity**, but the theory of gravitational attraction between bodies with mass is general. Newton's formulation of gravity has that the force arising from gravitational attraction between two bodies of mass m_1 and m_2 is given by

$$F = G\frac{m_1 m_2}{r^2}, \tag{3.2}$$

where $G = 6.674 \times 10^{-11}$ m^3 kg^{-1} s^{-2} is the *gravitational constant*, and r is the distance of separation between the two bodies. If we have multiple bodies we just add up the forces for each pair and proceed, but we will stick with two bodies for the discussion. A schematic of this is given in Fig. 3.1.

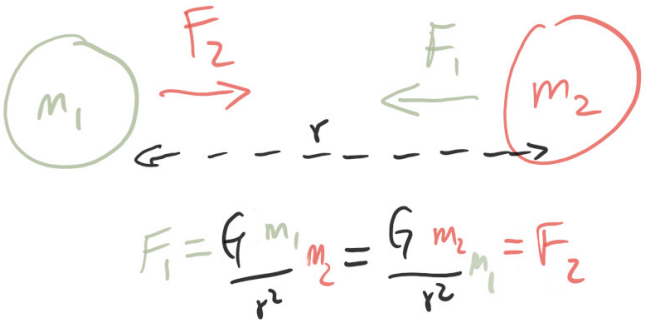


Figure 3.1: Schematic of gravitational attraction for two (supposed to be) spherical masses. If $m_1 \gg m_2$ (e.g. Earth and a pig) then forces on each body are equal, but its effect on one the pig is much larger than it is for the Earth (recall $F = ma$).

Three things to note about gravity are that:

- gravity is purely *attractive* (cf. *magnetism* that can attract and repel)
- the force goes like $1/r^2$, i.e. the force is weak at large distances
- there is an equal and opposite force (Newton's 3$^{\text{rd}}$ law): the larger mass attracts the smaller mass, and vice-versa

So where does g, the magnitude of **gravitational acceleration** come from? Taking an example, consider our friendly neighbourhood pig again as in Fig. 3.2. We group the terms in Eq. (3.2) as

$$F = m_{\text{pig}}\left(G\frac{m_{\text{earth}}}{r_{\text{earth}}^2}\right) = m_{\text{pig}}g, \tag{3.3}$$

i.e. g is the magnitude of acceleration towards Earth that the Earth causes a body (in this case a pig) to have. If we be a bit cavalier with degrees of accuracy and take $G \approx 6 \times 10^{-11}$ m^3 kg^{-1} s^{-2}, $m_{\text{earth}} \approx$

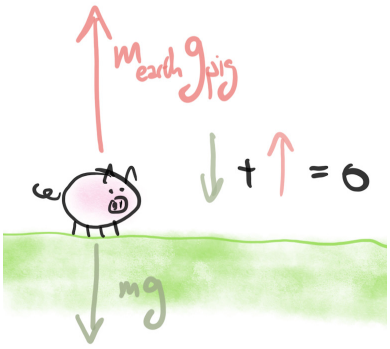


Figure 3.2: Gravity as applied on Earth on the friendly pig.

6×10^{24} kg, and $r_{\text{earth}} \approx 6400$ km $\approx 6 \times 10^6$ m, then (check the units agree)

$$\begin{aligned} F &= m_{\text{pig}} 6 \times 10^{-11} \frac{6 \times 10^{24}}{(6 \times 10^6)^2} \\ &= m_{\text{pig}} \frac{6^2}{6^2} \times 10^{-11+24-12} \\ &= m_{\text{pig}} 10 \\ &\equiv m_{\text{pig}} g, \end{aligned}$$

i.e. $g = 10$ m s^{-2}, which is an approximation often used in standard mechanics exercises to keep the numbers nicer (in reality $g \approx 9.81$ m s^{-2}). The weight of the pig on Earth is defined by F. If the pig is on the moon (lets assume for animal rights reasons it is wearing a space suit of negligible mass), then the mass of the pig would still be m_{pig}, but the pig's weight on the moon would be about $1/6$ of its value on Earth (because g_{moon} is about a $1/6$ smaller).

The schematic drawn in Fig. 3.1 assumes spherical bodies of uniform mass, and the Earth satisfies neither of those conditions! It is not spherical because it is spinning about an axis[1], and it is certainly not a body of uniform mass. An exaggerated version of how the Earth's mass is distributed is given in Fig. 3.3, which shows the variation in the *geoid height*. You can think of where the bulges are to be where there is more mass.

To define the geoid height properly we need to say what the **geoid** is first. So I personally always thought I knew what the geoid was, which is *the surface that gravity is perpendicular to everywhere*, i.e. the red line in Fig. 3.4, with the orange vectors (known as *plumb lines*) denoting $\boldsymbol{g}$. So it turns out this is not quite right, and what I just described is what is called an **equipotential surface**. Gravity is what is called a *potential force*, i.e. you can write $\boldsymbol{g} = -\nabla\phi$ for some ϕ (this is called the **geopotential**), and what I just described is a surface of $\phi =$ constant. Immediately you can argue on why the definition I was using doesn't quite work, because "the" geoid implies one, but there are an infinite number of geopotential surfaces depending on the constants I choose for $\phi =$ constant (I can shift the red line in Fig. 3.4 up in height accordingly and I can still get $\boldsymbol{g}$ to be perpendicular to it). So while the geoid is certain an equipotential surface, the actual definition is of the geoid is that *the shape that the ocean surface would take under the influence of the gravity and rotation of Earth alone, if other influences such as winds and tides were absent*. This definition certainly implies there should only be one such surface, but I don't know about you, but I think the one I've been working with is easier to remember...

So the **geoid height** is the signed deviation between the geoid and

[1] So the mass is 'flung out' near the equator, and is closer to an ellipsoid (see purple dashed line of Fig. 3.4), sometimes modelled as what's called a *Maclaurin spheriod*. After the Scottish mathematician Colin Maclaurin (1698-1746), who was a colleague of Newton. Until 2008 he held the record of being the youngest person to ever hold a professorship.

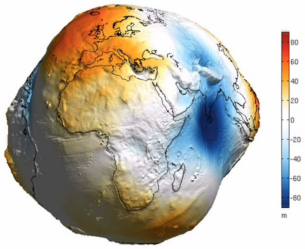


Figure 3.3: The "lumpy potato" Earth, variations in the geoid height magnified by several orders of magnitude to highlight difference. From Earth Gravitational Model 2008.

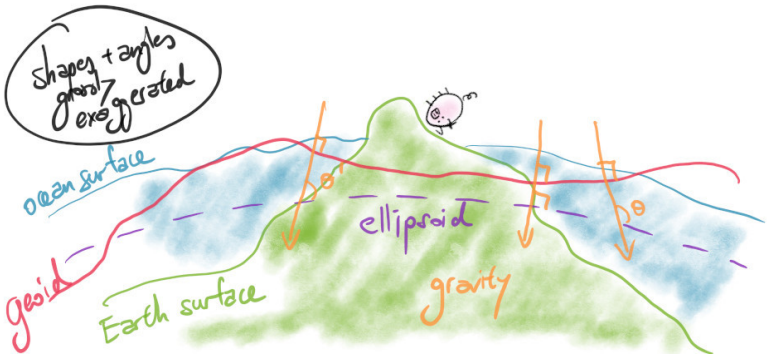


Figure 3.4: Schematic of the ellipsoid and geoid. While gravity is perpendicular to the geoid, it is not perpendicular to the ellipsoid.

the ellipsoid height (the difference between the red and purple line in Fig. 3.4), where the ellipsoid is chosen to be something. There are internationally agreed standards on these (e.g. WGS84) but since we are not going to be talking about *sea level* that much we will leave it as a detail for the reader to chase up. So Fig. 3.3 shows the (magnified) deviations of the Earth from an ellipsoid, arising from the fact that the Earth is not an uniform mass object, with the bulges (the red shading) where there is *more* mass, and dimples where there is *less* mass. Since gravity plays such an important role in the physics on Earth, we really do in fact need very accurate measures for the geoid, and this has only really been possible since the satellite era; we come back to the measure of gravity and the geoid in Ch. 7.

Before we leave this section, we mention some subtleties in relation to defining *sea level*, and even **sea surface height** (SSH). As you can be seen in Fig. 3.4, you could for example have

1. sea level above/below ellipsoid,
2. sea level above/below geoid,
3. sea level above sea floor.

Following Gregory et al 2019 Fig. 2, for the present purpose we will define 'vertical' to be relative to the *ellipsoid*, and the (**geodetic**) 'height' means distance measured in the direction perpendicular to the ellipsoid. With that, the **mean sea level** will be the time-averaged sea level above the ellipsoid, SSH will be the *instantaneous* sea level above/below the *ellipsoid*, and the **dynamic sea level** is the difference between the mean sea level and the geoid. Note that gravity is perpendicular to the geoid and *not* the ellipsoid, so dynamically speaking we are probably more interested in the height relative to the geoid. While in practice the differences between SSH relative to ellipsoid and geoid are small in magnitude, these subtleties are important for example when measuring SSH (see Ch. 7) or talking

about sea level rise, because different processes lead to different types of sea level changes.

3.1.2 *Pressure and hydrostatic balance*

We encountered *pressure* briefly in Ch. 2.3.3 when we were talking about density and the confusing things it can do to when defining temperature and density. Here we define pressure properly, and highlight the important link between density variations and pressure as well as its consequences for momentum.

If we imagine again we are trying squash a parcel of water (cf. Fig. 2.16), then we have to exert a force which is spread over some area. This is what **pressure** really is, the *magnitude of force exerted per area*, or

$$p = F/A, \tag{3.4}$$

where A is the area and $F = |\boldsymbol{F}|$ is magnitude of the force vector; note pressure is a scalar. Pressure has units N m^{-2}, sometimes in Pa (Pascals), and sometimes in (milli)bar in the atmospheric and oceanic literature[2]. For reference, sea level is sometimes defined as where the pressure is 1 bar $=$ 1000 mbar (millibar) $=$ 10^6 Pa, ocean depth is sometimes measured in bars (e.g. CTDs in Ch. 7), atmospheric weather charts is usually given in units of millibars (see Fig. 3.7 for an example), and atmospheric data is sometimes given in pressure coordinates measured in hPa (*hecto*pascal, 1 hPa $=$ 1 mbar) instead of height coordinates[3]. The lines of constant pressure are called **isobars**.

In the ocean, possibly as anticipated from the discussion in Ch. 2, the principal source of pressure comes from the weight of the seawater above a point, because the weight of seawater is so large. A schematic of this pressure due to the water column is given in Fig. 3.5, where the pig in the water bubble is at depth $-z$. If we assume that the pressure experienced by our pig demonstrator is proportional to the amount of water above it, then what this is saying is that, at depth $-z$, the pressure experienced is

$$p = mg = \left(\int_{-z}^{z_{\text{atm}}} \rho(z')\, \mathrm{d}z' \right) g, \tag{3.5}$$

where the mass of the water column is given by the integral of density from depth $-z$ to the surface. The principal contribution to pressure below sea level really is from this **hydrostatic pressure**, so we are not doing anything that drastic. If $\rho =$ constant then we recover maybe the more familiar form $p = \rho g z + p_{\text{atm}}$, with $p_{\text{atm}} =$ constant. If we take a derivative of the equation with respect

[2] Pascals is after French scientist Blaise Pascal (1623-1662), who also made contributions to probability theory as well as philosophy and theology. Bar as a unit was introduced by Vilheim Bjerknes (the father of the guy mentioned in Ch. 1.2.2)

[3] Because pressure is more relevant than height dynamically in the atmosphere, similar to how in the ocean we care more about along and across isopycnal directions.

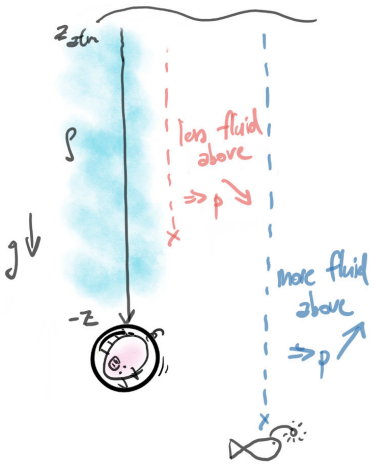


Figure 3.5: Schematic of hydrostatic pressure.

to z, then

$$\begin{aligned}\frac{\partial p}{\partial z} &= \frac{\partial}{\partial z}\left(\int_{-z}^{z_{\text{atm}}} \rho(z')\,\mathrm{d}z'\right) g \\ &= -g\frac{\partial}{\partial z}\left(\int_{z_{\text{atm}}}^{-z} \rho(z')\,\mathrm{d}z'\right) \\ &= -g\rho(-z),\end{aligned}$$

which is called **hydrostatic balance**, and you might recognise this already as Eq. (3.1b) where I've written out the argument of the function ρ explicitly[4]. If we are higher up in the water column, we experience less pressure, and lower down we experience more pressure, as in Fig. 3.5. The upshot is that we have a relation between density and pressure, so if we make the leap that pressure drives fluid motion, then density drives motion.

You should immediately object to what I just wrote, because pressure itself has no direction associated with it, and it is **pressure gradients** (and thus density gradients) that drives flows. Hence even though the numerical value of density does not vary much in the ocean, the point is somewhat mute because it is the gradients that matter. Fig. 3.6 shows schematically how pressure gradients drive flows in the ocean. Making the assumption that ρ = constant, the system is non-rotating, and the whole water column moves as one for simplicity, intuitively what we expect is that positive SSH regions will slump via the action of gravity, to fill the water columns that are in deficit, which is only achieved if there is a horizontal movement of the fluid. If we are look from a pressure point of view, then hydrostatic pressure is high where the SSH high, and hydrostatic pressure is low where the SSH is low. Intuitively we expect things at high pressure wants to go to low pressure to even out the pressure differences, so the action or force is pointed towards $-\nabla p$. A force pointing in the direction of the *negative pressure gradient* $-\nabla p$ leads to an acceleration in that direction (since there is a net force), and so the fluid moves from high pressure to low pressure. If the fluid surface is fully slumped (which will happen if there is damping and the system is non-rotating[5]), there is no pressure gradient and therefore no force, so nothing should move, as expected.

So far so good apart from the extra assumption about rotation. We expect that flow to be in the direction of $-\nabla p$, but is that seen in observations? Fig. 3.7 shows a weather chart and the contours are isobars, and I've taken the liberty and some license in adding the flow in. From the discussion of the geoid, or otherwise, $-\nabla p$ points *across* the isobars (marked in green in figure). So the answer is *no*, it doesn't really work, because most of the flow is *along*-isobars and not across isobars! The same thing turns out to hold true in the ocean

[4] The minus sign is just because I happen to have chosen to denote my depth as $-z$.

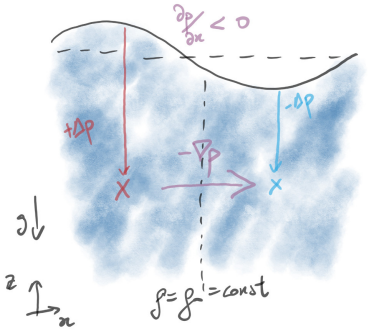


Figure 3.6: Horizontal effect because of hydrostatic pressure (assume ρ = constant and non-rotating for simplicity).

[5] Look up *geostrophic adjustments*, which we don't talk about here.

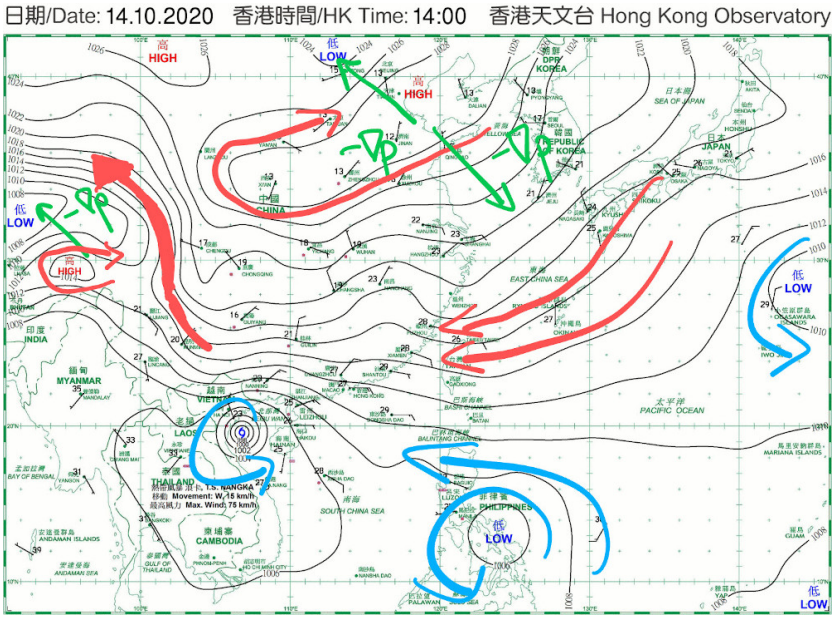


Figure 3.7: Atmospheric weather chart with isobars (in units of hPa = 100 Pa = mbar) and wind directions (the tails of the flags denote where the wind is *coming from*. From HKO.

for larger-scale flow. Fig. 3.8 shows what is called the **mean dynamic topography**, which is just the time-mean SSH. Now from Fig. 3.6 we argued that SSH is related to hydrostatic pressure, so again contours of SSH roughly correspond to the horizontal distribution of isobars in the ocean. The subtropical gyres (the ones adjacent to the equator) we know for example to be rotating anti-clockwise and clockwise in the Northern and Southern Hemisphere respectively (Ch. 1.2). Thus $-\nabla p$ points across contours of SSH, but the flow turns out to be largely along contours of SSH! So what is going on?

If you know some of this material, you will know I am actually cheating quite a bit here. If you look closely in Fig. 3.7, the wind direction is almost but not quite in the direction along the isobars (hence why I said I took some license in drawing the flow on...) We also don't really have such a large coverage of ocean current measurements, so the current I drew on in Fig. 3.8 is largely correct, but really it's done with hindsight through theory that we will go through in detail in the next section. Pressure really does contribute to driving the flow, but *rotation* plays a big role. It turns out the length and time-scales associated with the dynamics matter: on short time- and length-scales, rotational effects are weak, so pressure can dominate (e.g. pipe flows, your water tap, small-scale atmospheric flow), and flow can be in the direction of $-\nabla p$. On longer time- and larger length-scales, rotation effects dominante (e.g. atmospheric *jet streams*, ocean gyres, etc.), and there is a deflection of the flow away from the direction of $-\nabla p$.

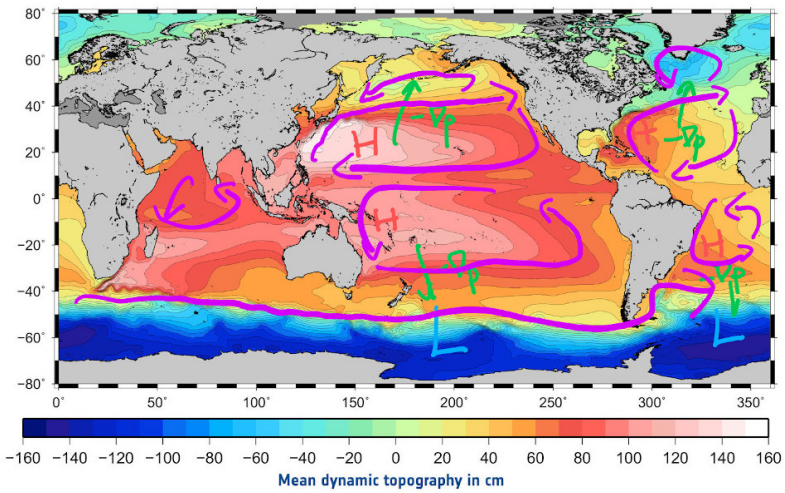


Figure 3.8: Time-mean global SSH (mean dynamic topography), with time-mean currents drawn on (notice the orientation around high/low SSH regions). Modified from Rio *et al.* (2011), J. Geophys. Res: Oceans.

For completeness, the **hydrostatic approximation** is the assumption a lot terms are small in the vertical momentum equation for the evolution of the vertical velocity w, such that we can throw away most of the terms except the ones relating to hydrostatic balance. This approximation is ok when the motion has small aspect ratio H/L, which is generally true in the ocean. From a computational point of view what this means is that instead of evolving three three-dimensional fields $\boldsymbol{u}_3 = (u, v, w)$, we only have to evolve $\boldsymbol{u} = (u, v)$, and find w through $\nabla \cdot \boldsymbol{u}_3 = 0$, which is more efficient.

3.2 *Coriolis effect*

The *Coriolis effect* is represented in the momentum equation Eq. (3.1) by

$$2\boldsymbol{\Omega} \times \boldsymbol{u}$$

So for a long time I knew how to use the symbols associated with the *Coriolis effect*[6] and what it does from mathematical point of view, but until I had to teach it I didn't realise how much I don't actually really understand it, and for a good while I couldn't explain it in a manner that I could convince myself. The drawing-a-line-while-rotating-a-piece-of-paper example (Fig. 3.10) was the one that really convinced me what it really is, and I highly recommend you try this out yourself; the Coriolis is probably one of those that you have to convince yourself is a 'real' effect, and other people telling you it is a thing doesn't really work... The Coriolis effect is going to be one of the central concepts that we use regularly until the end of the document, and it would help immensely in my opinion to spend

[6] After the French mathematician and scientist Gaspard-Gustave de Coriolis (1792-1843). Though well known in meteorology, he himself never worked in meteorology.

some time to learn and really understand this not entirely intuitive concept, before we utilise it extensive to do 'fun' (!) things with it[7].

So the Coriolis effect (note I use 'effect' and not 'force') is a *pseudo-force* or a *fictitious force*: it only arises because of the choice of perspective[8]. It turns out the choice is either choose an unwieldy perspective that gets rid of this effect (an *inertial frame*), or choose a more practical perspective but live with the Coriolis effect (an accelerating frame), and we actually choose the latter. If you already don't like the sound of this, I would still urge you to carry on, but for practical reasons, these are the key takeaways of this section:

- noting that the *work done* (Ch. 2) on a fluid parcel is $\boldsymbol{F} \cdot \boldsymbol{u}$, the Coriolis effect does no work (so it's a "fake" force from the work done point of view);
- if the system is not rotating, or there is no flow, there is no Coriolis effect;
- the force is to the *right* of intended travel in the Northern Hemisphere, and to the *left* in the Southern Hemisphere;
- the Coriolis effect on the (locally) *horizontal* flow is largest in magnitude at the poles, and vanishes at the equator;
- the *geostrophic flow* arising from a balance between pressure gradients and Coriolis effect is along isobars (cf. Fig. 3.7 and 3.8).

With that, lets dive in...

[7] If it helps, Coriolis effect also applies to atmospheres and generally large rotating bodies of fluid such as Jupiter, so you get extra applications for free!

[8] This is related to the talk about *inertial frames* in Ch. 1.4.1.

3.2.1 *Terminology and rationalisation*

Some terminology first. The **rotation axis** is the axis which the Earth rotates around, and is taken to be the (geographical) North Pole; see Fig. 3.9. Then $\boldsymbol{\Omega} = \Omega \boldsymbol{e}_{\hat{z}}$ is in the direction of the North Pole $\boldsymbol{e}_{\hat{z}}$ and Ω is called the **angular frequency** (units: s^{-1}), defined by

$$\Omega = \frac{2\pi}{T}, \tag{3.6}$$

where T is the **period** and the time it takes to complete one rotation, which is $360^\circ = 2\pi$ radians. So Ω is the rotation rate, i.e. the higher the Ω, the faster the body rotates. On Earth, $\Omega \approx 7.29 \times 10^{-5}\ \mathrm{s}^{-1}$.

One thing we will normally do is to choose a co-ordinate (or reference) that is locally vertical. What this means is that there is generally a mis-alignment of the direction between the local 'depth' co-ordinate $\boldsymbol{e}_z$ and $\boldsymbol{\Omega}$. This implies that the locally non-zero horizontal flow $\boldsymbol{u}$ will feel a different magnitude of the horizontal component of the Coriolis effect $2\boldsymbol{\Omega} \times \boldsymbol{u}$ as the latitude changes. It

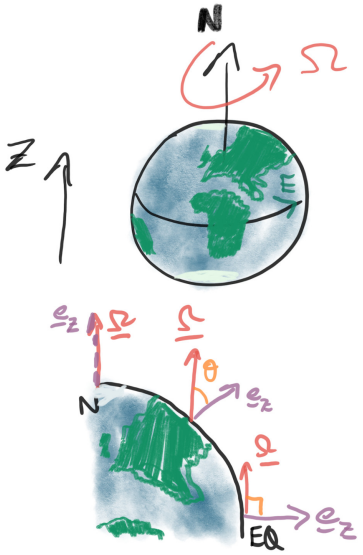


Figure 3.9: Rotation axis and angular frequency Ω. Generally speaking there is a mis-alignment of $\boldsymbol{\Omega}$ and $\boldsymbol{e}_z$ used locally for depth, giving rise to the Coriolis parameter $f = 2\Omega \sin(\text{latitude})$.

is convenient to take care of this mis-alignment by introducing the **Coriolis parameter** f where

$$f = 2\Omega \sin(\text{latitude}), \tag{3.7}$$

so the Coriolis effect in the local co-ordinate system is now given by $f\boldsymbol{e}_z \times \boldsymbol{u}$. Note then, by this definition, if we take (90S, 90N) to be $(-90\circ, +90^\circ)$ or $(-\pi/2, +\pi/2)$, then f is maximally positive at the North Pole, maximally negative at the South Pole, decreases in magnitude with latitude, and is zero at the equator.

So far we talked about rotation and have been alluding to the fact that the Coriolis effect does something to the flow, but have not really explained what it is. The way I convinced myself is to do a little demonstration with a piece of paper and a pen, and try and draw some straight lines while rotating the piece of paper underneath, as in Fig. 3.10. While I am going to describe the experiment and rationalise it, I highly recommend you try this yourself (it really really helped me).

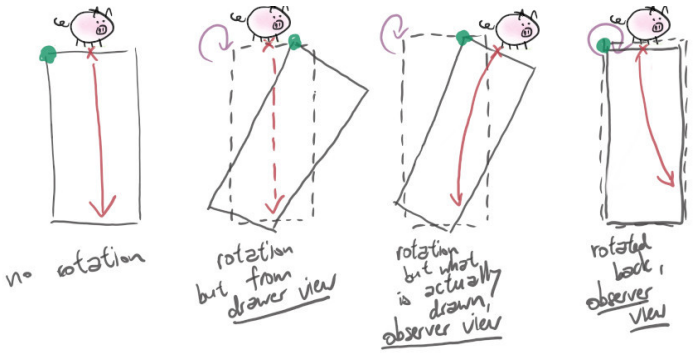


Figure 3.10: Schematic of apparent deflection from Coriolis effect. Here $\boldsymbol{\Omega}$ will be putting in or out of the page (as long as I choose my Ω accordingly).

So here you imagine there is a drawer (you), and an observer (put a plushie there if it helps). So when there is no rotation, you draw a straight line, and you see a straight line, and the observer also sees a straight line, great (1st column). Now, suppose you draw a straight line but rotate the piece of paper: from you the drawer's point of view, you are still drawing a straight line, no problem there (2nd column). However, what the observer (not you) should see is that there is now a deflection on the line drawn (3rd column). If you rotate the piece of paper to where you started from, then both you and the observer see that the line that came out is curved (4th column). The key thing here is that there is no contradiction: the action of line drawing is exactly the same, but different observers end up seeing things differently.

And now you can imagine if it was our pig demonstrator being pushed along the piece of paper (but WITHOUT being affected by

the friction of the piece of paper), and the underlying piece of paper was rotating. From our friendly neighbourhood pig's point of view, it is being pushed in a straight line, but from our point of view, the pig's path would trace out a curve. This apparently deflection is purely because we are looking at the action in a different perspective. To reconcile the two different results in the different perspectives, even though the physics should be exactly the same, we either have to add a correction to the forces in the pig's perspective, or we have to add it to our own. It turns out actually the pig's point of view is the 'right' one to take, so we add an extra (but "fake") force to correct the description we see as an observer, which turns out in this case to be the Coriolis effect $f\boldsymbol{e}_z \times \boldsymbol{u}$.

From a looking down on Earth point of view, what we normally want to do is consider the map as *fixed*; for argument sake lets choose the centre to be the zero longitude. So while the Earth is actually rotating in time, at any snapshot in time we recenter it to the zero longitude, so in effect we are rotating the piece of paper back like in the 4th column of Fig. 3.10. From this perspective, we would see the actual trajectory trace out a curve, because we are in the "wrong" (but convenient) perspective and we need the Coriolis effect to compensate us being in the "wrong" (but convenient) perspective.

What this ends up implying is that the deflection is to the *right* in the Northern Hemisphere, and to the *left* in the Southern Hemisphere; see Fig. 3.11. You can see this difference by doing the Fig. 3.10 experiment again, but rotating the piece of paper in a different direction. Also try imagining (or really try) to view the motion from *above*, so like the Northern Hemisphere point of view where you are looking down into the rotation axis, and view the motion from *below*, so like the Southern Hemisphere point of view where you are looking up but in the direction of the rotation axis. If it helps, try tracing a wet finger one of those brown basketballs or something like that while rotating it in a fixed direction, and see how the resulting water mark looks like. Another way to convince yourself that the Coriolis effect changes sign when you go from Northern to Southern Hemisphere is to put your right hand into the thumbs up position. If you are looking *into* the thumb, this is the Northern Hemisphere perspective, and your fingers are curving anti-clockwise from this point view, which is in the direction of Earth's rotation. If you are instead looking *along* the thumb with the little finger closest to your eye, this is the Southern Hemisphere perspective, then the fingers are now curving clockwise from this point of view, and effectively rotation has "reversed", hence the sign change in f.

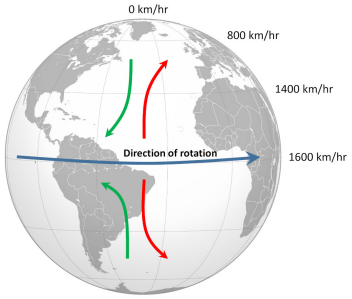


Figure 3.11: Deflection on Earth arising from Coriolis effect, from Vallis (2011).

3.2.2 *Rossby number*

From now on I'm also just going to drop the "deflection in the observer's point of view", because we will be working in this observer's point of view unless otherwise stated. First note that the Coriolis effect is not equally important across time- and length-scales. Going back to our friendly neighbourhood pig in Fig. 3.10, if rotation of the piece of paper is slow and/or the pig moves really fast, then the deflection might be negligible. If the rotation is fast and/or the pig takes its sweet time moving along the paper, then there would be significant deflection[9].

There is a competition between the *advective time-scale*, the time-scale associated with the body's motion, and the *rotation time-scale*, the time-scale associated with rotation of the system, which leads to whether the Coriolis effect is important or not. The measure is given by the **Rossby number**[10]

$$\mathrm{Ro} = \frac{U/L}{f} = \frac{1/T}{f} = \frac{\text{advective time-scale}}{\text{rotation time-scale}}. \tag{3.8}$$

Note that the Rossby number is a *non-dimensional* number. If T large so that $1/T$ small, i.e. motion on long time-scales, or f large, i.e. fast rotation, the $\mathrm{Ro} \ll 1$, Coriolis effect is important and motion is *rotationally constrained*. If $\mathrm{Ro} \approx 1$ then motion is *rotationally influenced*, and if $\mathrm{Ro} \gg 1$ then the Coriolis effect plays a minimal role in the dynamics.

As a numerical example, large-scale motion in Earth's atmosphere in the mid-latitudes (say 50°N) and $\Omega = 2\pi/\text{day}$ gives

$$\mathrm{Ro} = \frac{10\ \text{m s}^{-1}/1000\ \text{km}}{2 \times 2\pi\ \text{day}^{-1} \times \sin(50^\circ)} \approx \frac{10^1 \times 10^{-6}}{10^{-4}} = 0.1,$$

so large-scale motion in the atmosphere is rotationally constrained. For the above, just be careful with time and length units, i.e.

$$\text{day} = 3600 \times 24\ \text{s} \quad \Leftrightarrow \quad \text{day}^{-1} = (3600 \times 24)^{-1}\ \text{s}^{-1}.$$

In the ocean Ro depends somewhat on the scale of motion, since the time- and length-scale of the types of motion are usually anti-correlated (e.g. fast time-scale motion is usually small-scale motion, and vice-versa). In the Gulf Stream, we might have (for argument sake)

$$\mathrm{Ro} = \frac{1\ \text{m s}^{-1}/1000\ \text{km}}{2 \times 2\pi\ \text{day}^{-1} \times \sin(40^\circ)} \approx 0.01,$$

and the Rossby number is even smaller, because ocean flow tends to be slow, so there is ample time for the Coriolis effect to act. On the

[9] The path of the pig would actually trace out loops if the paper is allowed rotate multiple times, leading to what are called *inertial oscillations*; see Ch. 6.1.

[10] After the Swedish-American meteorologist Carl-Gustav Arvid Rossby (1898-1957), who studied under Vilheim Bjerknes. Rossby made fundamental contributions to modern day meteorology and geophysical fluid dynamics. See later also with *Rossby waves*.

other hand, if we are talking about *sub*mesoscale eddies, we may have

$$\mathrm{Ro} = \frac{0.1\ \mathrm{m\ s^{-1}}/1\ \mathrm{km}}{2 \times 2\pi\ \mathrm{day^{-1}} \times \sin(40^\circ)} \approx 1,$$

so the smaller-scale dynamics are rotationally influenced but by no means rotationally dominant. Going even smaller scales, you can imagine for example *gravity wave* motion will have large Ro, because they are fast as well as small-scale (see Ch. 6.1).

Some other astronomical examples (work these out yourselves maybe):

- Jupiter has these cloud bands associated with very fast jets that go up to $U = O(100\ \mathrm{m\ s^{-1}})$, but $\mathrm{Ro} \ll 1$, because while U is large, L is huge, and Ω on Jupiter is larger than Earth;
- Venus is similar size to Earth and can have very fast winds, but $\mathrm{Ro} \gg 1$, because Ω on Venus is tiny;
- the Solar interior has motion with large L, but Ω of the Sun is not that high, and $\mathrm{Ro} \approx 1$.

3.2.3 *Geostrophic balance*

So the flows in Fig. 3.7 and 3.8 are mostly going to be in the $\mathrm{Ro} \ll 1$ regime, but how does the Coriolis effect end up forcing the flow to be along rather than across isobars then? We recall that the momentum equation Eq. (3.1) (ignoring the forcing and dissipation) is given by (written in terms of f)

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + f \boldsymbol{e}_z \times \boldsymbol{u} = -\frac{1}{\rho_0} \nabla p.$$

The following procedure is called **dimensional analysis** or **non-dimensionalisation**: pull out all the dimensions associated with all the terms, gather everything, and see what you are left with[11]. The reason for doing this is that sometimes there might be multiple parameters in the system (e.g. Coriolis, density, flow speed, etc.), but it's not their individual values that matter, but a combination of their values. If we decide to re-scale as

$$t = Tt^*, \qquad (x, y) = L(x^*, y^*),$$

then we have, by considering the units of various terms,

$$\frac{\partial}{\partial t} \to \frac{1}{T}\frac{\partial}{\partial t^*}, \quad \nabla \to \frac{1}{L}\nabla^*, \quad \boldsymbol{u} \to U\boldsymbol{u}^* = \frac{L}{T}\boldsymbol{u}^*, \quad p \to Pp^*,$$

where we deliberately didn't do anything to f and ρ_0. Now, writing the equation in terms of the non-dimensional variables with stars and

[11] Think of it as switching units: a 100 kg pig would be 10^5 g, about 220 pounds, 0.1 tonnes, or 5.4 bajillion gazoboks (I made this one up), but it is still a pig with that amount of mass. When non-dimensionalising, you choose the measure so that the mass of the pig is '1', with the understanding that '1' means 100 kg if the choice of mass scaling is chosen to be 100 kg.

collecting factors accordingly, it may be shown that we have

$$\mathrm{Ro}\left(\frac{\partial \boldsymbol{u}^*}{\partial t^*} + \boldsymbol{u}^* \cdot \nabla^* \boldsymbol{u}^*\right) + \boldsymbol{e}_z \times \boldsymbol{u}^* = -\frac{P}{L\rho_0}\nabla^* p^*.$$

Now, in the $\mathrm{Ro} \ll 1$ regime, the evolution and nonlinear inertia term is definitely small, so we will look to throw those away. If the scale $P/(L\rho_0)$ is also really small and we throw that away, then we get nothing interesting, because it just says $\boldsymbol{u} = 0$. So what this means is that, in the $\mathrm{Ro} \ll 1$ regime, we expect the dominant *balance* between the "forces" (because Coriolis "force" is fake) in the horizontal direction is

$$f\boldsymbol{e}_z \times \boldsymbol{u}_g = -\frac{1}{\rho_0}\nabla p, \tag{3.9}$$

i.e. between the Coriolis effect and the pressure gradients. This is what is called **geostrophic balance**[12], and the flow $\boldsymbol{u}_g$ that satisfies to the geostrophic balance is called the **geostrophic flow**.

[12] For completeness, non-dimensionalisation with the vertical momentum equation is the formal way to obtain hydrostatic balance given by Eq. (3.1b), where the small parameter ends up being the aspect ratio.

A schematic of the implication of geostrophic balance in the Northern Hemisphere is given in Fig. 3.12. First we have our isobars (the red and blue dashed line), chosen so that $-\nabla p$ points up (north) and across isobars. To get 'force' balance (otherwise we have an acceleration), the 'force' arising from the Coriolis effect must be equal and opposite, so pointing down (south) and with a minus sign because it is in the opposite direction. Now, in the Northern Hemisphere, $\boldsymbol{\Omega}$ points out of the page (see the small insert of the globe to convince yourself). The questions to ask which direction is $\boldsymbol{u}_g$ pointing so that $-\boldsymbol{\Omega} \times \boldsymbol{u}_g$ is pointing down (so balancing $-\nabla p$). From the cross product discussion in Ch. 1.5.3, convince yourself $\boldsymbol{u}_g$ points right (east), then $-\boldsymbol{\Omega} \times \boldsymbol{u}_g$ is pointing south, and we have the balance we need.

The upshot is that (1) $\boldsymbol{u}_g$ is parallel to isobars, and (2) in the Northern Hemisphere, $\boldsymbol{u}_g$ goes to the *right* of $-\nabla p$. The same argument could be repeated for the Southern Hemisphere, still with $-\nabla p$ pointing north, and you find that because $\boldsymbol{\Omega}$ is now pointing into the page, the direction of $\boldsymbol{u}_g$ is swapped (i.e. to the west). The geostrophic flow is still along isobars, but now $\boldsymbol{u}_g$ is to the *left* of $-\nabla p$.

An equivalent mathematical way of doing it is to note that, taking (x, y, z) to be zonal-meridional-verical, then the schematic of Fig. 3.12 leads to Eq. (3.9) looking like

$$f\boldsymbol{e}_z \times \boldsymbol{u}_g \sim \boldsymbol{e}_y,$$

where $\boldsymbol{e}_{x,y,z}$ are the vectors pointing east, north and up respectively ($-\nabla p$ points north so is represented by $\boldsymbol{e}_y$). Then, in the Northern

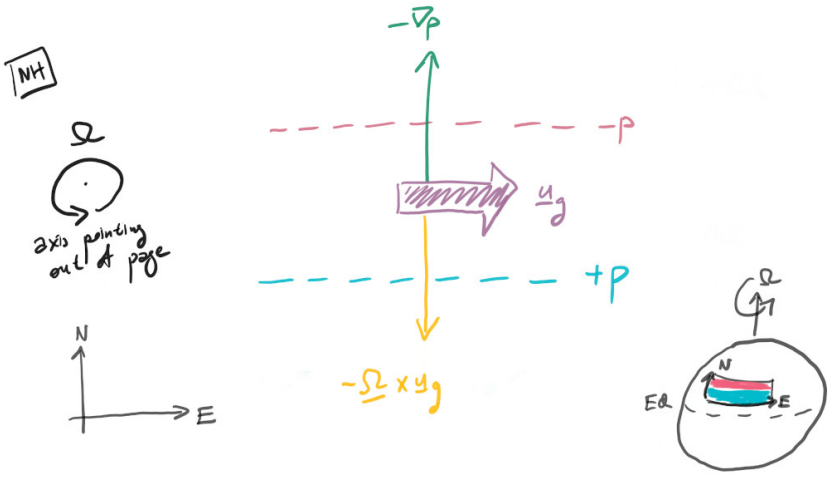


Figure 3.12: Geostrophic balance and resulting geostrophic flow $\boldsymbol{u}_g$ in Northern Hemisphere. Note $\boldsymbol{u}_g$ is **along** (rather than **across**) isobars.

Hemisphere, either by guessing, the cyclic rule associated with the cross product, or otherwise, we have to have $\boldsymbol{u}_g \sim \boldsymbol{e}_x$ since $f > 0$, i.e. pointing east, and to the *right* of $-\nabla p$. On the other hand, if $f < 0$, then by sign considerations, $\boldsymbol{u}_g \sim -\boldsymbol{e}_x$, i.e. pointing west, and to the *left* of $-\nabla p$ that is still pointing North. Try to convince yourself if $-\nabla p$ is pointing in other directions that similar arguments still hold (of course $\boldsymbol{u}_g$ will be pointing in a direction perpendicular to $-\nabla p$ depending on the hemisphere).

Note that at no point in the above discussion on geostrophic balance did we talk about the ocean or atmosphere, merely that the system is rapidly rotating in the sense that Ro $\ll 1$, and provides an explanation of the observed phenomena in Fig. 3.7 and 3.8 that the flow is largely along isobars or contours of SSH in the ocean. In fact this is one way of how we actually infer for the currents in the ocean (see Ch. 7.5.1), although we need another tool that we will visit in Ch. 5.1.3.

To close this section, we talk about how this works for *eddies* (as a noun), as in the schematic given in Fig. 3.13. Here the objective is to rationalise the flow orientation associated with these **geostrophic eddies**, taking the Northern Hemisphere case for concreteness. A geostrophic eddy with a bulge (high SSH) implies it is a high pressure in the eddy, and a low pressure outside. What this means is that $-\nabla p$ points *out* of the eddy, and since $\boldsymbol{u}_g$ points to the right of $-\nabla p$, this implies the eddy is circulating in a clockwise fashion. On the other hand, for a eddy with a depression (low SSH), the pressure is low in the core, $-\nabla p$ points *in* to the eddy, and again because $\boldsymbol{u}_g$ points to the right of the $-\nabla p$, this implies the eddy is circulating in an anti-clockwise fashion.

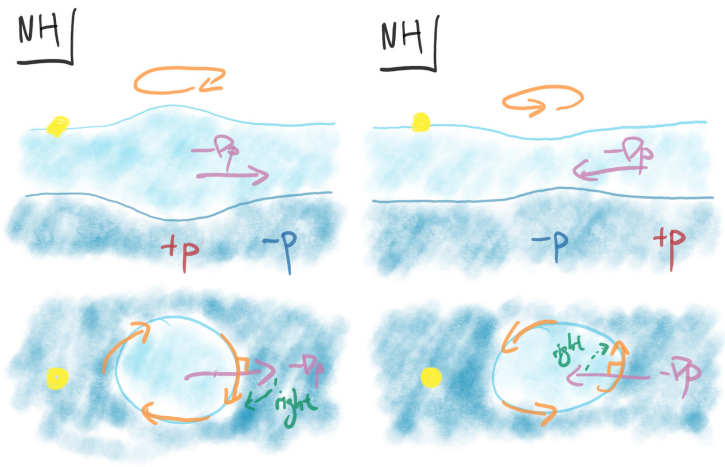


Figure 3.13: Schematic for an (left) anti-cyclonic warm core and (right) cyclonic cold core eddy, in the vertical view as well as a top-down view.

Just a little bit more terminology before we move on. In the Northern Hemisphere, since the Earth's rotation is regarded as rotating in the anti-clockwise sense (about the North Pole), something that is rotating in the same sense as the Earth is denoted **cyclonic**, and **anti-cyclonic** otherwise. In this regard, geostrophic eddies with high SSH are *anti-cyclonic* eddies, while eddies with low SSH are *cyclonic* eddies. It turns out the previous sentence holds true even in the Southern Hemisphere: high SSH, $-\nabla p$ points out, but $\boldsymbol{u}_g$ now to the *left*, so eddy circulation is anti-clockwise, but Earth is regarded to be rotating clockwise in the Southern Hemisphere, so the sense of rotation of the eddy is different to Earth, so we still have an *anti-cyclonic* eddy. The beauty is that this works also for the atmosphere: high pressures systems are **anti-cyclones**, while low pressure systems are **cyclones**, and this statement is true in both hemispheres. From a mathematical point of view, we tend to define the flow as cyclonic if the **vorticity** $\nabla \times \boldsymbol{u}$ has the same sign as f, and anti-cyclonic otherwise. The cyclonic/anti-cyclonic terminology as well as the concept of vorticity will be used with increasing frequency as we progress.

3.3 *Wind forcing*

From the previous section we noted that, in the atmosphere, high pressures systems are anti-cyclonic while low pressure systems are cyclonic. One thing you may know is that high pressure systems are *blocks*, which tend to lead to stable weather, while low pressure systems lead to unstable weather[13]. The reason is that associated with these systems are secondary up and down motion. Cyclones have associated with it upward motion in the eddy, encouraging

[13] Atmospheric *storms* are technically all cyclones.

moisture transport upwards, leading to cloud formation and thus precipitation. Anti-cyclones on the other hand lead to downward motion within the eddy, suppressing cloud formation and thus precipitation. Questions we are going to answer and explain in this section are:

- Are there analogues of up and down motion for ocean geostrophic eddies? (Yes)
- Can we replace $-\nabla p$ above with wind forcing instead? (Yes, and resulting *Ekman transport* is to the right or left of wind depending on f)
- Is there up and down motion associated with the wind forcing? (Yes, *Ekman suction* and *pumping*)

3.3.1 *Observed winds*

Lets start with what the atmopsheric observed winds actually look like first. Fig. 3.14 shows a schematic of the surface wind patterns, with the atmospheric pressure belts marked on. The observed surface winds are generally consistent with the discussion in the previous section, in that while the wind has a component in the direction of $-\nabla p$, it looks like it is mostly along isobars, with the deflection from the $-\nabla p$ in the relevant direction depending on the hemisphere. One reason the observed winds that are expected to be in geostrophic balance are not completely along isobars is because we are talking about *surface* winds, so *friction* (Ch. 3.4.3) arising from wind 'rubbing' against the ground is not entirely negligible (but turns out the effect of friction is consistent to what we see in the wind patterns). The other feature to notice is that the higher latitude winds are more along-isobars; this is consistent with $|f|$ being larger, so the Coriolis effect is relatively speaking stronger even in the presence of friction.

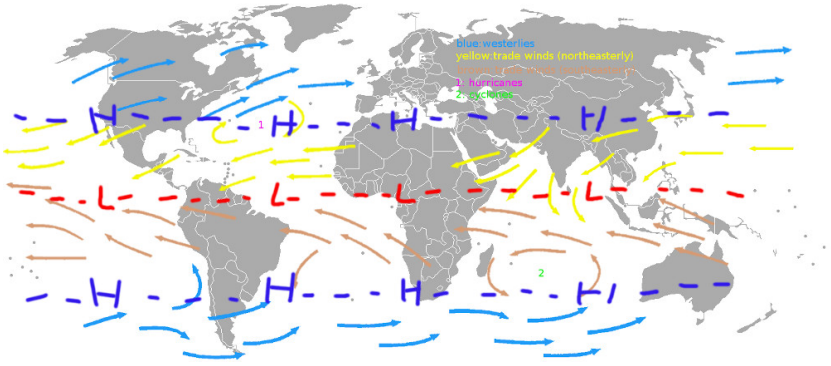


Figure 3.14: Patterns of global surface winds with pressure belts marked on. From wikimedia.org.to swap out with year-avg JRA55 data?

First of all why is there the observed atmospheric pressure patterns anyway? The pattern marked on roughly denotes the surface effects of the **Hadley cell**[14]. The Earth receives most heat around the Equator, and heating leads to air moving up vertically at the equator, hence the low pressure at the Equator. Air rises until it is neutrally buoyancy at some height (the *tropopause*), then it has to go somewhere. Since it can't go further up, down or in the zonal direction (why not?), it has to move meridionally away from the Equator. As the air moves higher in latitude, it is cooled because the surrounding temperature is cooler as the higher latitudes receives less heating, and thus it has to sink, leading to a high pressure at the surface. The return flow then manifests as the surface winds. One might have expected the circulation to extend all the way up to the poles with this thermodynamic argument, and it does if the Earth is not rotating. Being very hand-wavy, the argument is that the rotation leads to a deflection as the high altitude air moves to the higher latitudes. The distance the air travels is longer, thus allowing more time for the parcel to cool and sink back down to the surface at a lower latitude. There are more subtleties at play here that we will not need for our purposes so I don't really want to go into it... Anyway, the resulting downwelling region on Earth is around 30° N/S called the **subtropical highs**, where the surface winds are generally weak, and the weather is very stable[15].

[14] After English lawyer and meteorologist George Hadley (1685-1768). In the atmospheric literature there is also the *Ferrell cell*, but that is a can of worms I don't really want to open in this document here...see one of the digressions about *Deacon cells* in Ch. 5.1 for a related can of worms.

The main bit for our purposes particularly in Ch. 5 and Ch. 6 is the surface wind patterns. The **trade winds** refer to the wind patterns between the subtropical highs, which are generally Equator-ward with a westward component. Between the subtropic highs and to around the edge of the polar regions (around 60° N/S), the winds are **prevailing westerlies**[16], that are predominantly eastward winds. Note that there is a change in the sign of the wind *gradient* as we move up from the Equator towards to Poles; we will come back to this when we talk about the *wind stress curl* later.

[15] It is also called the *horse latitudes*, with one story being that when sailors used to take horses in their ships to travel across the Atlantic, to conserve drinking water, they tend to throw the horses abroad around this region as the ship is barely moving in this region (the winds are weak).

[16] In the atmospheric literature it is customary to call the winds by the direction they came *from* (e.g. eastward = westerlies). I will generally *not* take that convention here (I think I understand the rationale behind the terminology but I still don't like it).

For completeness, **monsoons** are seasonally varying winds that particularly affect the coastal areas of certain locations around the world (e.g. South and South-East Asia, sub-Saharan Africa, Central America), and very important for the regional climate with notable consequences for agriculture. The reasoning behind monsoons is to do with the heat capacity of seawater, with the schematic sketched out in Fig. 3.15. In the summer, the land warms up much more than the ocean because the ocean requires much more energy to heat up. This implies a $-\nabla p$ that points into the land, thus driving a wind from the ocean to the land (with appropriate deflections from the Coriolis effect), which carries a lot of moisture with it because the air is warm (so can hold more moisture) and the wind

is passing through the ocean. In the winter this situation is reversed, because the ocean holds heat much better than land, and results in a dry cold air blowing away from the land. More on the effects and consequences in Ch. 3.3.3.

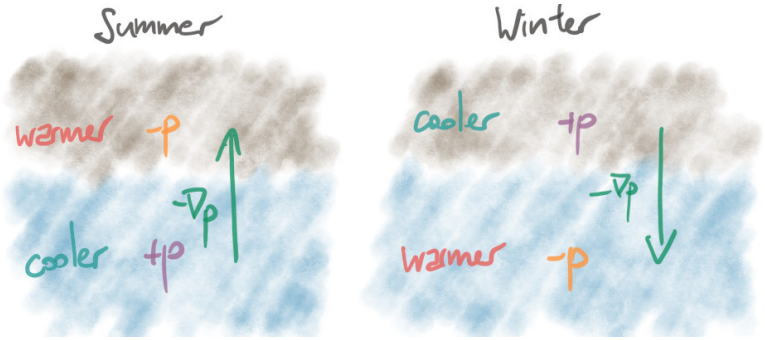


Figure 3.15: Schematic of monsoons, arising from changes in pressure gradients largely governed by heat. Actual wind direction slightly deflected because of Coriolis effect.

3.3.2 *Ekman layer, spiral and transport*

Winds forcing affects the momentum, and in fact wind forcing is the chief source of momentum into the ocean. Normally this is represented as a **wind stress** (units of N m^{-2}) in the momentum equation, so we may ask how is the momentum actually being put into the ocean; we leave this until the next section after we talked a bit about *diffusion* and *friction*. Accepting that wind does put momentum in via exerting a force over the top of the ocean, then we might expect the fluid's motion to be along the direction of the force. But we also know that by geostrophic balance the geostrophic flow should end up being perpendicular to the direction of intended motion, i.e. to the right or left of the direction of wind stress depending on hemisphere. So how do we reconcile this?

The answer is actually quite simple: the flow turns! The wind's influence has to have some limited vertical extent (e.g. Fig. 3.16) because the ocean is a fluid and not a solid, so layers of fluid can slide over each other easily, unlike in solids where there will be strong resistance from the material resisting the imposed force (see more in Ch. 3.4). Within the thin region near the ocean surface called the **Ekman layer**[17], geostrophic balance needs to be modified because forcing is not negligible.

[17] We give a better definition of the Ekman layer in Ch. 3.4.2.

What happens then can be seen in Fig. 3.17, and we take Northern Hemisphere for concreteness. Near the surface, the intended flow is in the direction of wind, but because of the deflection from the Coriolis effect the actual flow is at an angle to the intended flow (but not perpendicular to it; think adding vectors associated with the

Figure 3.16: Schematic of Ekman layer (boundary denoted by orange). The stuff underneath the Ekman layer could be regarded as being shielded from the direction influence of the wind.

wind stress and Coriolis effect together). However, as we gradually move deeper, the influence of wind weakens, the flow weakens, but we get closer to geostrophic balance since wind forcing is weaker, so the flow has to turn, in this case to the right of the wind. What results is an **Ekman spiral**[18].

The **Ekman transport** is the net transport of the resulting flow, i.e. the depth integral of the flow. While there is some component of the flow near the surface roughly in the direction of the wind, the Ekman transport is essentially perpendicular to the direction to the flow (to the right in Northern Hemisphere). While the flow is largest near the surface, it occupies a relatively small volume, so when it is integrated it is the small flow but over larger volume that dominates the final result.

[18] After the Swedish oceanographer Vagn Walfrid Ekman (1874-1954), whose theory was motivated by observations of iceberg motion not being in the direction of the wind after Vilheim Bjerknes got him onto the problem. Under simplifying assumptions one could actually get an analytical form of the Ekman spiral (as Ekman originally did in his doctoral thesis).

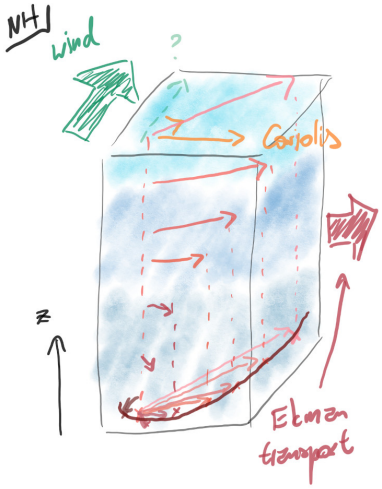


Figure 3.17: Schematic of Ekman spiral over the Ekman layer.

3.3.3 *Ekman pumping and suction*

Ekman dynamics also drive a vertical flow, which has important consequences for coastal dynamics and biogeochemistry. We illustrate two examples first, which may be useful references to keep in mind before we talk about the general cases of fluid *divergence/convergence* and relating the phenomenon with the *wind stress curl*.

In Fig. 3.18 we assume again we are in the Northern Hemisphere, and we have a surface wind blowing south along a coastline. According to the discussion above, the Ekman transport is to the right and so is off-shore. If we are looking at this from a two-dimensional point of view we immediately have a problem, because we can't continually have an off-short flow as we will run out of water to transport at some point. To maintain *mass conservation*, what we must have is water moving on-shore from the deep parts of the water column to replace the surface water that is being moved off-shore. This in turn implies there has to be an upward movement of water arising from the action of the wind, and is called **Ekman suction**.

Another example is at the Equator where the Coriolis parameter changes sign, depicted in Fig. 3.19. The trade winds converge at the

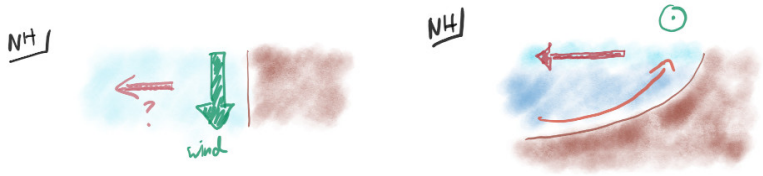


Figure 3.18: Schematic of Ekman suction near the coast. Wind is pointing out of the page.

Equator, and the corresponding Ekman transport leads to a flow *divergence* ($\nabla \cdot \boldsymbol{u}_g > 0$). Again, for mass conservation, this necessarily implies there is a circulation with fluid *convergence* ($\nabla \cdot \boldsymbol{u}_g < 0$) deeper down to replace the surface water diverging, and in turn implies an *upwelling* around the Equatorial region.

Figure 3.19: Schematic of Ekman suction around the Equator. Wind is pointing into the page.

The converse cases where the winds are reversed in the above will lead to a *downwelling*, i.e. **Ekman pumping**. A surface divergence of water leads to Ekman upwelling, while surface convergence of water leads to Ekman downwelling, so we want to know what kind of wind causes divergence and convergence in the flow in the general cases. Lets assume we are away from coasts, we notice that if we have a spatially uniform wind stress then there is no convergence or divergence, because the Ekman transport will be spatially uniform. Therefore we need a *shear* in the wind stress, i.e. a non-zero spatial gradient in the wind stress to get Ekman up or downwelling. Fig. 3.20 shows a schematic of the cases where the meridional shear in the wind stress is negative and positive respectively, assuming we are in the Northern Hemisphere. When the wind stress shear is negative, the wind is forcing harder at the south than at the north. This implies that there is a stronger southward Ekman transport at the south than at the north, so there is a flow divergence, and hence an upwelling. Conversely, for a wind stress shear that is positive, the wind is forcing harder at the north than at the south, the southward Ekman transport is stronger at the North, implying there is a piling on of water to the south, i.e. a flow

convergence, and therefore a Ekman downwelling. In symbols, we take the wind stress to be $\boldsymbol{\tau} = (\tau^x(y), 0)$, then we have for the first case that

$$\frac{\partial \tau^x}{\partial y} < 0 \quad \Rightarrow \quad \frac{\partial v_g}{\partial y} > 0, \tag{3.10}$$

and inequality signs are swapped for the second case.

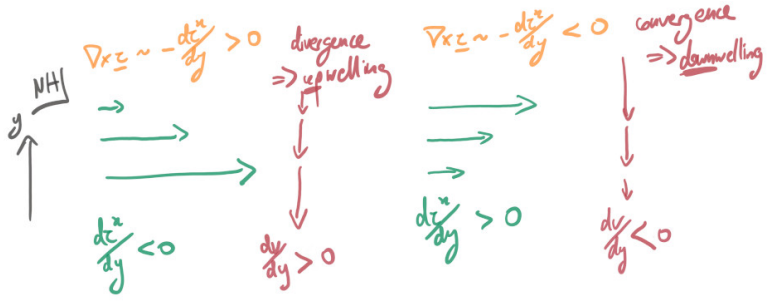


Figure 3.20: Schematic of wind shear (wind stress curl) with Ekman up/downwelling.

If we define the **wind stress curl** as $\nabla \times \boldsymbol{\tau}$, then it can be shown that the relevant component for the horizontal dynamics is the vertical component of the wind stress curl which is given by

$$\boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{\tau}) = \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y}, \tag{3.11}$$

where $\boldsymbol{\tau} = (\tau^x, \tau^y, \tau^z)$. Then we see that the above case in Fig. 3.20 has $\boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{\tau}) = -\partial \tau^x / \partial y$, so that wind stress shear and wind stress curl has a minus sign difference, implying that

$$\boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{\tau}) > 0 \quad \Rightarrow \quad \frac{\partial v_g}{\partial y} > 0, \tag{3.12}$$

and vice-versa. So the divergence of the geostrophic flow has the same sign as the wind stress curl, and hence positive wind stress curl implies an Ekman *up*welling, while negative wind stress curl implies an Ekman *down*welling, at least in the Northern Hemisphere. You can convince yourself that, in the Southern Hemisphere, positive wind stress curl implies an Ekman *down*welling (because the Ekman transport is in the other direction) and vice-versa, which implies a factor of f has to be in play. It turns out one can derive from the equations directly that

$$w_e = \frac{1}{\rho_0 f} \boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{\tau}), \tag{3.13}$$

where w_e denotes the Ekman vertical velocity (upwelling in $w_e > 0$ and vice-versa), and the verbal arguments above are consistent with (but is not a proof of) Eq. (3.13).

In Fig. 3.21 we revisit geostrophic eddies. Recalling that the vorticity of the geostrophic flow is defined to be $\nabla \times \boldsymbol{u}_g$ and we are in the Northern Hemisphere, the anti-cyclonic eddy (the one with the high pressure) is rotating clockwise, so the flow has negative[19] vorticity. If we use Eq. (3.13) but with $\boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{\tau})$ replaced by $\zeta = \boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{u}_g)$, then since $f > 0$, $w_e < 0$, i.e. an associated geostrophic downwelling. Similarly, a cyclonic eddy has $\zeta > 0$ in the Northern Hemisphere, so there is an associate geostrophic upwelling. In the Southern Hemisphere, an anti-cyclonic and cyclonic eddy *still* has an associated geostrophic downwelling and upwelling: the signs of ζ changes when we go into the Southern Hemisphere, but so does f, and there is a sign cancelling out.

[19] Traditionally in maths vorticity is positive if it is anti-clockwise. The best way to remember is either as a right hand in the thumbs up position (so fingers curl round in an anti-clockwise sense), or that angles are measure in an anti-clockwise fashion.

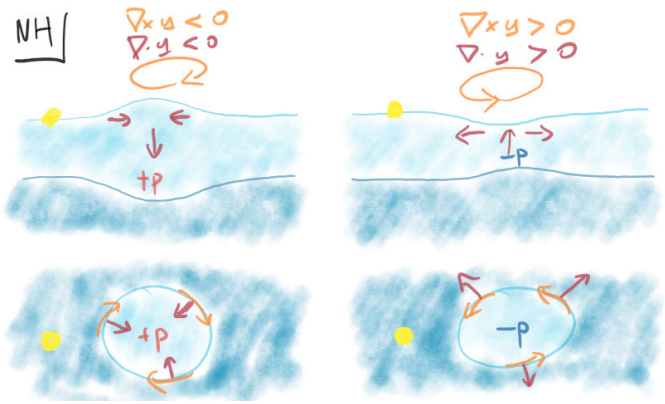


Figure 3.21: Up/downwelling associated with anti-cyclonic (left) and cyclonic (right) eddies (since we are in NH).

There are several ways to rationalise physically why cyclonic and anti-cyclonic eddies (in both ocean and atmosphere) are associated with upwelling and downwelling within the eddy. One way utilises friction at the bottom of the eddy (as 'rubbing against a medium with no motion', e.g. the ground, or a layer of fluid of no motion); see one of the chapter exercises.

some words on Ekman upwelling, Eastern Boudnary currents and bgc

3.4 *Diffusion, viscosity and friction*

We have been talking a bit about *friction*, and before focusing a bit more on friction I want to talk about *diffusion* first, which I think is a more fundamental concept.

3.4.1 *Diffusion example: milk in coffee*

Consider adding a bit of milk into say a cup of coffee as in Fig. 3.22. If you don't *stir* the coffee, then you realise it actually takes ages for the coffee to lighten in colour, because the milk just sits there minding its own business, albeit *spreading* but very slowly. On the other hand, stirring the coffee moves the coffee around, which carries the milk around, and you find the rate of spreading is substantially faster.

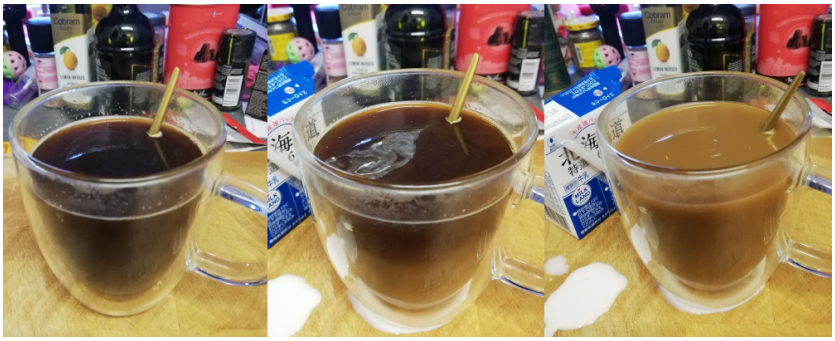


Figure 3.22: Making a mess of coffee + milk...

So here **diffusion** is just going to be referred to as the action that leads to spreading of 'stuff'. The actions above are to be distinguished as *molecular diffusion* and *effective diffusion* respectively, which we say more in due course. Diffusion results in erasing of *gradients*. When you add the milk into the coffee there is a gradient in the milk concentration, and molecular diffusion acts to spread/erase the concentration gradients, albeit really slowly. When you stir it, you involve dynamics that increases the rate of gradient removal, and concoction becomes *well-mixed* sooner, i.e. there are essentially no discernible gradients anymore. There is of course nothing special about milk and coffee, and the arguments work similarly for tracers within a fluid (e.g. chemical concentrations, momentum, heat, salt etc.[20]).

Lets start with **molecular diffusion**. Imagine there are a bunch of particles in a box, so like Fig. 3.23, but imagine even larger numbers (I didn't want to draw too many dots), and we tag the particles by colours. Suppose you start it off so that there is a gradient in the colours, and assume nothing goes out of the box. Particles randomly move around through **Brownian motion**[21], the particles are jiggling about but mixing away from the purely separated configuration. If you leave it long enough (and it can take very long...), and take a look at the box again, what you expect to see is that the colours should be fairly mixed up. Now, this *microscopic* phenomenon has a *macroscopic* effect. If we for example consider one colour as -1 and

[20] Although we may want to distinguish *passive* and *active* tracers, where the former basically moves around the with the flow, but the latter can evolve according to its own equations (e.g. chemicals) or feedback onto the flow (e.g. momentum).

[21] After the Scottish botanist Robert Brown (1773-1858). Describing Brownian motion is actually also one of Einstein's major contributions to science, even if he is mostly remembered for $E = mc^2$.

the other as $+1$, and work out the sum of per horizontal section, initially we have something with a large vertical gradient, but as time goes on and the particles do their dance, the macroscopic gradient gets gradually erased.

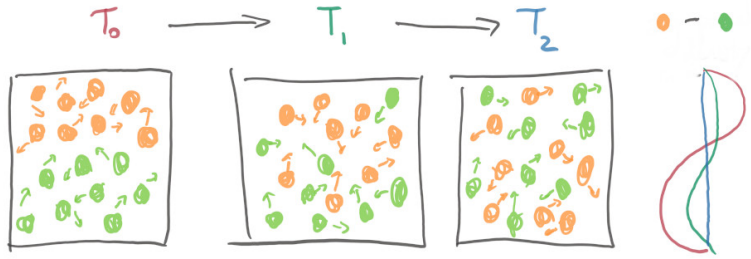


Figure 3.23: Schematic of *microscopic* motion leading to *macroscopic* diffusion. Here it is actual particle types, but can also imagine particles carrying "stuff", bumping into each other and transferring "stuff", and eventually the distribution of "stuff" evens out (but the total is conserved).

The thing I just want to clarify here is that this is not saying there can be no *un*mixing: it is just extremely unlikely. One analogy might be to think about the consequence of the random motion as how many different ways are there to rearrange the different types of particles in this box. In Fig. 3.23 there are eight particles of each colour, and suppose I start it off as a regular array and restrict the action so that the particles can only be swapped around. There are precisely $16!/(8! \times 8!) = 12870$ arrangements, accounting for redundancies and repeated patterns[22]. Now, what you can imagine is that, after you leave the particles doing its thing long enough, they keep swapping positions. When you take a peek into the box, you are sampling one of the large number of possible configurations. But since there is only essentially two configurations where the colours are all on one side, but so many more where the configuration 'looks' mixed, from a probabilistic point of view you are probably going to encounter once of those 'mixed' states. In reality, instead of 16 particles you are going to have about 6×10^{23} particles per mole of 'stuff'[23], and the number of configurations become astronomical! Yet you still only have two configurations where the colours are completely separated, compare to a ridiculous amount of states where it is 'mixed'. So it isn't things cannot *unmix*, it is just extremely unlikely to observe it, so for all intents and purposes we consider this kind of diffusive action *irreversible*.

[22] $n! = n \times (n-1) \times (n-2) \ldots \times 1$ is the *factorial*, and it gets big very quickly.

[23] Look up the *Avogadro constant*.

Staying with molecular diffusion and taking temperature T for concreteness, the mixing rate depends on the magnitude of the tracer gradients ∇T, as well as a (molecular) **diffusivity**, normally denoted κ ('kappa', units of $\mathrm{m}^2\ \mathrm{s}^{-1}$). This definition is analogous for salt, and we distinguish the diffusivities by κ_T and κ_S respectively[24] Without going into too much detail, for the fluid itself there are internal stresses that lead to spreading of momentum, but taking various

[24] Different tracers need not have the same diffusivities.

material	κ_T	κ_S	ν
seawater	10^{-7}	10^{-9}	10^{-6}
air	10^{-5}	—	10^{-5}
honey	10^{-6}	—	10^{-2}
lava	10^{-7} (!)	—	depends, 10^{0}?
steel	10^{1}	—	big (!?)

Table 3.1: Table of molecular diffusivity/viscosity values at some control conditions. All numerical entries have units of $m^2\ s^{-1}$.

simplifying assumptions this spreading can also be represented as a diffusion, except the diffusivity has a different name and is called the **viscosity**, denoted ν or $\mu = \rho\nu$ (I will use ν, 'nu'). Viscosity measures how *sticky* something is: cement has very high viscosity, but air and sea water not so much. Diffusion tends to show up in equation form as

$$\kappa_T \nabla^2 T, \qquad \nu \nabla^2 \boldsymbol{u}, \tag{3.14}$$

where $\nabla^2 = \nabla \cdot \nabla = (\partial^2/\partial x^2 + \partial^2/\partial y^2)$ is the *Laplacian operator*, which is a scalar operator consisting of second derivatives (and includes $\partial^2/\partial z^2$ if we are dealing with three-dimensions). If the diffusivity is spatially varying or more diffuses in different directions different, then they may be written as

$$\nabla \cdot (\kappa \nabla T), \qquad \nabla \cdot (\mathbf{K} \nabla T), \tag{3.15}$$

where $\mathbf{K}$ denotes a *tensor* (not going to say more about that).

The molecular diffusivities or viscosities of various tracers are material dependent and *we know what they are* (e.g. from lab experiments or by inference; see example later). Table 3.1 shows some representative values taken from various places. The exact values tend to depend on temperature and pressure, and only order of magnitudes have been given. So the diffusivities and viscosities for seawater and air are small. The viscosity of honey can actually be measured in the lab, and is lower when it is warm. The viscosities for lava is inferred for by measuring the flow rate (see later for a seawater example). Steel is technically a solid but of course if we are to be awkward, we could argue that on long time-scales everything flows in some way, so we could in principle assign a viscosity[25].

[25] As they say in *rheoloical sciences*, "*everything flows*".

One concept we use to highlight why molecular diffusivity is usually not the relevant one we care about when we are dealing with large-scale dynamics is the **diffusion time**. Since the diffusivity has units of $m^2\ s^{-1}$, we can define a time-scale as

$$t = \frac{L^2}{\kappa}, \tag{3.16}$$

where we interpret t as the time it takes for 'stuff' to diffuse a length L if the diffusivity was κ. This is one way to infer for diffusivity of

lava for example: if you know how long the lava took to travel some distance, and removing say the contributions due to other forces[26] such as gravity, then you can get back out a diffusivity.

[26] This is actually much harder than I am making it out to be...

For our purposes, lets say we are interested in knowing how long it takes for some heat (as measured by temperature) to diffuse vertically by about 100 m (to roughly below the mixed layer say). Then, if we were to use the molecular diffusivity κ_T value in Table 3.1, then

$$t_m = \frac{100^2}{10^{-7}} = 10^{11} \text{ s} \approx 3000 \text{ years!}$$

There are various reasons why we know the associated time-scale is substantially faster than the 3000 years we just computed (e.g. chemical tracers, pollutants, etc.), so there is something amiss here. The reason is of course we are not taking into account the effect of *stirring* at all when we are dealing with molecular diffusion. Just like our milk can mix and spread much faster when stirring is involved as in Fig. 3.22, the ocean dynamics is generally seen to lead to an increase in an **effective diffusivity** κ_e (sometimes *turbulent diffusivity* or *eddy diffusivity*). In general $\kappa_e \gg \kappa_m$, where κ_m is the molecular diffusivity, and thus results in $t_e \ll t_m$, i.e. the effective diffusion time is much smaller than the molecular diffusion time. If we take $\kappa_e = 10^{-2}$ m^2 s^{-1}, then repeating the calculation above leads to

$$t_m = \frac{100^2}{10^{-2}} = 10^6 \text{ s} \approx 10 \text{ days},$$

which is substantially faster and probably closer to what we might expect.

The main problem and, really, one of the perpetual problems not just in physical oceanography but generally in fluid dynamics, is that we have no idea what this effective diffusivity should be! We can get some evidence from measurements (Ch. 7) or make some educated guesses with approximated theories, but fundamentally we don't really know what it should be from first principles. The effective diffusion is entirely dynamics dependent because of the stirring aspect, and it is context dependent[27]. In the open ocean, it is generally accepted that, away from *boundary layers* (see Ch. 3.4.2) we may have

[27] The dynamical stirring does not necessarily have to be completely diffusive either! Not opening this giant can of worms here...

$$\kappa_{e,z} = 10^{-4} \text{ to } 10^{-5} \text{ m}^2 \text{ s}^{-1}, \qquad \kappa_{e,h} = 10^{-1} \text{ to } 10^{3} \text{ m}^2 \text{ s}^{-1},$$

where the effective vertical (or diapycnal) diffusivity $\kappa_{e,z}$ is much smaller than the horizontal (or along-isopycnal, or neutral) diffusivity because the ocean is density stratified, so motion in one direction is severely inhibited. Again, these depend on dynamics. Additionally, why this is saying is that it is actually very difficult to move water

up and down, so there are questions as to how the MOC functions. While we may know how water goes down, how does it come back up again? The *abyssal upwelling* problem is briefly touched up on now when we talk a bit about *boundary layers*, and in more detail in Ch. 5.2.4.

3.4.2 *Non-dimensional numbers and boundary layers*

Just like the Rossby number (Ch. 3.2.2) defines the relative importance of the Coriolis effect as the ratio of dynamical to rotational time-scales, we can similarly define non-dimensional numbers to measure the importance of diffusion and viscosity relative to the dynamics. Doing analogous non-dimensionlisations but for the tracer equations Eq. (2.1), the **Péclet number**[28] is defined as

$$\mathrm{Pe} = \frac{UL}{\kappa} = \frac{\text{advective transport}}{\text{diffusive transport}}, \tag{3.17}$$

[28] After the French physicist Jean Claude Eugène Péclet (1793-1857). He was Coriolis's brother-in-law.

so if $\mathrm{Pe} \ll 1$, diffusion is important at the length and velocity scales concerned. The equivalent to the Péclet number for viscosity has its own name, called the **Reynolds number**[29], and is defined as

$$\mathrm{Re} = \frac{UL}{\nu}. \tag{3.18}$$

[29] After the British fluid dynamicist Osborne Reynolds (1842-1912), who made some of the first systematic studies of laminar to turbulent flow transitions.

Again, if $\mathrm{Re} \ll 1$ then viscosity is important at the length and velocity scales concerned. The expectation is that high Re flow is *turbulent*; if $\mathrm{Re} \gg 1$, the nonlinear inertial terms in the momentum equation Eq. (3.1) are not negligible, and nonlinear effects are important.

When rotation is involved then one can ask about the relative importance of rotation versus viscous effects. This ratio is defined by the **Ekman number**, given by

$$\mathrm{Ek} = \frac{\nu}{fL^2} = \frac{(U/L)/f}{UL/\nu} = \frac{\mathrm{Ro}}{\mathrm{Re}}, \tag{3.19}$$

where $\mathrm{Ek} \gg 1$ means viscous effects dominate. The Ekman number is normally a more fundamental property in rotating fluid dynamics that is used often in geophysics and planetary sciences.

Another way to think about the above numbers is that, given for example ν, this implies a length-scale L_δ below which diffusive effects start dominating over dynamics, i.e. where Re, Pe or Ek ≈ 1. For large-scale dynamics, generally we have $\mathrm{Pe} \gg 1$, $\mathrm{Re} \gg 1$, and $\mathrm{Ek} \ll 1$, i.e. viscous effects do not dominate[30]. However, there are locations where the effective diffusivities are higher, and usually these occur near *boundaries* in a thin region, called a **boundary layer**. A way to define the boundary layer is to denote the regions near the boundary where Re, Pe or Ek ≈ 1. The boundary layer extent is then given by the appropriate L_δ.

[30] From a computational point of view this is actually a bad thing, because it means to really represent all the dynamics permitted by the choice of diffusivities and viscosity, we need to a spatial (and temporal) resolution fine enough to resolve down to roughly where these numbers are $O(1)$.

A schematic of the ocean bottom boundary layer is given in Fig. 3.24. Here, if we regard the ground as stationary while the ocean is moving over it, then the flow of the ocean has to go to zero as we approach the bathymetry. Away from the boundary layer, viscous effects are unimportant and the flow does whatever it is doing. As we get into the boundary layer, viscous effects start dominating, leading a stronger diffusion, arising mostly because of the dynamics going on in the region, causing the flow to decrease. From the point of view of the ocean, the ocean is experiencing *friction* provided by the ground, and is losing momentum into the ground. Depending on the bathymetric features, the effective boundary layer may have a larger vertical extent: the ocean may see the ground as more 'rough', and the corresponding friction might be larger. The bottom boundary in this instance is a region of larger diffusivity (and, in particular, *diapycnal diffusivity*) and a place where momentum is lost; we revisit these two important points in Ch. 5.

Figure 3.24: Momentum diffusion and friction. Friction arises because there is relative motion, so it might look like there is momentum loss if we are looking only at the ocean as it's own system. But really what is happening is that the ocean is transferring momentum into the land, and from the ocean + earth system point of view there is no momentum lost.

A similar argument can be made for the Ekman layer (Ch. 3.3.2) and shown schematically in Fig. 3.25. The atmosphere tends to move much faster than the ocean, but there is a region around the atmosphere-ocean interface where diffusive effects are important. The atmospheric flow from this point of view sees the ocean as a drag but, conversely, the ocean sees the atmosphere as a source of momentum.

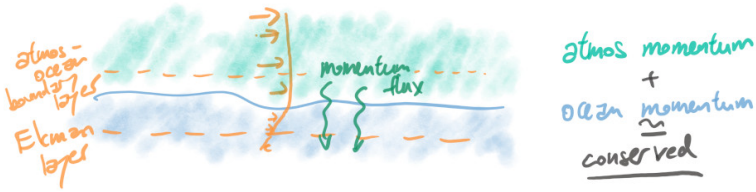


Figure 3.25: Same as above but for atmosphere and land. Atmosphere acts as a **source** of momentum for ocean most of the time (equivalently, the ocean acts as a **sink** of momentum for the atmosphere).

Noting that

$$\mathrm{Ek} = \frac{\nu}{fL^2} \quad \Rightarrow \quad L_{\mathrm{Ek}} = \sqrt{\frac{\nu_e}{f}}, \tag{3.20}$$

if we take $f \sim 10^{-4}\ \mathrm{s}^{-1}$ and $\nu_{e,z} = 10^{-2}\ \mathrm{m}^2\ \mathrm{s}^{-1}$, $L_{\mathrm{Ek}} = 10$ m.

However, note that $\nu_{e,z}$ itself maybe a function of depth, so the Ekman layer depth is only an estimate.

3.4.3 *Friction vs. diffusion*

Friction is the resistance to relative motion, so we only have friction if things are moving; contrast this with diffusion where things will spread as long as there is a gradient. Our favourite pig friend in Fig. 1.22 has *ground friction* and *air drag* acting against it when it is in motion, because it is moving relative to the ground and air (which doesn't want to move). Friction can also manifest as internal stresses against relative motion within a body itself. If you ever tried picking a cat up, you will probably get the result in Fig. 3.26. The cat clearly does not want to move, but there is a force acting on it, so the middle part moves up, but since it is connected to the other parts that are not moving, there is an internal stress that is trying to resist the motion. These internal stresses occur generally in solids and fluids, and in fluids the internal stresses manifest as a viscosity for example[31].

[31] As we know cats are liquid, so maybe we should assign viscosities to them.

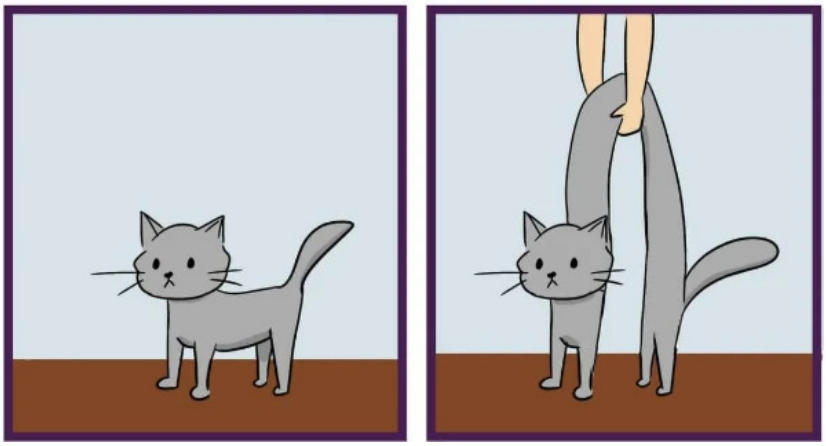


Figure 3.26: Cat physics. Picture from Meowingtons.

There are multiple ways to represent friction. Two examples are

$$-r\boldsymbol{u}, \qquad -C_d|\boldsymbol{u}|\boldsymbol{u}, \tag{3.21}$$

which are called **linear drag** and **quadratic drag** respectively, with some coefficients r and C_d (these have units but omitting them here). These function certainly are zero when there $\boldsymbol{u} = 0$, and act in the opposite direction to $\boldsymbol{u}$. The main point however I want to highlight here is that these functions remove velocity differences (because they are larger for larger flows), which affects momentum (because momentum is mass times velocity). Viscosity (or momentum diffusion) on the other hand does not remove momentum, but redistributes in, i.e. drag can act as a sink of momentum but diffusion cannot.

But you may also ask that, since Newton's laws effectively says we should have momentum conservation, then does drag violate conservation given it removes momentum? The subtlety here is that momentum is conserved in a *closed* system, and the ocean by itself is not a closed system. Consider the cases in Fig. 3.24 and 3.25, where the friction is going to be over the boundary layer. While from the ocean's point of view the ground provides a drag and thus is a **sink** of ocean momentum, from the ground's point of view the ocean is acting as a **source** of momentum. Friction removes relative motion but in such a way that the combined system momentum is conserved. Similarly for the atmosphere and ocean, the atmosphere sees the ocean as a sink, but the ocean sees the atmosphere as a source of momentum. From that point of view there is no contradiction. When we model the ocean however we usually don't model the solid Earth at the same time[32], so we have what looks like a momentum loss out of the ocean, but it's because we don't explicitly account for where it is going.

[32] E.g. when talking about sea level the ocean loading onto land and the land rebounds make a difference, so you do need some land physics there.

Summary and further reading

Changes in density manifest in changes of pressure mostly via hydrostatic balance, and negative pressure gradients are forces and drives flows via its contribution to the momentum equation. A slight complication arises because the Earth is rotating, and we make the standard and convenient decision to view the dynamics in a practical perspective, but one such that we need to include a deflection by the Coriolis effect. Again, the Coriolis effect only arises because of a difference in perspective.

Accepting this slight complication, we argued if the flow is in geostrophic balance (where $\mathrm{Ro} \ll 1$), then the geostrophic flow actually travels to the right of the direction of intended travel in the Northern Hemisphere, and to the left in the Southern Hemisphere (because the Coriolis parameter f changes sign). The same argument works mostly for wind forcing: although we have an Ekman spiral, the Ekman transport is largely in the direction of the geostrophic flow. The Ekman upwelling and downwelling was argued to be related to the wind stress curl and f, and the corresponding geostrophic upwelling and downwelling was argued to be related to the fluid vorticity and f. The concept of diffusion and friction was introduced, highlighting the differences between the know but too small molecular diffusion, and the larger but unknown effective diffusivity that depends crucially on dynamics. While diffusion redistributes momentum, friction removes momentum if we are considering the ocean as its own system.

One of the main things to note is that the boundary layers are regions where most the momentum transfers happen between the ocean and other Earth system components, these regions may also be regarded as disproportionately important regions[33]. The difficulty that we don't touch on here is that, because the boundary layers are relatively shallow, the dynamics are very hard to understand in these regions! There are many tools and tricks that simply don't work in these regions (e.g. geostrophic balance). Lots of small-scale dynamics are at also play, with nonlinear effects and feedbacks all over the place. To understand the boundary layer behaviours and the dependence of eddy diffusivities on dynamics, we need to understand the contributing small-scale dynamics, some of which are introduced somewhat in Ch. 6.

[33] This is in line with the mathematical observation that if you change the boundary condition you are probably going to change everything.

Chapter exercises

1. From the example surrounding Fig. 3.2, doing the calculations and not dropping decimal places everywhere like I did, show that $g \approx 9.81$ m s^{-2}.

2. From the example surrounding Fig. 3.2, work out g_{moon}.

3. From the example surrounding Fig. 3.2, without working out g_{pig}, do you think this is large or small? Taking $m_{\text{pig}} = 100$ kg, work out g_{pig}.

4. Convince yourself (pictorially or mathematically) that if $\boldsymbol{g} = -\nabla\phi$, then $\boldsymbol{g}$ is perpendicular to the surfaces of $\phi =$ constant.

5. If you are at sea level you experience a pressure of 1000 mb, or 1 atmospheric pressure (1 atm), i.e. there is one Earth atmosphere's worth of pressure pushing down on you. Work out via hydrostatic balance the pressure you experience just from having 10 m of sea water above you (with no atmosphere above you), assuming for simplicity that $\rho_{\text{sea}} = 1000$ kg m^{-3} and $g = 10$ m s^{-2}. What is the pressure in units of mb if you are 1000 m below the ocean (with no atmosphere above you)?

6. Using the above calculation and assumptions, calculate the largest hydrostatic pressure (in atm) due to just sea water experienced by Captain Nemo's *Nautilus* in Jules Verne's novel. (Wikipedia will help you here; the answer is not $20000 \times 4 \times 1000/10 = 8 \times 10^6$ atm!)

7. Demonstrate mathematically that the Coriolis effect does no work.

8. In the general form the Coriolis effect is given by the term $-\boldsymbol{\Omega} \times \boldsymbol{u}$. Using the properties of the cross product, without calculations (so draw pictures or reason it out), rationalise that, for a horizontal flow parallel/tangent to the Earth's surface at the Equator, there is no horizontal Coriolis effect acting on this flow (hence why $f = 0$ at the equator).

9. Show the above mathematically (probably want to use cylindrical co-ordinates for this).

10. What does the Coriolis parameter f look like for a cylindrical Earth (with rotation axis pointing out of the circle of the cylinder)?

11. Try the Fig. 3.10 experiment but draw the line really slowly, and allow the piece of paper to rotate multiple times, and convince yourself you do get looping motion (cf. *inertial oscillations*).

12. Look up some numbers to work out the Rossby numbers of Jupiter, the Solar interior[34] and Venus. They should be tiny, about one, and large respectively.

13. Comment on whether there is any scientific merit to the claim that "*when you flush the toilet in the Southern Hemisphere, the water flows the other way compared to the Northern Hemisphere because the Coriolis effect is of a different sign*". Back up your claims, and provide some estimates accordingly.

14. Show mathematically that, for a flow in geostrophic balance, i.e. $f\boldsymbol{e}_z \times \boldsymbol{u}_g = -\rho_0^{-1}\nabla p$, $\boldsymbol{u}_g$ is always perpendicular to $-\nabla p$.

15. Normally we hear about storms forming in the ocean and then hitting the south-eastern North America and north-western Europe, but not north-western parts of North America or south-western Europe, why is that?

16. What happens to geostrophic balance when friction is involved? Extend the pictorial arguments in Fig. 3.12 and, assuming friction acts against the direction of $\boldsymbol{u}_g$, show that resulting angle between $\boldsymbol{u}_g$ and $-\nabla p$ has to be strictly less than 90° ($\pi/2$), i.e. the resulting geostrophic flow is tilted into the direction of $-\nabla p$. From this deduce that anti-cyclonic eddies have associated with it divergent flow at the bottom of the eddy, implying a downwelling, and cyclonic eddies have associated with convergent flow, implying an upwelling[35].

[34] Try looking for *Solar tachochline* maybe

[35] Note at no point here have I said anything about the ocean or the atmosphere.

4 Gyre circulation and western intensification

"All models are wrong, but some are useful.'

– attributed to George Box

some EBUS stuff here?

Here we give a first example on how to put a good portion of what we have gone through so far to provide *a*[1] dynamical explanation for gyres and Western Boundary Currents. The focus will be on the classical theory focusing on the role of wind forcing, but see the end of chapter for some extensions and complementary descriptions. The wind driven theory to me is a particularly good example of a good simple *model*: you learn stuff because the model works, but you *also* learn stuff because the model *doesn't* work 100%[2].

[1] Note the use of *a* rather than *the*.

[2] It is also one of the ones that I can go through essentially by drawing pictures, hence it's here.

4.1 Recaps and spoilers for wind driven gyre theory

Recall from Ch. 1 that we have the anti-cyclonic subtropical gyres and cyclonic subpolar gyres in both hemispheres, and associated with these gyres are Western Boundary Currents (WBCs) that tend to flow polewards (at least from surface observations). Recall also from Ch. 3 that, starting from the Equator and going towards the Poles (in both hemispheres), we have the trade winds and prevailing westerlies that lead to an alternating wind pattern over the subtropics and mid-latitudes. It turns out there are also the surface *polar easterlies* over the polar regions that are westwards (similar in orientation to the trade winds, converging onto the Arctic and Antarctic circles). You can convince yourself that the subtropic and subpolar gyres are flowing in the same orientation as the wind forcing (and yes, we are going to start using *wind stress curl*). These observations are perhaps not entirely surprising, since from intuition it's kind of hard to see that it could be anything else since, from an energetic point of view, flows should be expected to work *with* and not *against* the wind.

However, the energetic argument does not explain why we should have *western intensification* and WBCs: we could for example have a broad flow which might in face be dynamically more favourable (e.g. less *unstable*; see Ch. 6). Here we provide a largely pictorial argument as to why we should have intensification and, in particular, why we the intensification should be on the west and not on the east. Note that you get maybe a plausible argument to the *why*, but not the quantitative aspects (e.g how narrow the WBC might want to be, for which you need some more physical arguments and maths) although it does suggest where to look. The main ingredients and broad logical deductions we will go through are as follows:

1. dissipative/diffusive effects are weak away from land boundaries (Ek $\ll 1$ and away from boundary layers, see Ch. 3.4.2);
2. *Sverdrup balance* relates the wind stress curl with the interior flow (cf. geostrophic balance in Ch. 3.2.3);
3. *mass conservation* implies configuration with essentially two orientations;
4. *vorticity balance* sets the orientation, and intensification needs to be on the west over a boundary layer.

4.2 Wind driven theory

4.2.1 β-plane and model set up

For the intended purposes we first start by formalising a concept that we have actually implicitly been using already in Ch. 3 (e.g. Fig. 3.12), called the **β-plane**[3]. What you imagine is something like Fig. 4.1, where instead of considering the sphere (which is curved and a bit annoying to deal with), you consider a point on the sphere, and imagine everything is actually flat around this point[4], so instead of dealing with co-ordinates on a sphere, you can use the standard Cartesian co-ordinates (x, y) and use the usual $\{\boldsymbol{e}_x, \boldsymbol{e}_y\}$ basis (see discussion around Fig. 1.25) as a substitute for longitude and latitude respectively.

So, again, we have (x, y, z) to be zonal-meridional-verical, but we need to change the Coriolis parameter $f = f(\text{latitude})$ to be in line with the new co-ordinates. Since y is essentially the measure of latitude, we should have $f = f(y)$, and the simplest thing to do is to take

$$f = f_0 + \beta y, \tag{4.1}$$

where f_0 is a constant and $\beta = \partial f/\partial y$ measures the change of the Coriolis parameter with latitude[5]. The value of f_0 depends on the

[3] Not to be confused with the coefficient of haline contraction β in Ch. 2.3. Context should be clear.

[4] It's what is called a *tangent plane*, i.e. the rectangular plane that only touches the sphere point of reference.

[5] If we take $f = f_0$ then this is called an *f-plane approximation*.

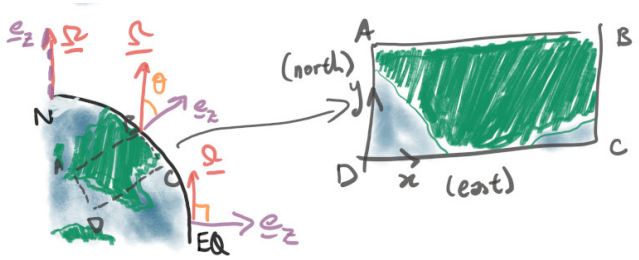


Figure 4.1: β-plane schematic. (lon, lat) $\rightarrow (x, y)$ with $f = f_0 + \beta y$ on the plane. Recall Fig. 3.9.

reference latitude chosen, but on Earth this is roughly $f_0 \approx 10^{-4}$ s^{-1}, and $\beta \approx 10^{-11}$ m^{-1} s^{-1} (note that y is now measured in meters instead of degrees/radians, and is probably going to be big, so β is small to compensate).

We set up an idealised model up with some simplifications to get a simple theory for gyres and WBCs:

- Northern Hemisphere β-plane ($f = f_0 + \beta y > 0$), with a rectangular domain[6];
- subtropical wind profile, assumed to be in the zonal direction but only varying in the meridional direction;
- ρ is constant, or we depth-integrate, and that there is no vertical dynamics and model is *depth-independent*[7] with no bathymetry;
- friction dominates only over the lateral frictional boundary layers.

Schematically the model is depicted in Fig. 4.2. In the β-plane formalism, we denote the wind forcing as $\boldsymbol{\tau} = (\tau^x(y), 0)$, i.e. zonal wind with meridional shear only. Here the winds going westwards and eastwards represent the trade winds and prevailing westerlies respectively (Ch. 3.3.1), and note that $\partial\tau^x/\partial y > 0$. Since we are in the same situation as the set up in Ch. 3.3.3 (zonal wind with meridional shear only), the wind stress curl is $\boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{\tau}) = -\partial\tau^x/\partial y < 0$. You might have anticipated this already, since the wind is blowing in a clockwise sense, so by convention this is a negative curl.

[6] We don't actually need the rectangular aspect for the pictorial argument, but if it is square you can actually solve the resulting model by hand! See bottom of Ch. 4.2.3 for a sketch, and Vallis [2006] Ch.14 for more details.

[7] Sometimes instead of *depth-independent* or *depth-integrated* the term *barotropic* is used, but I don't like barotropic used like this, because barotropic has the very specific meaning that $\rho = \rho(p)$ and refers to the property of the fluid. A barotropic fluid could have depth-independent dynamics, but a depth-integrated model does not necessarily mean the fluid concerned is barotropic. This 'issue' will come up again in Ch. 6 when we talk about *vertical wavenumbers*.

4.2.2 *Sverdrup balance*

To keep the discussion clean I'm going to sketch out some equation manipulations. First we start with the horizontal momentum equation Eq. (3.1). For the purposes here we hit the equation on both sides with a curl $\nabla \times (\cdot)$, which gets rid of the $-\nabla p$ term (because $\nabla \times \nabla\phi = 0$ for any function ϕ). Taking the curl of a vector gives a vector, but since we make the depth-independent assumption with everything being only a function of the horizontal co-ordinates, it

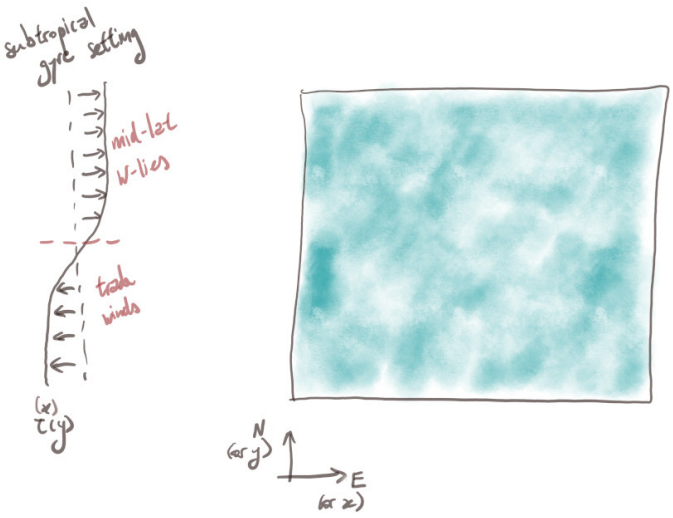


Figure 4.2: Schematic of wind-drive model (NH, assume subtropical wind profile, β-plane, square, homogeneous in density)

.

turns out the only non-zero entries are in the vertical component, and we get the scalar equation

$$\frac{\partial \zeta}{\partial t} + \underbrace{\boldsymbol{u} \cdot \nabla \zeta}_{\text{inertia}} + \underbrace{\beta v}_{\text{Coriolis}} = \underbrace{-r\zeta}_{\text{drag}} + \underbrace{F_\tau}_{\text{wind}}, \tag{4.2}$$

where $\zeta = \boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{u})$ is the vertical component of the vorticity (end of Ch. 3.2.3). Here we assumed linear drag (Ch. 3.4.3), no diffusion, and we denote $F_\tau = \boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{\tau})$ as the wind-stress curl. Only the β term survives the derivative from the curl (because derivative of $f_0 =$ constant is zero).

Now, recall that if we are in the Ro $\ll 1$ regime then the time-derivative and inertia terms are small. Dropping those terms results in

$$\beta v \approx -r\zeta + F_\tau. \tag{4.3}$$

This is sometimes referred to as **Stommel's model**[8].

Away from boundary layers, the assumption is the drag is unimportant, so we have the dominant balance

$$\beta v \approx F_\tau, \tag{4.4}$$

which is called the **Sverdrup balance**[9]. This is not dissimilar to geostrophic balance, except here we don't have $-\nabla p$ since we curled it away, so the balance is between Coriolis effect and wind. The flow v that is implied is sometimes referred to as the **Sverdrup interior**, and for the subtropical gyre wind configuration, since F_τ is the wind stress curl and we already argued this is negative, the

[8] After the American oceanographer Henry Stommel (1920-1992), widely recognised as one of the most influential physical oceanographers. This model forms the basis of Stommel's pioneering work on explaining western intensification.

[9] The same Sverdrup as the unit Sv.

Sverdrup interior is $v < 0$, i.e. a southward flow. This is drawn on as the yellow arrows in Fig. 4.3.

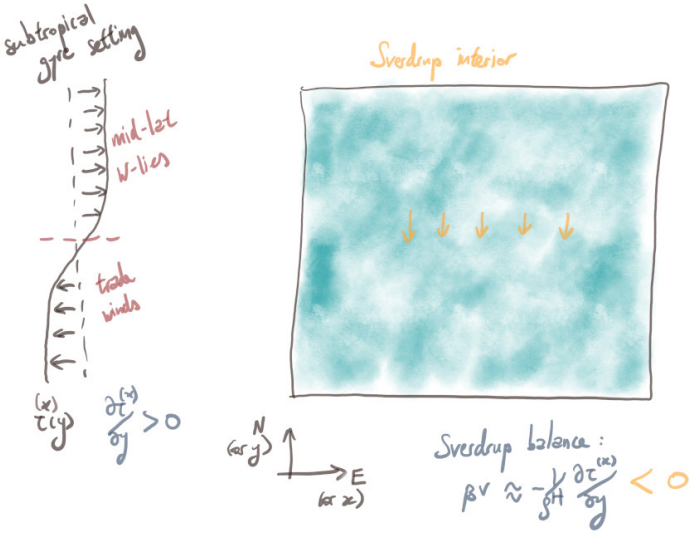


Figure 4.3: Sverdrup balance in the interior for the Northern Hemisphere single gyre case. Sverdrup interior is southwards.

Now, because mass conservation (cf. Ch. 3.3.3) we have to have an intense northward flow somewhere returning all the mass moving to the south, but Sverdrup balance doesn't tell us where this return flow is located though. For example you can convince yourself that in Fig. 4.3 you can have the return flow could be along the western boundary, so the gyre is in a clockwise rotating sense (Fig. 4.4), but an equally consistent configuration would be having the return flow along the *east*ern boundary, so the gyre is in an anti-clockwise sense (Fig. 4.4). You could even have both cases, with a clockwise rotation on the left half and an anti-clockwise rotation on the right half! Intuitively we expect it should be western boundary case, because (1) we actually know the answer, but (2) the winds are blowing in a clockwise sense, so it makes sense for the flow to not be working against the wind.

4.2.3 *Vorticity balance*

The following argument using vorticity sources and sinks to tell us why the northward return flow cannot be on the east in this set up. The argument is based on having a balance of sources and sinks: if the sinks cannot remove inputs from the sources then solution should not be physically realisable (because it doesn't satisfy the governing equations and no balance is possible). By assumption, the sinks arising from friction is only strong over a boundary layer at the

boundaries, and we ask whether these sinks can act to remove the vorticity being put into the system.

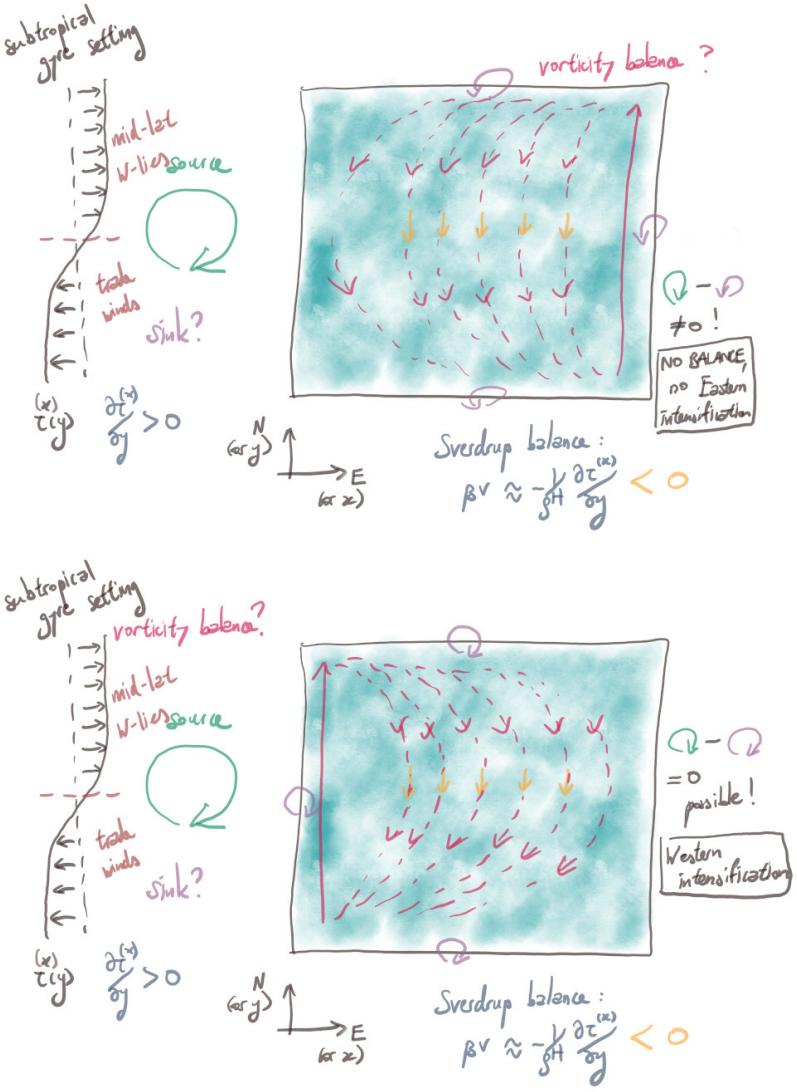


Figure 4.4: Vorticity balance in a Northern Hemisphere single gyre model. The hypothetical eastern intensification scenario is not realisable because there is no vorticity balance. The western intensification case is ok.

Consider the schematic drawn in the top panel of Fig. 4.4 for the case of eastern intensification that shouldn't work. Now, the wind is putting in negative (clockwise, anti-cyclonic) vorticity into the system. *If* we had eastern intensification, then the flow will possess positive vorticity (anti-clockwise, cyclonic). As it rubs against the boundaries the boundaries will remove vorticity, but in this case it only removes positive vorticity because the flow only has positive vorticity. But the wind inputs negative vorticity, which is not removed by the sinks available because of the flow configuration, and

thus there is no possible balance in the vorticity, which implies this state cannot be physically realisable in this set up. By contrast, for the case of western intensification drawn in the bottom panel Fig. 4.4, the flow will possess negative vorticity (clockwise, anti-cyclonic), and thus friction at the boundaries removes negative vorticity also. In this instance, a balance is at least possible, so the configuration could at least be physically realisable, within the approximated system we (well Stommel) cooked up. So we conclude we should have WBCs arising from this kind of boundary intensification, and this has to happen on the west.

This bit is an aside: If the domain is rectangular it is in fact completely possible to obtain an analytical closed-form solution to Eq. 4.3. The argument proceeds as above, and you find the *interior solution* satisfying Sverdrup balance Eq. 4.4, and a *boundary layer solution* that holds over the boundary layer. Then you demand that the interior and boundary layer solution match up over a transition region (this is a standard technique in *asymptotic analysis*). You end up with a western and eastern intensified solution, but it turns out in this setting there is only one boundary condition you can put on, and you have to put it on the west for physical consistency, so you have to select the western intensification solution[10].

[10] Similar principles are employed when selecting the *branch* for *Kelvin waves* solutions, so that Kelvin waves have to decay when moving away from the coast; see Ch. 6.1.

4.2.4 *Double gyre analogue*

The arguments presented above carries over to a double gyre configuration as depicted in Fig. 4.5. Here, we include the trade winds, the prevailing westerlies and the polar easterlies, which is supposed to drive the subtropical and subpolar gyres. While we impose three 'types' of winds, the vorticity argument above uses the wind stress curl, i.e. gradients in the wind, so we actually have negative wind stress curl over where the subtropical gyre would be, and positive wind stress curl (anti-clockwise, cyclonic) over where the subpolar gyre would be. Svredrup balance already told us in the subtropical gyre we should have a slow, broad southward flow in the interior. In the subpolar gyre, since the wind stress curl is of the other sign, the Svredrup balance implies a Svredrup interior that is northward. The same argument with vorticity balance in the subpolar gyre also implies you cannot have eastern intensification. If the flow in the subpolar gyre is going to be eastern intensified, then the flow has negative vorticity (as in the subtropical gyre case), so friction takes out negative vorticity, but the wind is putting in positive vorticity, and there is no vorticity balance. On the other hand, for the western intensification case, there is at least the possibility for vorticity balance, and thus setting the gyre orientation.

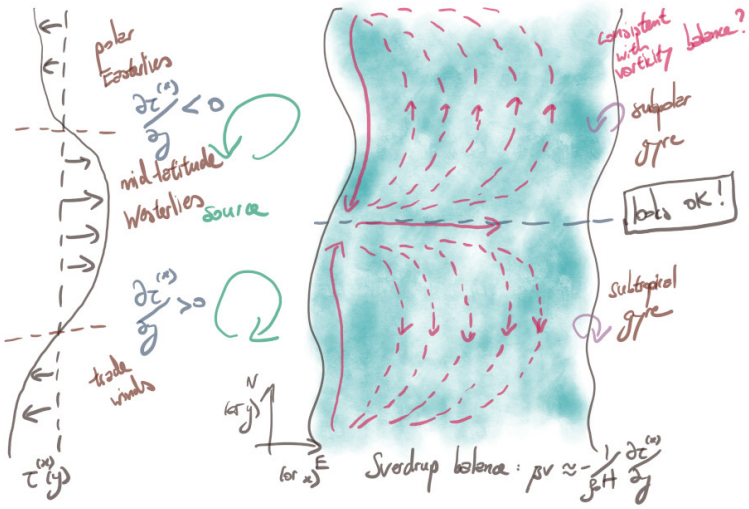


Figure 4.5: Vorticity balance in double gyre model.

To close this part, we note that vorticity is one way of arguing for gyre orientation and western intensification. We could have used momentum arguments too, but it is perhaps not as clean as the vorticity argument, partly because you need to keep track of pressures and see where their gradients are pointing. One of the main points though is that the vorticity argument 'works' for both the subtropical and subpolar gyres. One argument I have seen floated around is that you have the subtropical gyre configuration because you have a WBC going eastward, the Coriolis effect will lead to a deflection to the right in the Northern Hemisphere, leading to the subtropical gyre configuration. This 'short circuiting' argument doesn't really work because (1) the argument is back to front (you assume the answer first and then justify the origin of the answer, using the answer itself, which is wrong), (2) it doesn't work for the subpolar gyre configuration, so the argument doesn't stand up to scrutiny.

To clarify, the vorticity argument is not to say we cannot have *eastern boundary currents*, but the currents that are intense with somewhat significant vertical extent ($\sim$ 1000 m) starting at the surface are all on the west. We do have eastern boundary currents that are on the east near the surface, but they are usually slower and substantially more shallow ($\sim$ 250 m say), and these are driven by other means. Nevertheless, these surface eastern boundary currents are important because they are associated with Ekman upwelling and have important consequences for biogeochemistry; more here?.

4.3 *Beyond wind driven theory*

In the previous arguments I have tried to be a bit pedantic about stating the conclusions are consistent and at least valid within the model set up chosen (depth-independent, no bathymetry, zonal wind forcing with meridional shear, no nonlinearity from Ro $\ll$ 1). Of course these are only convenient simplifications to help us learn about the underlying processes somewhat. If you just use the simplified presented model above, the gross features might be qualitative consistent, but the quantitative aspects might disagree (e.g. WBCs are along latitude lines when the real world ones have a *pole*ward deflection, the resulting transport is too large is 'realistic' values for friction coefficients are used). Here we look at what aspects of the arguments presented above change if we include some of the additional features relevant to the ocean.

4.3.1 *Role of topography and bottom pressure torque*

For the moment we will keep the depth-independence and ignore the nonlinear terms, and see what happens when we have bathymetry. Fig. 4.6 shows intuitively what could happen. If you have bathymetric features such as seamounts or slopes, one consideration is whether flow would go over or around. While energetically speaking you may expect the flow to not want to work against gravity, so with a preference to not got over the mound, this somewhat depends on the characteristics of the bathymetric feature. If the seamount for example is relatively flat but wide, it might actually be preferable to go over (taking a shorter distance) instead of going around (not working against gravity but have to travel further). So what we expect is that the overall arguments relating to gyre circulation (such as orientation) is not modified in the presence of small bathymetric features, although the details may have local modifications.

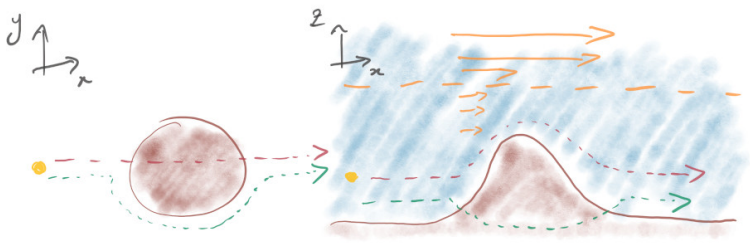


Figure 4.6: Possible paths when facing an obstacle: go over or go around. Depending on H as well one might be more preferable than the other.

The same of course cannot be said if we have large-scale bathymetric features such as continental slopes or mid-ocean ridges, because these will now steer and place strong constraints on what the flow

can or cannot do. In particular, we know gyres and WBCs have to be affected by continental slopes because they touch the boundaries. One concept we will introduce for this discussion is that of **potential vorticity**, which for the purposes here is defined as[11]

$$q = \frac{f + \zeta}{H}. \tag{4.5}$$

[11] This are more general/analogous definitions depending on context but we will not touch on those here.

It turns out it is this quantity that is conserved (up to forcing and dissipation) as the fluid moves around, and the conservation of q places strong constraints on the dynamics. One example is shown in Fig. 4.7, where we make the assumption that $f = 0$ or $|f| \ll |\zeta|$, so $q = \zeta/H$. The column of fluid has a particular spin measured by the vertical component of the vorticity ζ. As the fluid column moves into a region of shallower water, q is conserved, H is decreasing, so this implies ζ has to also decrease, i.e the fluid has to spin *slower*.

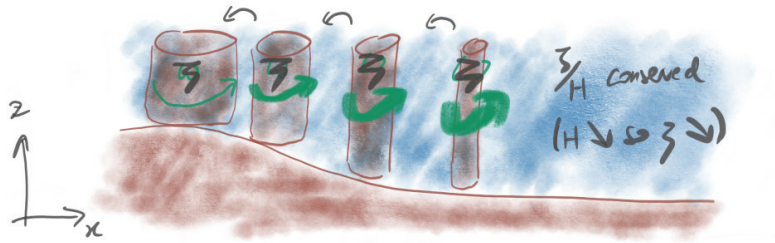


Figure 4.7: Conservation of $q = \zeta/H$ (assuming $|\zeta| \gg |f|$ for illustration). The width of the spin arrows denotes the magnitude of ζ and the volume of the fluid column is 'wider' to roughly denote volume conservation. As H decreases, the spinning rate ζ decreases such that q is conserved.

This is similar to the *ballerina effect* and is a manifestation of **angular momentum** conservation (cf. linear momentum which is mass times velocity, analogous definition for rate of spinning motion). You can try this for yourself with a spinning chair: sit in the chair, spin it around, and if you stick your leg out the rotation slows down, while if you tuck yourself in then you end up spinning faster. When you are rotating around you can imagine you sketch out an effective circle, which then has a radius associated with it. As you stick your leg out, you increase this effective radius, but correspondingly your rate of rotation decreases, such that a combination of the two is constant[12]. The fluid is doing something similar: as it moves into shallow regions, it gets squashed in some sense, and to preserve volume, the effective radius of the cylinder has to increase, so it means the spinning rate decreases.

[12] Or, $L = mur$, where L is the angular momentum, u the *angular velocity* (i.e. related to how fast you are spinning), and r the effective radius. L conserved and m being constant means if r increases u has to decrease.

A forcing that changes the spin of a body is called a *torque* and, in this setting, the change in the bottom depth because of the presence of bathymetry could be thought of a **bottom pressure torque** acting on the fluid[13]. What this means is that, in the presence of bathymetry, we have an extra effect on top of the ones given in Eq. (4.3) that may be important. As we have noted and alluded to,

[13] Remember that we have Newton's 3rd law, that while the ocean can 'push' on the ground, the ground equivalently 'pushes' back on the fluid. We will revisit this again later when talking about *topographic form stress* (Ch. 5.1.2).

the effect from bathymetry is related to the conservation of q. It turns out that, for sufficiently large-scale dynamics, $|f| \gg |\zeta|$ (see exercise) and so $q \approx f/H$, and what this implies is that *the large-scale flow should largely follow f/H contours*.

But we know the f/H contours, because these are system parameters! If there is no bathymetry, H is constant so the contours are parallel to latitudes, since it is f that changes with latitude. This is somewhat illustrated in Fig. 4.5, where the WBC is following the latitude lines. The Sverdrup interior doesn't satisfy the constraint as such because q is not completely conserved when forcing is involved, but again remember Sverdrup interior flows are weak, so we only have a mild violation of the conservation constraint. The conservation constraint doesn't apply in boundary layers because friction and dissipation is large there.

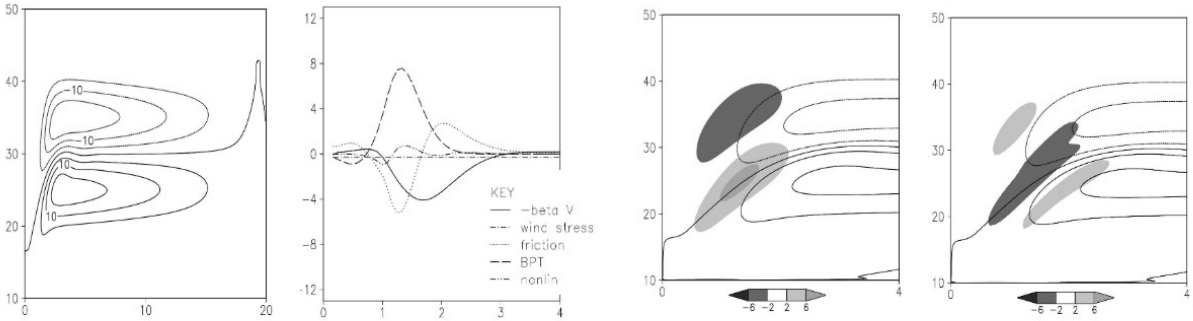


Figure 4.8: Results form a depth-independent gyre model with a slope on the west, showing (*a*) streamfunction, (*b*) balances between various terms over the western part of the domain, (*c*) bottom pressure torque forcing, and (*d*) frictional forcing. Taken and adapted from Jackson, Hughes & Williams, 2006, *J. Phys. Oceanogr.* (their Figs. 1–4). Also see Williams and Follows [2011], Ch. 8.3.

Now we see what we can get if we have a rectangular domain in the Northern Hemisphere still, but with a continental slope protruding from the west for a bit but is otherwise flat in the interior, the ocean is shallow on the western boundary (e.g. slops configuration like in Fig. 4.7), and the bathymetry is a function of longitude but not of latitude. As we go from the west to the east at fixed latitude, f/H decreases (because H increases), while as we go south to north at fixed longitude, f/H increases (because f increases). So if you start in the middle of the domain and you keep tracking where f/H is constant, you can convince yourself that the contours are latitude lines in the interior, but as you hit the continental slope this contour has to deflect to the *south*. The flow and in particular the WBC should follow the f/H contour as it splits from the boundary layers, i.e. go north-eastwards as it splits from the western boundary. This is indeed what happens in the results from a numerical experiment shown in Fig. 4.8 (first panel), plotting what are called

the **streamlines**, with the flow mostly following the streamlines (the WBC can be regarded as the streamline that detaches from the western boundary). The vorticity balance now is not necessarily just between wind stress curl, friction and Coriolis effect, and bottom pressure torque. The latter should be significant particularly near where the continental slopes are, and indeed this is seen in the second panel of Fig. 4.8, showing the diagnosed contributions to the vorticity balance, zoomed in over the slope region. The third and fourth panel shows the contribution of bottom pressure torque and friction to vorticity. Friction acts to slow down the spin within the WBC, but bottom pressure torque is leading to an increase in the spin, which is consistent because as the water column moves to a deeper region it spins faster so that q remains roughly conserved.

4.3.2 *Nonlinearity and baroclinicity*

Note that Fig. 4.8 shown above if for a case where the model was forced to be depth-independent, but is allowed to have *eddies* in it (there is a nonlinear contribution in the second panel). One can ask what happens if you include the eddy contribution, and/or what happens in you include allow for depth variation, i.e. density is no longer constant, and the system is now generically **baroclinic**. It turns out the nonlinearity and the baroclinicity by themselves do not modify the qualitative aspects in gyre circulations that much[14], although they do matter for the quantitative aspects. In the depth-independent case the eddies don't actually contribute very much to the overall balance (as can be seen in second panel of Fig. 4.8). When the model is baroclinic, there can now be a disconnect between the near surface and bottom flow (e.g. Fig. 4.6)[15], but the bottom pressure torque is still of relevance because conservation of an analogue of q places a strong overall constraint on the flow dynamics.

One thing we will highlight here is the effect when you allow for

[14] The same statement is not true for the Antarctic Circumpolar Current; see Ch. 5.1.1.

[15] Despite measures such as *JEBAR* stating otherwise; see e.g. Cane, Kamenkovich & Krupitsky, 1998, *J. Phys. Oceanogr.* for an explanation).

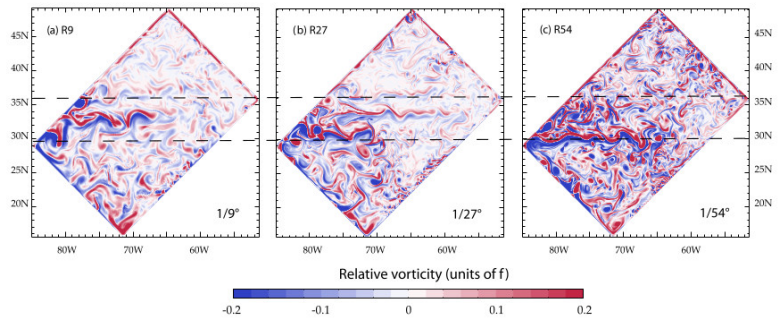


Figure 4.9: Snapshots of surface relative vorticity of a double gyre model at different resolutions. From Lévy et al. [2010] (modified from their Fig. 3).

both barolinicity and the possibility for eddies. This combination allows for *baroclinic instability*, which is one of the fundamental processes in atmospheric and oceanic dynamics. Fig. 4.9 shows results from a numerical simulation within a tilted baroclinic gyre model where they progressively increase the resolution of the model, going from a *mesoscale* eddy 'resolving' model (eddies arising from standard baroclinic instability, eddy scales around 50 to 100 km or so) to a *sub*mesoscale eddy 'resolving' model (mesoscale eddies as well as submesoscale eddies of 10 km scale or less arising from *symmetric instabilities*). While the overall characteristics for gyre circulation and WBC may not be so drastically different, the details starts to really matter. As can be seen from the model, the WBC extends *further* as well as detaches at a *lower* latitude as we allow a larger range of dynamics in the numerical model. The observations are somewhat related to the discussion in Ch. 3.4, where we mentioned in passing that while stirring by eddies often leads to enhanced diffusion, these are not the only things they do, and they can (and do) feedback on the flow that generated them in the first place. We will revisit the role of baroclinic eddies again in Ch. 5.1 and Ch. 6.2.

4.3.3 *Buoyancy forcing*

The arguments given so far make no mention of buoyancy forcing, does it not matter? The 'classic' theory is for wind-driven gyres and there are some arguments as to why buoyancy should matter for the details but maybe not so much for the overall structure[16]. We do have buoyancy forcing in the ocean, and the effect of buoyancy on gyre circulation has been more revisited again recently. Fig. 4.10 shows results from one such study, where they find that buoyancy forcing has a stronger influence on the subpolar gyre. The study also finds that the buoyancy forcing influence is more notable when the model has a resolution that permits the mesoscale baroclinic eddies.

[16] Certainly if the model is depth-independent there is not much to do.

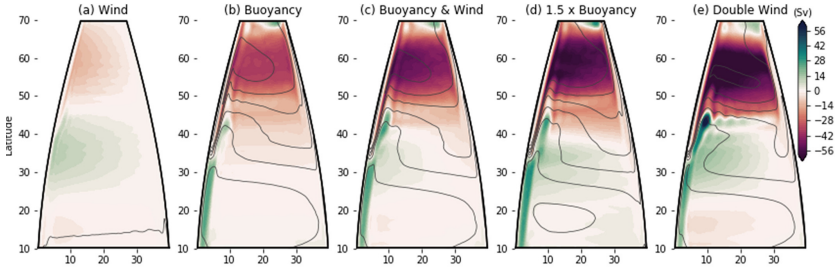


Figure 4.10: Streamfunction associate with depth-integrated flow (rows) for a few experiments, showing the streamfunction (as shading) and the SST (as gray lines). From Hogg & Gayen (2020), *Geophys. Res. Lett.* (modified from their Fig. 3).

Summary and further reading

In this chapter we examined the 'classic' theory of wind-driven gyres, as well as some extensions and/or alternatives. In the standard wind-driven gyre theory on Earth, vorticity balance between wind input and frictional dissipation on the boundary layers imply that intensification cannot be on the east, and it has to be on the west, with a Sverdrup interior depending on the wind stress curl. Bathymetry places a strong constraint on the flow via potential vorticity (related to angular momentum) conservation, and we find that the WBC in the Northern Hemisphere should deflect to the north-east as it detaches from the boundary and moves away from continental slopes, which is consistent with snapshots in e.g. Fig. 1.6 for the Gulf Stream (bathymetry also steers the current). Other effects such as nonlinear eddies, baroclinicity and buoyancy forcing further modify the details, though perhaps not so much on the overall characteristics.

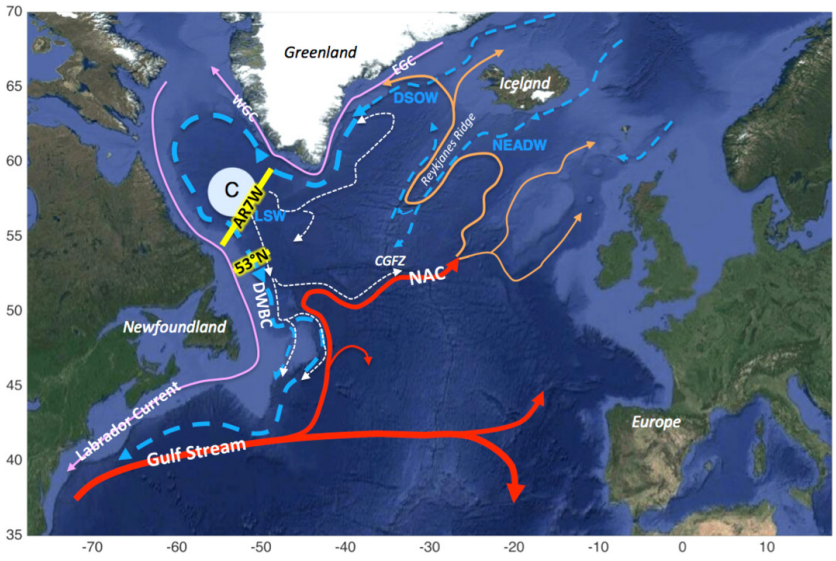


Figure 4.11: Schematic of WBC (Gulf Stream, the red line) and the deep WBC (the blue-dashed line). From Handmann *et al.*, 2018, *J. Geophys. Res: Oceans* (their Fig. 1).

One detail we are skipping over here completely is that, in the depth-independent model here, the WBC detaches but has to return at the eastern boundary. In the Atlantic, we appear to have a net motion of water moving towards the north, with the WBC continue going polewards towards the North Pole (Fig. 1.6), and we don't in fact have such a strong return near the eastern boundary as we have here. When we have baroclinicity what we have are **deep WBCs**, with a schematic given in Fig. 4.11. Water convects to the deeper parts of the Atlantic mostly within the Labrador sea area, and returns via a deep return flow below the WBC. We do not touch on the related

Stommel–Arons theory ref here, but we will say that the Stommel–Arons theory is one of the rarer occasions in physical oceanography where the prediction was made before the observation (and in fact guided the observation).[17]

[17] TODO: May add this in for completeness in the future.

One thing we do highlight is the existence of **deep Eastern Boundary Currents**. Like their deep WBC counterparts, these currents seem to exist at depth (below 2 km) and are *pole*wards. Unlike the arguments above, these deep Eastern Boundary Currents seems to only exist when baroclinicity and bottom pressure torque contributions (i.e. bathymetric effects) are included Yang et al 19.

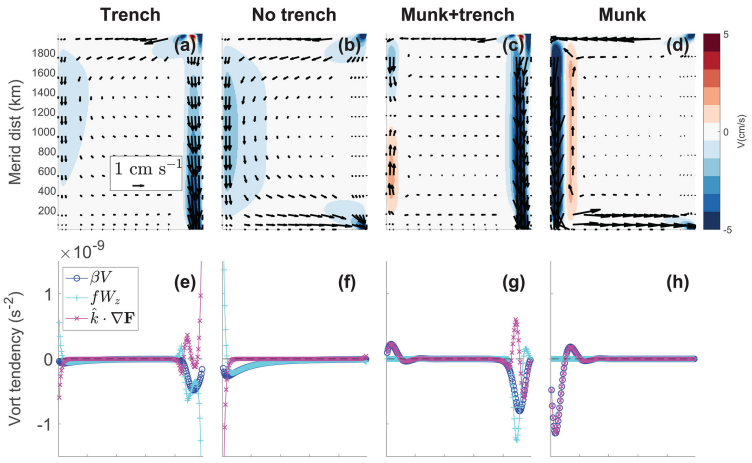


Figure 4.12: Meridional flow and vorticity budget and bathymetry (rows) from Yang, Tziperman & Speer, 2020, *Geophys. Res. Lett.* (modified from their Fig. 3).

The main slightly philosophical point is that theories are continually updated, either as new observations, tools and/or ideas come in. The models/theories created should not be regarded as the truth, the whole truth and nothing but the truth, but really more as a caricature/representation of the world. The value of models and theories should be measured in both their accuracy in representing the world, as well as what you actually end up learning from them. Simple theories/models in this aspect are important because you learn something from them, in contrast to having more complex theories/models that reproduces the things you care about, but the causalities are so entangled making attribution difficult. One might be more suited for learning and the other for predictions, but it's not that one is 'better' than other so to speak, maybe they should be seen as different tools for different problems.

I don't think anyone really believes the gyres and WBCs are really completely wind-driven (certainly the wind has a very strong influence on them). As we have seen, bathymetry, nonlinearities, baroclinicity as well as buoyancy forcing all have to play a role in some way[18]. Similar arguments about what process is the most

[18] I feel like I don't have enough evidence to satisfy myself that there is complete truth to declare allegiance to, so I remain neutral on these things (I don't really do beliefs)...

important for the Southern Ocean circulation is indeed an ongoing research topic. We will talk a bit about those in the next chapter, with a view to connect the discussion to the global Meridional Overturning Circulation.

Chapter exercises

1. What happens to the argument leading to Fig. 4.5 if you are in the Southern Hemisphere? What if the Earth was moving around the rotating axis the other direction? How about if the Earth was moving around the Sun in the other direction?

2. For the double gyre picture in Fig. 4.5, draw on or state the implied Ekman up and downwelling over the open ocean. Relate this accordingly to the Sverdrup interior. Repeat this for a Southern Hemisphere scenario.

3. There are some people who believe the Earth is in fact, flat (e.g. International Flat Earth Research Society). Repeat the wind driven gyre arguments presented above for a case of a flat Earth, noting any similarities and/or differences. What about for a cylindrical Earth (for the International Cylindrical Earth Research Society[19])? (Hint: drawing a picture of what the hypothetical Earths look like might help.)

4. For $q = (f + \zeta)/H$, carry out a dimensional analysis and argue that, for horizontal length-scales that are sufficiently large, $|\zeta| \ll |f|$ (recall that $\zeta \sim \nabla \times \boldsymbol{u}$). Estimate the horizontal length-scale at which the vorticity is comparable to $f \approx 10^{-4}$ (take the horizontal velocity scale $U = 10^{-1}$ m s^{-1} for simplicity).

5. Give a pictorial argument as to why f/H contours deflect to the south in the Northern Hemisphere as you are moving along latitude lines onto the western continental slopes from the open ocean. Is the deflection still int he same direction if you are moving instead onto the eastern continental slope? What if you are in the Southern Hemisphere? Justify your answers accordingly (pictorial or verbally).

6. Sketch out the gyre streamlines we may have if the continental slope is on the east instead of the west. Provide arguments to back up your answers.

7. What do you think happens if the bathymetry is so high that f/H contours are 'blocked' (i.e. the f/H contours are not able to cross the bathymetric feature)?

[19] I am the chairperson for this one (humour me).

5 Southern Ocean, and the Meridional Overturning Circulation

The previous chapter on gyres and WBCs largely focused on a depth-independent theory in the presence of lateral boundaries, touching on the role of topography, barolinicity and nonlinearity in passing. In reality all of the processes should play a role although there is of course arguments as to which one is the most important one. Here we talk a bit about the Southern Ocean circulation and argue in a very hand wavy way that everything is important, although we are mostly going to talk about the role of wind forcing and baroclinic eddies. Additionally, we will introduce the Meridional Overturning Circulation (MOC) more formally here than in Ch. 1, particularly focusing on how the Southern Ocean influences the global MOC, and why the Southern Ocean really might be seen as the 'centre' of the world's oceans.

5.1 Southern Ocean

To recap, the Southern Ocean connects to all the other major ocean basins in the world (except the Arctic), and this is seen in Fig. 5.1. The perhaps unconventional map projection used is called a *Spilhaus projection*[1], and this maps is centred on Antarctica and wraps the land and oceans around it (in an angle preserving or *conformal* fashion, though not area preserving). A schematic of the MOC is also drawn on, and visually it is suggestive that the Southern Ocean really is 'disproportionately important' from the global MOC point of view, since any modifications to the Southern Ocean circulation can have an effect on the *global* ocean circulation. We formalise this argument a bit more later (it's to do with the connections through density stratification).

Unlike most of the other oceans in the world, there are *open latitudes* in the current day Southern Ocean, i.e. latitudes that are unblocked by land boundaries. This has important dynamical consequences, notably the lack of zonal pressure gradients over unblocked

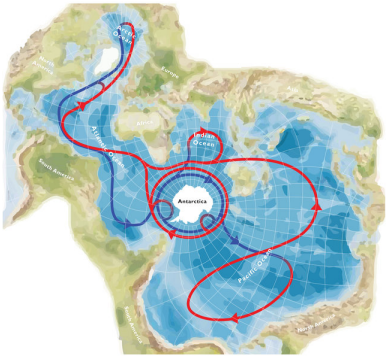


Figure 5.1: Spillhaus projection with a focus on the oceans and, in particular the Southern Ocean, with a schematic of the MOC put on (red = warm surface waters, blue = cold deep/abyssal waters). Atlantic is on the upper left part of the map. From Mike Meredith (BAS), from his Challenger Medal ceremony talk in 2018.

[1] After the South African-American geophysicist, oceanographer and cartoonist Athelstan F. Spilhaus (1911-1998).

latitudes above the bathymetric features: you can think of this as there not being any land that the water can pile up against to build a zonal pressure gradient, or from consideration that the zonal integral of $\partial p/\partial x$ over the periodic domain has to be zero. All this leads to very different dynamical balances, and one dynamical consequence is that the Southern Ocean possesses a very strong current system called the **Antarctic Circumpolar Current** (ACC), with a transport of around 130 Sv $= 130 \times 10^6$ m^3 s^{-1}. Contrast this to the Gulf Stream with its transport of around 30 Sv: while the Gulf Stream is more intense in terms of flow speeds, the ACC occupies a much larger volume, so even if it is relatively 'sluggish' it moves more water around. The large transport we will argue to be linked to the stratification later on, using a combination of hydrostatic balance (Ch. 3.1.2) and geostrophic balance (Ch. 3.2.3).

While there are open latitudes on the surface there are of course bathymetric features below the surface that influences and steers the flow. Fig. 5.2 shows the bathymetry around the Drake passage region (between South America and Antarctica), which is one of the choke points for the ACC. The bathymetry is expected to influence the large-scale flows by affecting the f/H contours, argued in Ch. 4.3 to place an important constraint on what the flow can and cannot do via potential vorticity conservation (up to forcing and dissipation). Other notable features such as the East Pacific Rise and the Kerguelen Plateau deflect the f/H contours, leading to a deflection of the flow.

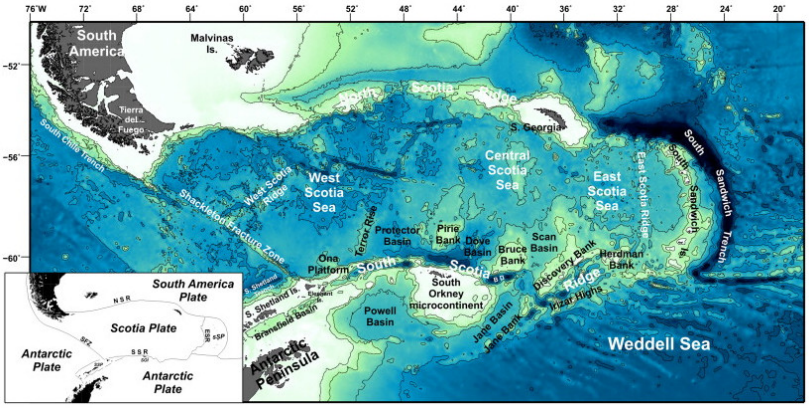


Figure 5.2: Bathymetry around the Drake passage. Figure modified from Civile *et al.* (2012), *Tectonophysics* (top half of their Fig. 1)

.

The ACC is subject to both strong wind and thermodynamic forcing. The Southern Ocean is subject to some of the strongest surface eastward (westerly) winds on the globe, injecting eastward momentum into the ACC. The atmosphere is relatively speaking much colder than the ocean, and there is a buoyancy loss certainly in the southern part of the Southern Ocean (i.e. water loses heat to

the atmosphere and gets denser), but possibly buoyancy gain in the northern part of the Southern Ocean depending on the upwelled watermass (i.e. water gains heat and gets lighter; more in Ch. 5.2.2). Additionally there is the presence of *ice*, which also affects the resulting thermodynamic and mechanical forcing. Most of features these have seasonal as well as longer time-scale shifts, and as we know changes in forcing is expected to lead to changes in the circulation. The one we will sketch out here is the response to the changes in the *wind* (because it is easier to talk about with the tools we have); see the end of the chapter on further discussions on the changes arising influences by thermodynamic forcing and/or ice.

We may be tempted to try creating depth-independent theories for the ACC, and while the resulting broad structures might be reasonable, the associated transport tends to be way too large for realistic values of friction (e.g. Gill 68). The problem here is trying to reconcile the zonally symmetric (think of a re-entrant channel where water flowing to the east returns from the west) theories, where the transport is excessively large, with Sverdrup balance, which holds when meridional continental barriers are available. The balance is different in the ACC setting compared with the gyre setting, and the it turns out *eddies*, particularly those arising from *baroclinic instability* (Ch. 6.2), play a crucial role. Eddies are expected in this region since the flow is relatively strong. As mentioned in Ch. 4.3, baroclinic mesoscale eddies *need* the baroclinicity, and this is perhaps one reason why depth-independent theories do not 'work'[2].

[2] There are depth-independent-esque theories although they do invoke the baroclinic eddies in some form or another; see for example the work of Marshall et al. [2016].

5.1.1 *Overturning, and stratification point of view*

Lets start with a slightly easier way to view the Southern Ocean dynamics via considering how the density stratification could change under wind forcing, using the schematic in Fig. 5.3 (there is a westward wind over the shelves but we don't need that here). The argument proceeds by considering Ekman dynamics, which strictly speaking should only apply over a certain vertical extent, but lets roll with it and see what happens. So again we take the β-plane setting and consider $\boldsymbol{e}_y$ as pointing north and $\boldsymbol{e}_x$ pointing east as usual, and for simplicity we again consider a zonal wind that has meridional shear, i.e. $\tau^x(y)\boldsymbol{e}_x$. From this point of view, we see that as we move north from the southern part of the Southern Ocean, we have $\partial\tau^x/\partial y > 0$, i.e. positive shear in the wind stress, and that as we move further north pass the location of the maximum wind stress, we start having $\partial\tau^x/\partial y < 0$, i.e. negative shear in the wind stress.

From a wind stress curl point of view, convince yourself that we have negative wind stress curl in the south and positive wind stress

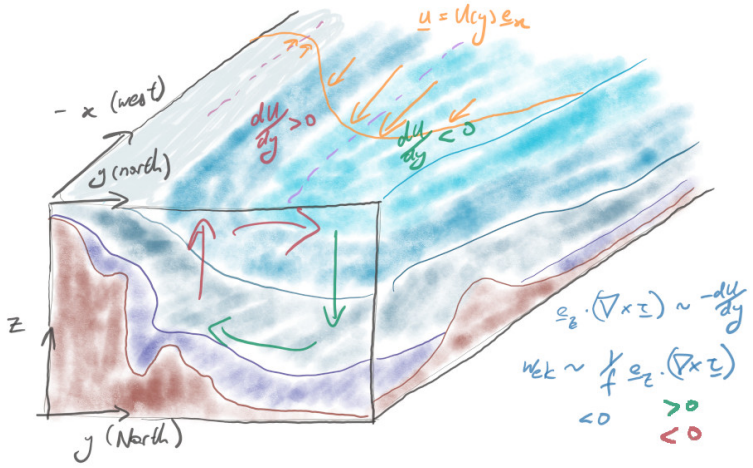


Figure 5.3: Schematic of wind forcing over Southern Ocean and associated Ekman circulation. $f < 0$ because we are in the Southern Hemisphere. Diagram based on Olbers

curl in the north, either by working out the curl explicitly, or noting that the positive wind shear gives you a clockwise spin, which by convention is negative curl (remember the definition of the curl is independent of the sign of f). Now, Ekman up/downwelling is related to to the wind stress curl but up to a factor of f, either from Eq. (3.13) or from considering whether the wind forcing leads to a convergence or divergence in the flow (noting that $f < 0$ in the Southern Hemisphere, so the Ekman transport is to the *left*). Either way, we have an upwelling to the south, and downwelling to the north, as illustrated in Fig. 5.3.

Fig. 5.4 considers what happens to the isopycnals under the kind of overturning implied by Ekman dynamics. If we suppose the isopycnals were initially flat, then the Ekman overturning will start tilting and steepening the isopycnals as water is being moved around by the Ekman overturning. The surface Ekman wind forcing moves the lighter water from the south to the north at the surface, tilting the isopycnals, so that the isopycnals eventually **outcrop** (i.e. connect to the ocean surface) to the south.

But presumably Ekman overturning cannot be the only thing at play, because if there wasn't something counteracting this steepening effect then the isopycnals will become vertical and eventually tilt over, implying there is *static instability* where denser water is above lighter water, leading to massive overturns in the Southern Ocean over a large vertical extent. In reality we know the isoypcnals are certainly tilted, but not vertical (e.g. Fig. 2.15). From an energetic point of view, the wind forcing is inputting kinetic as well as potential energy into the system: imagine you move a pen on the table to somewhere higher, you have to have done work against gravity,

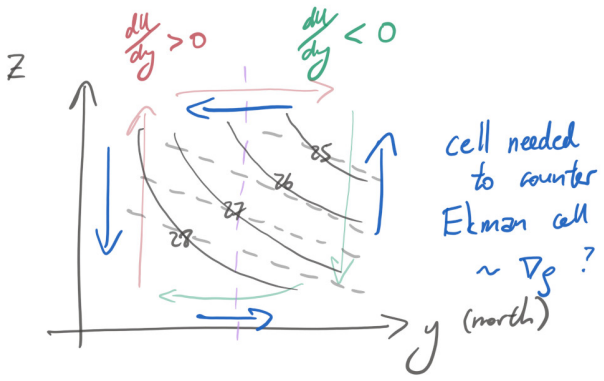


Figure 5.4: Ekman overturning and its consequences.

which means some energy has to be put in. In this case the wind is making the isopycnals steepen, when it otherwise wants to slump, so work is done and the ocean gains potential as well as kinetic energy. If there is nothing counteracting or releasing this input of energy then this is like saying you can indefinitely put energy into a closed and finite system, which surely is not physically realisable.

A resolution here is *eddies*, and in particular *baroclinic eddies*. We will say more in Ch. 6.2, but for now we just note that baroclinic eddies could be seen as the ocean system's way of moving / releasing this potential energy being put in to the system to some other form that fuels smaller scale dynamics[3]. From a mechanical point of view, the thing that is happening is that the collective effect of baroclinic eddies leads to an *eddy induced overturning* that wants to slump the isopycnals (and thus reducing the system's potential energy), counteracting the steepening effect of the Ekman overturning circulation. The isopycnal configuration we see is a balance between the Ekman and eddy induced overturning.

The thing to note is that while I've drawn the schematic so the overturning circulations span the whole depth, this is just for schematic reasons, and in reality the two processes vary spatially and possibly have regions where the effects do not overlap significantly. The other thing is that if you find some data, take a zonal average at fixed depth (as per usual) and compute the zonal average streamfunction (cf. Fig. 5.4), you appear to find this kind of large overturning cell. You don't really have that in reality, though the reason is more complicated[4].

[3] You could think of baroclinic instability as a conduit for moving energy from the large-scale wind forcing to scales smaller than baroclinic eddies. It turns out it is probably the primary conduit for this energy transfer between scales (e.g. Ferrari and Wunsch 2010).

[4] The result from doing a zonal average of the velocity at fixed height to get the overturning circulationin the Southern Ocean leads to what is called the *Deacon cell*. Similar synthesis in the atmosphere in the mid-latitudes leads to the *Ferrel cells*. In some sense both are artifacts of not taking the 'right' perspective, since the choice of averaging ignores the fact that dynamics wants to be doing stuff along-isopycnals, rather than horizontally (if isopycnals are flat then the two are equivalent).

5.1.2 *Form stress, and momentum point of view*

Lets maybe take an alternative viewpoint and see how Southern Ocean circulation functions, this time from the momentum view-

point, and in passing talk about what happens to the momentum that is put into the ocean by the wind. For this we introduce the concept of **form stress**, which are schematically drawn on in Fig. 5.5.

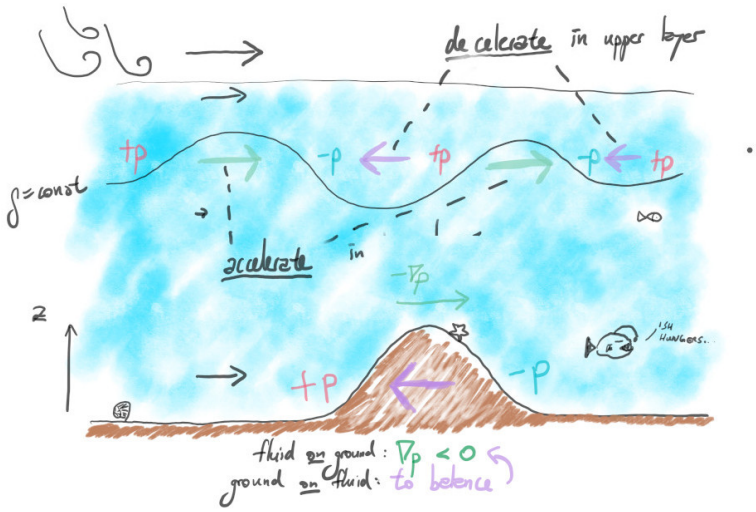


Figure 5.5: Schematic of interfacial and topographic form stress.

Intuitively what we expect is that if there is a flow at the bottom. then water is rubbing against the sea floor, there is friction, and friction acts to remove some of the momentum from the ocean (really it is just transferred into the solid Earth). Friction would presumably occur regardless of whether there are bathymetric features or not. A complementary way of transferring momentum into solid Earth would be via **topographic/bottom form stress**, which is probably the easier one to start with when trying to explain form stress. What you imagine is that, for argument sake, the wind is putting right-ward momentum in, and in turn drives a right-ward flow (the black arrows at the bottom of Fig. 5.5), ignoring Ekman dynamics for the moment. Then what you can think of is that water is being piled up on the left side of the seamount, leading to a larger pressure on the left of the seamount. On the other hand, assuming the flow is also going to the right throughout the domain, water is being moved away from the seamount, so there is less water pressing on the other side, and the pressure on the other side of the seamount is consequently lower. This pressure difference results in $-\nabla p$, i.e. a force, that is pointing into the seamount, and in the process moves the Earth a little bit, leading to a momentum transfer. There is of course an equal and opposite reaction from the Earth, where the Earth is pushing the ocean back, slowing the ocean flow down. This may all sound a bit odd that the ocean is moving the Earth and vice-versa, but again remember that momentum is $p = mu$, so since $m_{\text{Earth}} \gg m_{\text{ocean}}$,

the same momentum transfer means the ocean moves the solid Earth only very slightly, but the effect of the solid Earth slowing the ocean could be fairly significant.

Interfacial form stress is similar, again thinking about pressure gradients but across isopycnals instead of across bathymetry. If we go up the water column in Fig. 5.5 and consider an isopycnal draw on as the wavy black line, assuming there is a right-ward flow above the isopycnal, this sets up pressure gradients accordingly, with positive pressure anomalies on the left side of the peaks and negative pressure anomalies on the right side of the peaks. Then, drawing on the negative pressure gradients, we see see that the $-\nabla p$ is always pointing to the left in regions *above* the isopycnals, and to the right *below* the isopycnals. Above the isopycnal, $-\nabla p$ is acting against the direction of flow, thus leading to a *deceleration*, while below the isopycnal, $-\nabla p$ is pointing in the direction of flow, suggesting an *acceleration*. This can of course be thought of as the interfacial form stress leading to a transfer of right-ward momentum *vertically* downwards across the isopycnals.

It is no coincidence that interfacial form stress is also called **eddy form stress** because it is the baroclinic eddies that principally give rise to the perturbations in the isopycnals, resulting in the bumpyness in the isopycnals that allow for the $-\nabla p$ configurations to be set up. So from the momentum point of view, we have momentum input from the wind with removal of momentum by bottom form stress, with the vertical transfer mediated by eddy form stress. It turns out that with 'realistic' values of frictional coefficients such as those in Eq. (3.21), frictional removal of momentum is rather small, and the dominant balance in momentum really seems to be between wind and bottom form stress removal, certainly within the Southern Ocean. Again, the importance of eddies is highlighted through its role in the vertical transfer of momentum: if the eddies are artificially killed off, momentum transfer is weak in a baroclinic system, and you could envisage the momentum being stuck near the surface with no way of removal, leading to very large transports via significant increases in the flow velocity.

5.1.3 *Thermal wind, and how the two views are the 'same'*

It turns out both of the ways presented above are intimately linked with each other, and we briefly sketch out how. Recall that in Ch. 3 we have both hydrostatic and geostrophic balance, which within the approximations utilised here are given respectively by

$$\frac{\partial p}{\partial z} = -\rho g, \qquad f\boldsymbol{e}_z \times \boldsymbol{u}_g = -\frac{1}{\rho_0}\nabla p. \tag{5.1}$$

Upon looking at the equations, the thing you might considering doing is substituting for the pressure accordingly. This could be done by taking a ∇ of the hydrostatic balance and a $\partial/\partial z$ of the geostrophic balance, resulting in

$$f\boldsymbol{e}_z \times \frac{\partial \boldsymbol{u}_g}{\partial z} = \frac{g}{\rho_0}\nabla\rho. \tag{5.2}$$

Writing this out in component form explicitly, we have

$$\frac{\partial v_g}{\partial z} = -\frac{1}{f}\frac{g}{\rho_0}\frac{\partial \rho}{\partial x}, \qquad \frac{\partial u_g}{\partial z} = \frac{1}{f}\frac{g}{\rho_0}\frac{\partial \rho}{\partial y}. \tag{5.3}$$

The general equation Eq. (5.2) is called the **thermal wind shear relation** (or sometimes just *thermal wind balance*), and what this says is that, under the assumptions that hydrostatic balance and geostrophic balance hold (so $H/L \ll 1$ and $\mathrm{Ro} \ll 1$), *horizontal gradients of density are directly related to the vertical gradients in the geostrophic velocity*.

It is perhaps no surprise then momentum and stratification point of view presented are consistent, because in some ways they are different ways of saying the same thing. Since we have $\partial\rho/\partial y > 0$ and $f < 0$ in the Southern Ocean, so the second part of Eq. (5.3) states that $\partial u_g/\partial z > 0$, implying an eastward geostrophic flow that increases as we move up the water column towards the ocean surface. Note also that the Southern Ocean is one of the few places with strongly tilting isopycnals (e.g. Fig. 2.15), and so it is also entirely consistent that the Southern Ocean has the largest transport observed over the globe. A large-scale momentum forcing leads to large-scale flows and in turn is equivalent to saying there is a significant tilting in the isopycnals through thermal wind shear relation. Since baroclinic mesoscale eddies arise as perturbations to the large-scale flow, and that mesoscale eddies are sufficiently large-scale they can still be considered to be in geostrophic balance, the associated perturbation in the velocity from the baroclinic mesoscale eddies naturally leads to perturbations in the density, thus justifying the use of the term eddy form stress when talking about interfacial form stress.

Before we move on to the global overturning circulation, note that a hint of this thermal wind effect was already seen in the vertical section of Fig. 2.7, in the middle panel for the profile over the Kuroshio. The larger temperature gradients over the top 1000 m is suggestive of a strong flow in the region. Further, the thermal wind shear relation is a more general phenomenon that is a characteristic of large-scale rotating stratified fluid dynamics. The form of thermal wind shear relation given above in Eq. (5.2) makes use of the Boussinesq approximation, but turns out you don't really need it

to get thermal wind shear relation (and you can't really use it for example in the atmosphere over large vertical extents), although you need some other machinery. The term 'thermal wind' actually comes from the atmospheric literature, because in the atmosphere it is the temperature that controls the density. One example is that as you go from the mid-latitudes up to the polar regions in the Northern Hemisphere, you rapidly go from warm to cold regions, which means you go from a region with lighter air to denser air, and so $\partial\rho/\partial y > 0$. Since $f > 0$, $\partial u_g/\partial z > 0$ by the second part of Eq. (5.3), associated with the strong eastward flowing mid-latitude/polar *jet stream*. Similar ideas apply to the eastward flowing *subtropical jet stream*, jet streams in the Southern Hemisphere, and similar systems in planetary atmospheres for example.

5.2 The MOC

It is perhaps becoming clear how the Southern Ocean might be disproportionately important in the global overturning circulation. From Fig. 5.1 it is clear that geographically this is the case, but now we have a dynamical link through the thermal wind shear relation: the global ocean circulation is linked through connecting isopycnals, and since isopycnals are intimately linked to the geostrophic flow, any processes that leads to the modifications to the Southern Ocean circulation can certainly modify the *global* stratification on long time-scales. We explore this global connectivity in the **Meridional Overturning Circulation** (MOC) a little further in this section.

Recall that the Earth is tilted and the dynamics on Earth is largely driven by the Sun through uneven solar radiation (Ch. 2). By the fact that there is an uneven solar heating of the Earth, we naturally expect that there has to be movement of energy around in some form of another, because we have latitudinal gradients in heat, so minimally there should be some diffusion via conduction that leads to transport (Ch. 3.4). Of course the transport by conduction is extremely inefficient, while the Earth is not so warm that large amount of energy can be moved around by radiation, but of course the atmosphere and ocean are low viscosity fluids (Ch. 3.4), so it is perhaps natural to expect the atmosphere *and* the ocean to both have a MOC that to redistributes the energy so as to attempt to erase the latitudinal energy gradients.

Fig. 5.6 shows the amount of heat (interpreted as an energy) being moved around by the atmospheric as well as oceanic MOCs as a function of latitude (appropriately summed over space and averaged over time). The atmosphere does most of the redistribution of heat over the globe, accounting for at least two-thirds of the total.

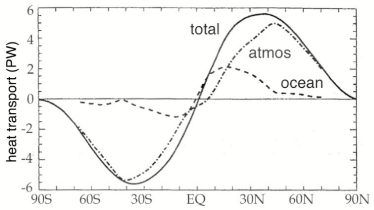


Figure 5.6: Total transport of energy to the *north* as a function of latitude, in units of *petawatts* (10^{15} W, where $1\ \mathrm{W} = 1\ \mathrm{J\,s^{-1}}$). The total is denote by the solid line, and the atmosphere and ocean component is respectively denoted by the dot-dashed and dashed line. Reworked from Trenberth & Caron, 2001, *J. Climate*.

The ocean from this point of view moves less heat around, but the amount is not entirely negligible over the Northern Hemisphere, and the magnitude is certainly bigger in the tropics. An interesting feature to note is that while the total (and to a lesser extent the atmospheric component) is effectively symmetric about the Equator, this is not the case for the ocean, where there is a preference for the energy to be moved *north*wards.

The northward energy transport is largely through the WBCs and in particular the **Atlantic Meridional Overturning Circulation** (AMOC), which I'm going to take to cover everything in the Atlantic that leads to a meridional transport after a zonal average, which could include the gyre systems and the WBCs (both the surface and the deep ones) as covered in Ch. 4, the eastern boundary currents, and anything else that is in between. Fig. 5.7 (as well Fig. 5.1) provides a schematic of the AMOC and the dominance of heat being moved northwards by the surface currents in the Atlantic (but not so much in the Pacific). Again, we have seen this already in Ch. 1.2.1 and Ch. 4. The Gulf Stream moves the warm surface Equatorial water towards the North Pole, releasing some of this heat to western Europe. When the water recirculates it gets *transformed* into denser waters via cooling especially around the Lab sea, which fuels the deep WBC in the Atlantic (Ch. 1.3.3, Ch. 4, Fig. 4.11). The deep WBC moves water at depth towards the South Pole.

Some of this water turns out to get upwelled over the Southern Ocean (Ch. 5.2.4) and circumnavigates around the globe as part of the Southern Ocean circulation. A description of the global MOC however is deferred to the end of this chapter, partly because I want to use a different figure than Fig. 5.7, which definitely requires few more bits of terminology, and the dynamics contributing to the eventual description would probably be helpful too[5].

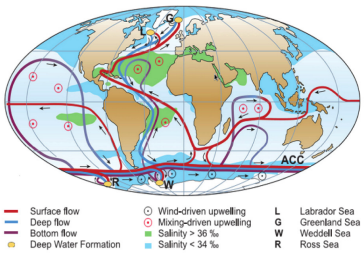


Figure 5.7: Schematic of the global MOC (red: surface warmer waters; blue: deeper colder waters; purple: abyssal cold waters). From Rahmstorf (2002), *Nature*, figure in Box 1.

[5] It's Fig. 5.17 if you want to see it now.

5.2.1 *MOC vs. global conveyor belt vs. thermohaline circulation*

The currents drawn on in Fig. 5.7 (and to a lesser extent in Fig. 5.1) are sometimes referred to as the **global (ocean) conveyor belt**[6]. My biased point of view is that these kind of pictures should be viewed as useful *schematics*, as the overturning circulation itself is not really a conveyor belt as such. The MOC should really be seen as an *average* (over space and/or time) because it is a climatological phenomenon, rather than some well-defined persistent current system as implied by the conveyor belt picture. It is certainly true that the MOC has features such as the WBCs where the current systems are well-defined in space and/or time, but these should be seen as exceptions rather than the norm. Another term you might see in place of the

[6] The American geochemist Wally Broecker (1931-2019) seems to be the one attributed to coining this term.

MOC is the **thermohaline circulation**, which again I am not going to use out of pedantry: the term implies that the circulation is driven by gradients in temperature and salinity, but we know that is not the complete picture (e.g. wind and tides are important), and also there can be flows driven by density gradients that do not contribute to the MOC.

I personally prefer 'the MOC' because it is technically correct, but also sufficiently non-committal.

5.2.2 *Watermass properties*

To talk about the MOC we need to talk a bit about **watermass properties**, in order to track water being moved around so as to infer for the sense of water flow arising from the MOC. For illustration, we stick with the AMOC for the moment. Watermass properties refers to the water having a specific *signature* in the *tracer* property that provides a suggestion of where the water came from (cf. a 'tag' for the water). Fundamentally this relies on the mixing to be 'weak' (Ch. 3.4), otherwise the distinguishing tracer features gets eroded. Ideally we would throw in a dynamically / chemically / biologically inert and *conservative* 'dye' that tracks the waters as it moves around the global ocean. In practice we can't really do that for money and/or environmental reasons (e.g. plastics would actually quite a good tracer because of its properties), so we make do with other 'dyes' that are not necessarily completely passive, but is good enough for most intents and purposes[7].

[7] You could actually do it in a numerical model but that's not the actual ocean so to speak.

It is perhaps easiest to explain the concept with an example. Fig. 5.8 shows a meridional section of salinity in the Atlantic (cf. Fig. 2.11). What we see is that there is a blob of water with higher salinity intruding in from the north, connecting the northern higher latitudes with the interior waters at depth. This watermass is known as the **North Atlantic Deep Water** (NADW), starting its life as initially as warm salty tropical water carried northwards by the Gulf Stream before being *transformed* to denser waters in the Lab Sea, returning south from the northern higher latitudes as part of the deep WBC, while roughly preserving the salty characteristic as it is moved around. The Med Sea salty overflow waters also contribute to this deep water. See more about watermass transformation in Ch. 5.2.3.

Salinity can also be used to identify the **Antarctic Intermediate Water** (AAIW), as seen in the relatively fresher tongue of water coming from the south in Fig. 5.8. One origin of the AAIW is from the salty NADW being upwelled in the Southern Ocean (Fig. 5.3), which is then subjected to mixing and dilution with the fresher Antarctic waters within the turbulent Southern Ocean, and then moved northwards by the Ekman overturning circulation, and subducted by the

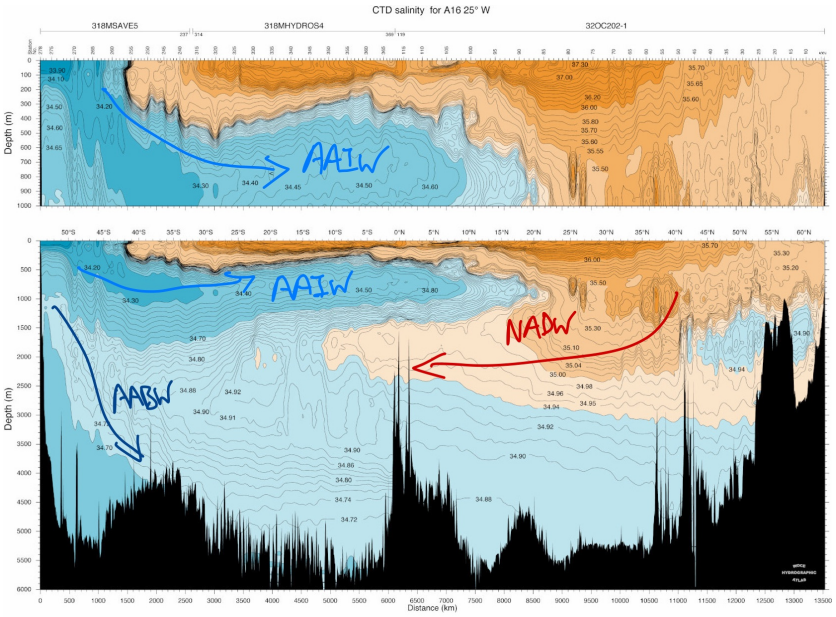


Figure 5.8: Salinity profile at 25°W, with a zoom near the surface. Highlighting the North Atlantic Deep Water (NADW), as well as the Antarctic Intermediate Water (AAIW). From WOCE.

Ekman pumping (see Fig. 5.3). The resulting intermediate water is located higher up in the water column than the deep water (but below the *surface waters*) because of its density characteristics.

We could also use temperature as well, as shown in Fig. 5.9. Note however visually the waters associated with NADW and AAIW is not as clear. The temperature signal however shows very cold *bottom* waters originating from the Antarctic, sinking to the bottom and filling up the very bottom of the ocean. This watermass is called the **Antarctic Bottom Water** (AABW), characterised by the very cold and somewhat fresher characteristic relative to the NADW (Fig. 5.8). This water is formed in the smaller seas of near the Antartic (e.g. Weddell and Ross seas), where it is subject to extremely cold atmospheric temperatures, leading to its characteristic cold signature.

So far we have used the not quite dynamically passive or conserved tracers, but they do do they job well enough[8]. We could of course use other tracers. Fig. 5.10 shows the same meridional sections as above but showing dissolved oxygen, and we have a rather strong signal that seems to be able to distinguish the NADW, AAIW and AABW. The oxygen content of the various watermasses could be attributed to the last time the water has been in contact with the atmosphere, as there is no resupply of oxygen if we are sufficiently deep below the ocean surface (below the *euphotic zone*, where there is not enough sunlight to power photosynthesis). The AAIW has higher oxygen content because it originates from the surface, which has

[8] One could argue salinity is expected to be slightly better than temperature as a tracking tracer because the diffusion associated with salinity should be smaller.

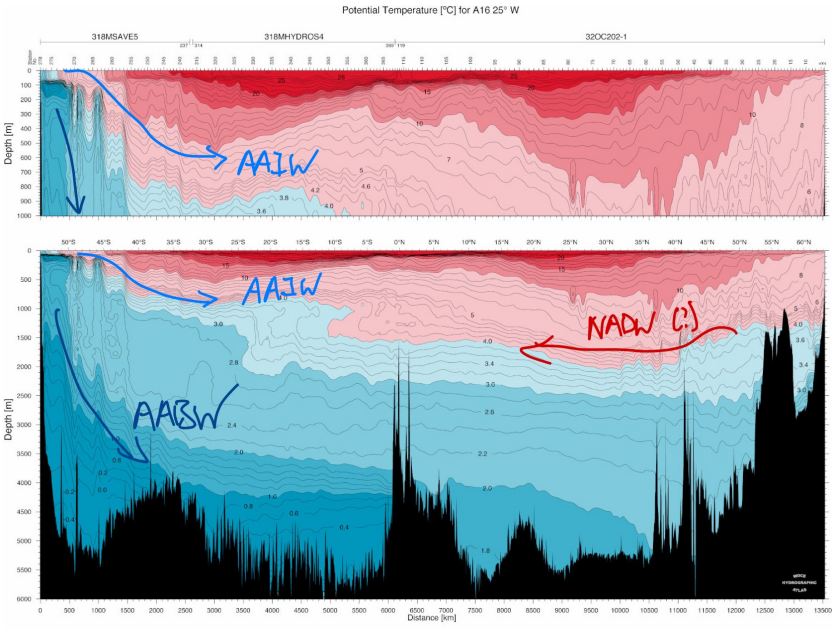


Figure 5.9: Potential temperature profile at 25°W, highlighting the Antarctic Bottom Water (AABW). From WOCE.

only 'recently' been subducted by Ekman downwelling. The bottom and deeper waters are also relatively high in oxygen, because the surface water carrying the oxygen sinks before the oxygen can really be used up by the organisms undergoing respiration, and at depths the oxygen usage is low because of a relative absence of life.

Anyhow, hopefully the above figures have illustrated that the AMOC follows a schematic like the one shown in Fig. 5.11. As a summary, the Gulf Stream brings warm and saline water from the tropics towards the North Pole. This water gets cooled at the higher latitudes and sinks, but maintaining the saline character. The cooler, saline and more dense water returns as the NADW, and transverses the Atlantic as deep water. The NADW gets upwelled in the Southern Ocean, which circumnavigates the globe, mixing accordingly with the fresher water from the Antarctic, losing its saline character and becoming lighter. Some of the Southern Ocean surface water (with higher oxygen content) returns polewards and is recycled a part of surface water, while some are cooled and subducted as slightly denser AAIW, maintaining its relatively higher oxygen content. On the other hand, some of the Southern Ocean water goes towards the South Pole, which then can get significantly cooled by the very cold atmosphere, becoming very dense and identified as the AABW, which sinks to the bottom of the ocean spreads along the seafloor.

There are other tracers that could be used (e.g. neutral density,

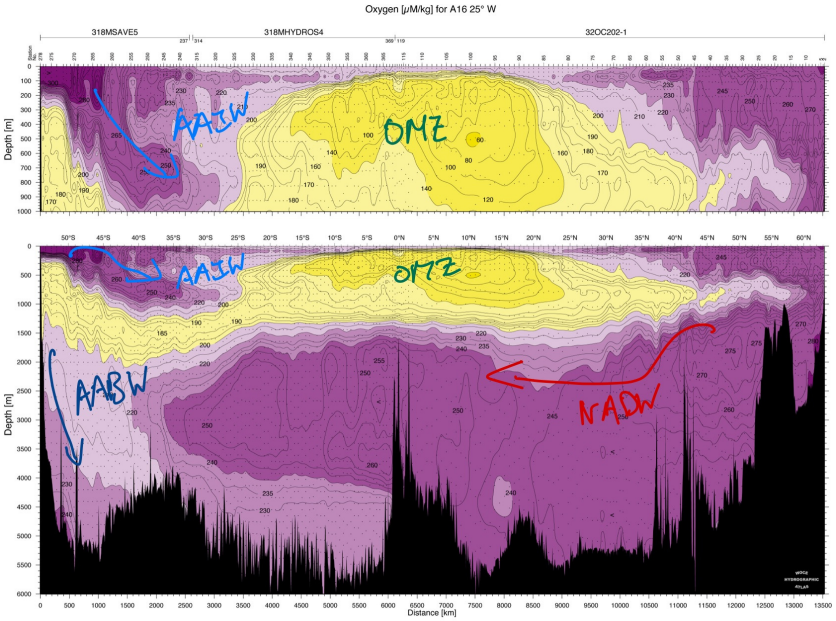


Figure 5.10: Dissolved oxygen concentration profile at 25°W, highlighting the Antarctic Intermediate Water (AAIW) high oxygen tongue at intermediate depths in the South. From WOCE. The oxygen lows are marked on as the OMZ (Oxygen Minimum Zone).

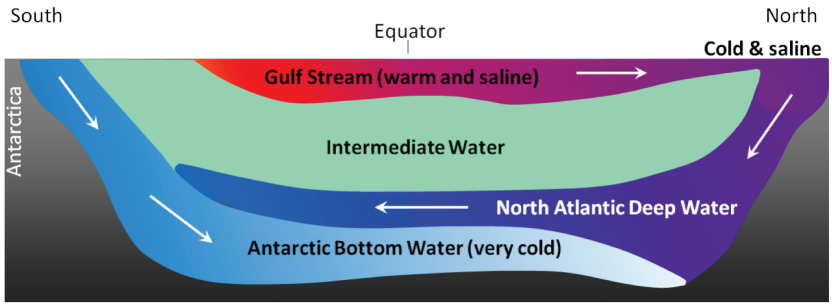


Figure 5.11: Schematic of the MOC and watermasses in the Atlantic. From Stephen Earle's open source textbook, Figure 18.17.

tritium, carbon isotopes etc.) for watermass identification; see for example the World Ocean Atlas meridional section gallery[9]. There are also many other watermasses around the globe that could be labelled, and different choices / combinations of tracers may be more suitable for different watermasses. The main focus here is on how these waters are formed, and how these waters are cycled around as part of the MOC, so we refer the reader to the book of Talley et al. [2011] for a much more comprehensive overview of watermass characteristics around the globe.

[9] www.ewoce.org/gallery/index.html

5.2.3 *How does water go down...?*

We tackle the physical schematics relating to downwelling and upwelling of the waters associated with the MOC, before we revisit the schematic of the global MOC at the end of the chapter. The two main questions that arise from the preceding discussion are (1) how are the deeper/bottom waters formed in the first place, and (2) how are the deeper/bottom waters subsequently lightened and 'recycled', if at all?

Recall that, unlike in the atmosphere, the ocean is effectively heated from above (cf. Fig. 2.4). The set up reinforces the density stratification of the upper ocean, and since flow is largely along rather than across isopycnals, by construction the physical set up is stacked against us having extended vertical / diapycnals transport of water. While there is Ekman downwelling, the vertical extent is rather limited. As such we need specialised conditions to have a significant vertical motion, in particular we need places to form waters dense enough to sink sufficiently deep (to depths greater than about 2000 m). Notably we require significant **watermass transformation** to occur, i.e. conditions to change the temperature and/or salinity characteristics of the watermass, which can happen when the water is subject to strong thermodynamical forcing from the atmosphere (but also near rivers and ice regions). Two schematics are given in Fig. 5.12 that shows how the atmosphere can work to create dense water classes. Intense cooling near the surface causes the water to be denser, and this *lifts* the isopycnals, until at some point there is static instability and significant sinking of the cold water. Similarly, evaporation can do the same thing, though normally this is combined a drop in the temperature to form the cooler and saline dense water.

In addition to a combination of intense cooling or evaporation (which restricts our attention to higher latitudes and the subtropics), the presence of bathymetric features to hold the denser water back before it floods the deeper parts of the ocean is highly desirable

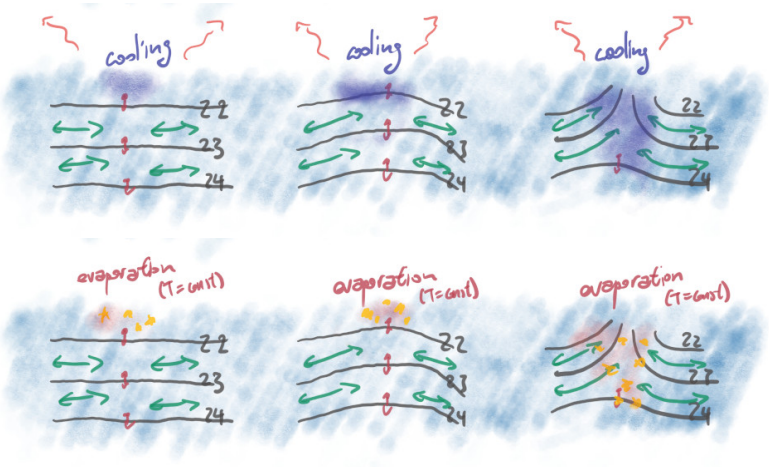


Figure 5.12: Watermass transformation by changing temperature and/or salinity. Marked on are representative isopycnals associated with different density classes.

(or even necessary), because if there isn't a sufficiently large pool of water, the water can undergo mixing as it sinks, thus limiting how deep it can sink. There are only a few places on Earth where the conditions are satisfied, notably the Lab Sea, the Weddell Sea and the Med Sea, as highlighted in Ch. 1.3.2 and 1.3.3. In the Lab Sea and the Weddell Sea, intense cooling occurs that leads to the forming of NADW and AABW respectively. In the Med Sea, the high evaporation and low precipitation leads to very saline waters, and occasionally cold bursts of continental air leads to a cooling of the waters, forming dense water and eventually joining the NADW. All three locations have bathymetric features holding the dense water back, before it slides down the continental slopes as **overflows**[10]. How deep the dense water overflow sinks depends on the water density, characteristics of the bathymetry in the region, and possibly also on the volume of the overflow water. If for example the slopes are fairly gentle, then the water slides down the slopes without necessarily mixing that much with the surroundings, maintaining its dense characteristic, and sinking to the deeper parts. If on the other hand the slopes are rough, then there could be significant mixing with the surrounding lighter waters, diluting the water of its dense characteristic before it has a chance to sink that much. Fig. 5.13 shows a schematic of this kind of process for the AABW, through transformation of the NADW in the southern regions of the Southern Ocean where significant cooling can occur (e.g. through gaps in the ice such as *polynyas*).

[10] Think underwater waterfalls.

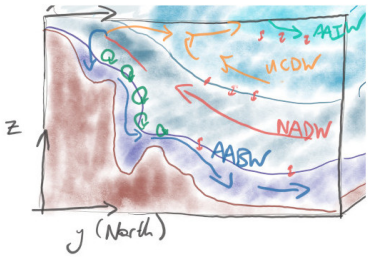


Figure 5.13: Schematic of watermass movement in the Southern Ocean, showing NADW, AABW, *Upper Circumpolar Deep Water* (UCDW) and AAIW. See also Fig. 5.3.

5.2.4 *How does water come up...?*

For the upwelling of deep water, lets talk about the NADW one first using Fig. 5.13. The NADW is upwelling is partly due to Ekman upwelling, but this is only part of the story. Again, since motion is largely along isopycnals, the NADW can come back up presumably because there is the Southern Ocean isopycnal tilt allowing the connection between the NADW with the atmosphere in the Southern Ocean via outcropping isopycnals. The tilting isopycnals primarily exist because there is a possibility for a strong ACC (cf. thermal wind shear relation), so one could wonder what happens if the Southern Ocean did not have open latitudes, or if we had different land configurations like in prehistoric times, and what kind of effect this could have on the overall Earth climate. We do not pursue this question here[11], apart from stating that a large portion of upwelling of deep water happens on the northern part of the Southern Ocean and correlated with where the wind stress curl is positive (see discussion in Ch. 5.1.1).

[11] A plug however for the work of Munday et al. [2015] studying the influence of Drake passage and Tasman gateway influence in an idealised numerical model though.

But what about the upwelling of *bottom* water? As mentioned very briefly in Fig. 1.4.2, we do have geothermal fluxes at the bottom that is heating the water from below, although it turns out this is not enough to explain the overall upwelling. The 'classical' argument considers a broad diffusive upwelling arising from diapyncal diffusivity, based on the now classic "abyssal recipes" paper of Munk [1966] (with a follow up in Munk and Wunsch [1998]). The argument is roughly that *we know how much water is going down, so what is the diffusivity we need to get the upwelling required to keep the overturning circulation going*? We may not need a diapycnal diffusivity that is too large, because remember we have a very large volume to play with. What was deduced is that we need a diapyncal diffusivity κ_d of around 10^{-4} m^2 s^{-1} throughout the ocean to maintain an overturning circulation (Munk [1966]), and is attributed to breaking *internal tides* (more on this in Ch. 6.3). Field measurements however find that generally we have $\kappa_d = O(10^{-5}$ m^2 s$^{-1})$, so we seem to have a mismatch. A follow up in Munk and Wunsch [1998] concluded something similar from a diffusivity point of view, but instead interpreted the desired value of $\kappa_d = O(10^{-4}$ m^2 s$^{-1})$ to be an average, i.e. there are regions where the diapycnal diffusivity is large and disproportionately important for upwelling (i.e. the *efficiency* matters). They also calculate the energy required to carry out the mixing, and find that we need around 2 *terrawatts* of power[12] to stir the water accordingly to get the upwelling we need to maintain a MOC, which should probably come from tides and wind driven mixing. An interesting point to note is that 'only' 2 TW of power is

[12] 1 TW = 10^{12} W. For reference, the largest power plant in the world generates around 10,000 MW = 10^{10} W = 0.01 TW of power.

needed to maintain a MOC that has a power of roughly 2 PW = 2000 TW (cf. Fig. 5.6).

That was a 'classic' picture for a long while, but science evolves and the broad diffusive view has been challenged somewhat recently (e.g. Ferrari et al. [2016], de Lavergne et al. [2016a,b, 2017]). Their arguments depend on the observation that the diapycnal diffusivity has a vertical structure (e.g. Fig. 5.14), and is largest near the top and bottom boundary layers (recall Ch. 3.4.2). The broad diffusive upwelling picture essentially doesn't really work for the bottom waters, and the argument is shown schematically in Fig. 5.15. If $\partial\kappa_d/\partial z > 0$, as in the case for the top water parcel, then imagine you have a blank parcel, but because of the sign of the gradient in κ_d, you mix more *lighter* water than *denser* water into the parcel, so the overall parcel lightens and upwells. However, as you go deeper toward the bottom, the gradient sign changes and $\partial\kappa_d/\partial z < 0$, as in the middle parcel, and with the same thought experiment you are forced to conclude that you mix more *denser* water into the blank parcel, so the water parcel has to *sink*!

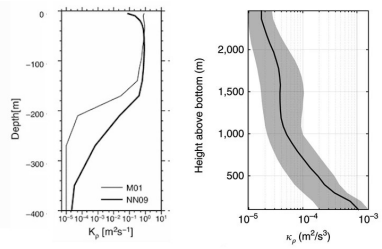


Figure 5.14: Observed κ_d in the upper ocean (left) and deep ocean (right). Figure taken from Watanabe & Hibiya (2013), *J. Phys. Oceanogr.* (left, their Fig. 5*d*) and Mashayek *et al.* (2017), *Nature Comms.* (right, their Fig. 2*c*).

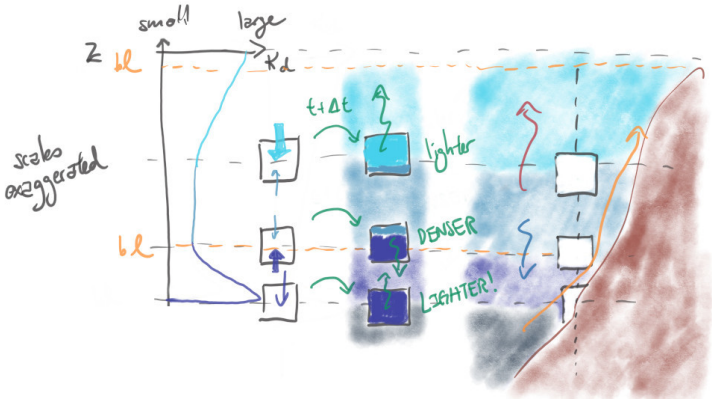


Figure 5.15: Schematic of the boundary intensified diffusive upwelling.

The way to get around this is if you were *close enough to the bottom boundary*. Then there is no water to mix in from the bottom, so the only water being mixed in is the water above, which has to lighter, so the water parcel gets lighter. What this implies is that we should have a *boundary intensified upwelling*. Furthermore, the geometry of seafloor distribution plays an important role in shaping the upwelling of the bottom waters. Some evidence for this is given in Fig. 5.16 using radiocarbon content; see de Lavergne et al. [2017] for more details.

If we are dealing with bottom waters, then the only thing that can really do mechanically stir the water to result in the mixing has to be something to do with *internal tides*, which we introduce in Ch. 6.3. Again, the discussion highlights the importance of dynamics and in particular small-scale dynamics in driving the large-scale MOC,

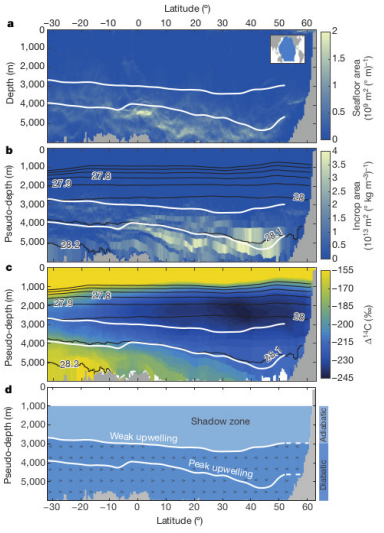


Figure 5.16: Observational evidence for *shadow zones* and boundary intensified upwelling. Taken from de Lavergne et al. [2017] (their Fig. 3).

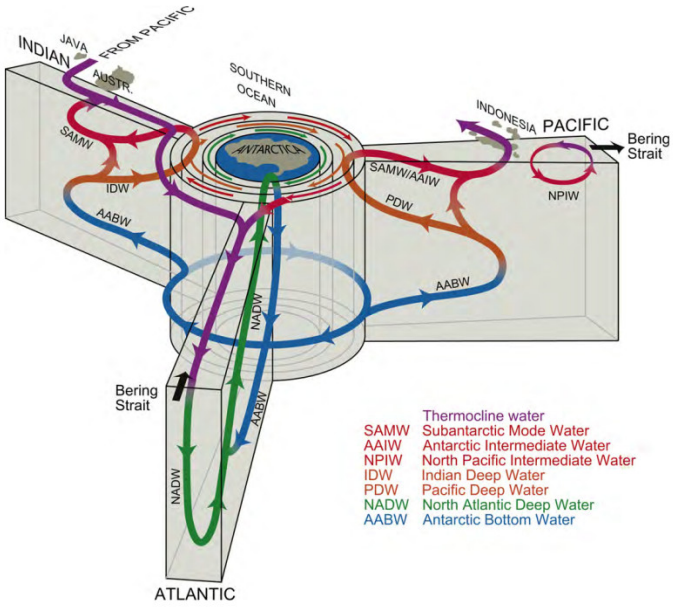


Figure 5.17: Schematic of the 3d MOC with watermass distributions. From Talley et al. [2011]; see more in their Fig. 14.11. Format after Arnold Gordon (1991)?.

which is the subject of the next chapter.

Summary and further reading

As advertised, instead of the two-dimensional schematic of the (A)MOC as in Fig. 5.11, the three-dimensional schematic of Fig. 5.17 is perhaps more comprehensive, highlighting the three-dimensional and circumpolar aspect of the global MOC. The AMOC aspects we talked about already, with the downwelling by watermass transformation, eventually upwelled in the Southern Ocean from Ekman upwelling, enabled by the outcropping isopycnals. The water either gets carried around the globe as part of the ACC, moves north and gets returned northwards as intermediate or surface waters in the basins, or moves south and gets transformed into AABW, sinking all the way to the bottom and filling up the ocean. This bottom water gets upwelled *along the boundaries*, supplying the deep waters in the basins, which then either gets upwelled again in the Southern Ocean, or through the broad diffusive upwelling towards the surface, but on a slow diffusive time-scale. The cycle repeats, and we have our global MOC, with various watermasses that could be identified via distinguishing characteristics, be they dynamical tracers (e.g. temperature, salinity, density), chemical tracers (e.g. dissolved oxygen, radiocarbon), or others.

As highlighted in this chapter, there are several places on the globe that are disproportionately important for the global MOC, such as

the sites for deep water formation (which are generally known), the locations of the upwelling (which is an active area of research), but perhaps no location is as central to the MOC as the Southern Ocean. Through the thermal wind shear relation, we have highlighted that the circumpolar circulation is fundamentally linked to the Southern Ocean stratification, and the circumpolar transport goes hand-in-hand with the tilting isopycnals in the Southern Ocean. The Southern Ocean is linked to the global ocean via the connecting isopycnals, it implies that if the Southern Ocean circulation changes, this can have consequences for the *global* stratification, which in turn affects the global MOC. Ongoing research questions include what sets the Southern Ocean circulation and stratification (cf. the arguments in Ch. 4): is it wind, buoyancy, bathymetry, and/or even eddies (or something else)? The works of Ferrari et al. [2014] and Jansen [2017] for example focus on buoyancy forcing and *ice* influences on the Southern Ocean and global stratification: ice formation can decrease the pycnocline depth, with associated decreases the AMOC strength, but more bottom water formation can lead to substantial carbon being carried into the bottom of the ocean. Wind controls on the Southern Ocean circulation was perhaps first explored in global circulation models in Toggweiler and Samuels [1995] and Toggweiler et al. [2006], and have been subsequently revisited again investigating the role of eddies (e.g. Munday et al. [2013], Mak et al. [2018]; see Fig. 5.18). Ongoing research updates what is regarded as 'standard' and 'classic' theories, and it may be that a few years down the line whatever is written here will require an update. We have also not talked about other aspects that control the MOC, and the role of the Southern Ocean and/or the MOC in biogeochemical cycles. We refer to reader to the books of Talley et al. [2011] and/or Williams and Follows [2011] as two possible avenue to look further.

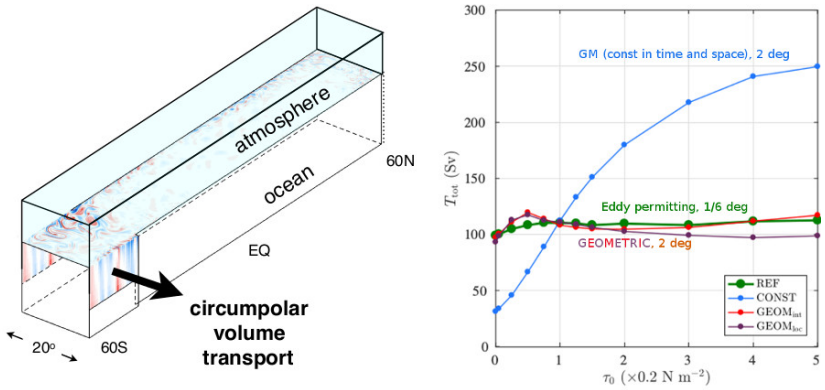


Figure 5.18: Idealised sector model from Munday et al. [2013] and results on ACC transport (related to Southern Ocean overturning) sensitivity with changes of wind depending on mesoscale parameterisation from Mak et al. [2018].

Again, one of the things that is highlighted is the fundamental

role of dynamics, and in particular smaller-scale phenomena, in controlling the large-scale structures via feedback mechanisms. Baroclinic eddies have been highlighted here and in Ch. 4 to be important for shaping the circulation and/or the stratification, and contributing to eddy form stress, which is important for transferring momentum put into the ocean by the atmosphere towards the bottom where it can be transferred out of the ocean. Upwelling of waters fundamentally depend on the (effective) diapycnal diffusivity, which in turn mostly depends on breaking *internal waves* arising from *tidal forcing* and on local shear instabilities contributing to the diapyncal mixing. Formation of deep waters require static instability. Mixing between atmopshere and ocean depends on the turbulent dynamics with the boundary layers. The next chapter looks at some of the smaller-scale dynamical features in a bit more qualitative detail.

Chapter exercises

1. here be tumbleweeds...

6 Dynamics

Hopefully the last few chapters have convinced you the importance of *small(er)-scale dynamics* in a multi-scale system such as the ocean. Motion at the large-scales can lead to small-scale motion via *instabilities*, and *waves* can be excited by these instabilities or other forms of forcing. To add to the mix, the small-scale motion can interact and lead to feedback onto the large-scale motions, modifying the things that generated the small-scales in the first place. This overall topic of *wave/eddy-mean interaction* is of theoretical interest as a fundamental problem in geophysical fluid dynamics, but with important practical consequences (e.g. *parameterisation of sub-grid physics* in numerical models used in predicting ocean/atmosphere weather and/or climate). Here we are only going to talk about how the mean (large-scale motion) generates the waves/eddies (small-scale motion), and only mention the feedbacks in passing.

6.1 Waves

6.1.1 Concepts

We start with *waves* first because we will be using them as building blocks for talking some of the *instabilities* later. A **wave** is just some periodic and/or oscillatory signal/motion. The physical properties of waves depend on the physics and forces associated with the system. Some examples are shown in Fig. 6.1, and these are:

- *Alfv́en waves*, which are waves travelling along magnetic field lines with the Lorentz force as the restoring force;
- *gravitational waves*, which are ripples in spacetime, recently reported by the Laser Interferometer Gravitational-Wave Observatory (LIGO) and the Virgo interferometer in 2016 ref;
- *electromagnetic waves*, already encountered in Ch. 2.2.1, with visible light as a special case;
- *sound waves*, propagating through colliding particles, through the

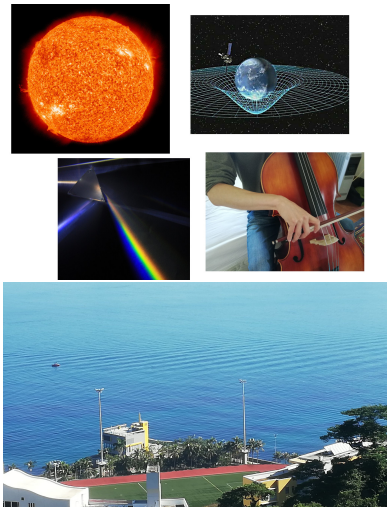


Figure 6.1: Example of some wave and systems supporting waves (sun, spacetime, electromagnetic waves, sound waves, surface gravity waves). All figures from Wikipedia except the cello one and gravity waves one (personal photos).

solid cello soundboard of the cello which, in turn, vibrates the surrounding air before reaching our ears;

- surface gravity waves, propagating on the atmosphere-ocean interface, restored by buoyancy forces.

Generically, a wave can be described by features labelled in Fig. 6.2. If we consider a wave leading to a *displacement* (relative to the dashed line), then the maximums and minimums of the displacements are called the **peaks** and **troughs** (or anything else really describing a bump and a depression). If this waves is not travelling but oscillating about some equilibrium (e.g. the dashed line), then it is a **standing wave**, and **nodes** are where the displacement is zero, while the **anti-nodes** are where the displacement is maximum.

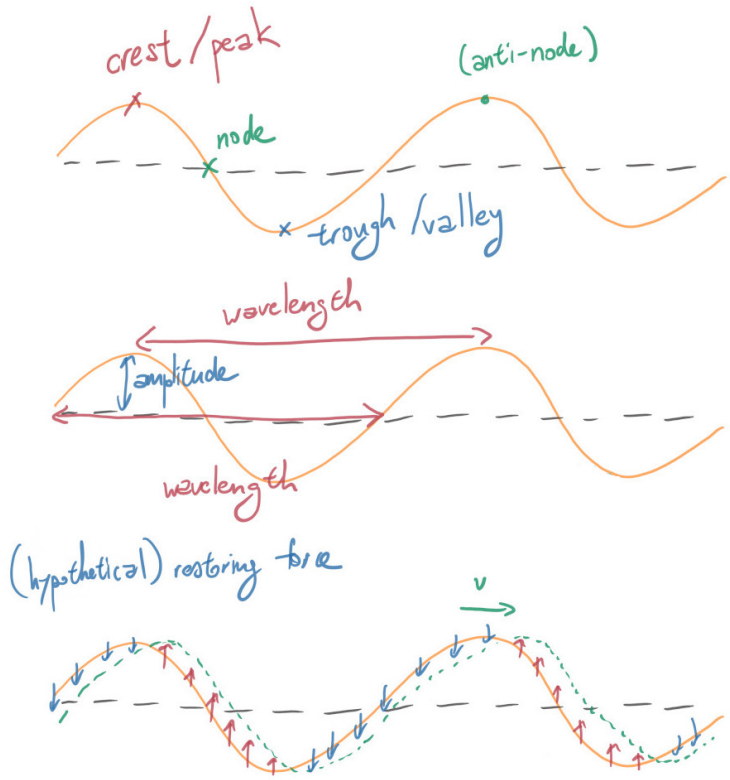


Figure 6.2: Schematic of wave features. A transverse wave is drawn here for easier illustration.

The magnitude of the wave is called the **amplitude**[1]. The **wavelength** λ ('lambda') measures the spatial periodicity, while the **frequency** γ ('gamma') measures the temporal periodicity[2], and is measured in Hz (1 *Hertz* = 1 oscillation s^{-1}). If a wave propagates, then the speed it propagates at is the **phase speed**, denoted v on the diagram but will be generally denoted c_p; think of this as the velocity

[1] The amplitude then becomes measure dependent. Here we use the displacement, but we don't necessarily have to; wave energy could be another one.

[2] Normally it's f but f I'm saving for Coriolis parameter.

that the peaks propagate at (or any other point of reference on the wave).

As the wave evolves the **phase** changes: the phase is the wave's position in the wave cycle, usually measured between 0° to 360° (0 to 2π radians) or $-180°$ to 180° ($-\pi$ to π radians) depending on the convention. Fig. 6.3 shows the orange and green waves that are **in phase** (the peaks and troughs coincide with one another, zero phase shift) and **in anti-phase** (the peaks of one overlap with the troughs of the other, and vice-versa, 180°, π or half a wavelength phase-shift). The waves are in quadrature if they are $\pm 90°$ ($\pm\pi/2$ radians) or a quarter of a wavelength out of phase. The waves themselves can **interfere**, i.e. you can add them together to get a collective effect. The in phase and and in anti-phase cases here lead to **constructive** and **destructive interference** respectively.

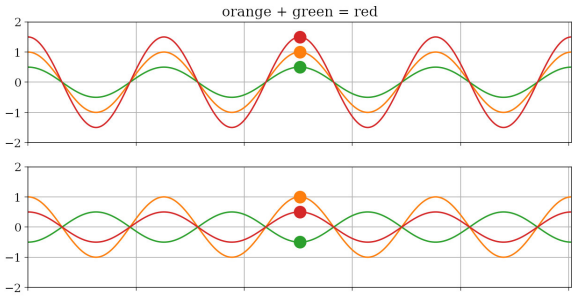


Figure 6.3: Interference of waves, where the red wave is the result of adding the orange and the green wave. For waves in phase and in anti-phase.

To link this up to some symbols, taking η as the displacement, a wave can normally be described as

$$\eta \sim A\cos\left(2\pi\left(\frac{x}{\lambda} - \gamma t\right)\right) = A\cos\left(\frac{2\pi}{\lambda}\left(x - c_p t\right)\right), \qquad (6.1)$$

where (x, t) are the space and temporal co-ordinate, A here would be the amplitude, and the frequency and wavelength is related by

$$\gamma = \frac{c_p}{\lambda}. \qquad (6.2)$$

If we had $\eta \sim A\cos(\ldots + \theta_0)$, then we would have a phase shift of θ_0.

It is customary to get rid of the 2π factor by employing the **wavenumber** k and **angular frequency** ω, defined as

$$k = \frac{2\pi}{\lambda}, \qquad \omega = 2\pi\gamma, \qquad (6.3)$$

so that the wave is written as

$$\eta \sim A\cos\left(kx - \omega t\right). \qquad (6.4)$$

The wavenumber you can sort of think of as how many waves fit in a box, although as defined it is only an integer if the box has length

2π; you should just think of it as an inverse measure of wavelength. In Fig. 6.2 and Fig. 6.3, we would have $k = 2$ and $k = 4$ respectively (because there are two and four waves that fit in a box, assuming the box is of length 2π). I will almost exclusively be using wavenumber throughout the text, with the understanding that the wavenumber is inversely proportional to wavelength ($k \sim \lambda^{-1}$), and wavelength itself is related to frequency ($\lambda \sim \gamma^{-1}$).

If we are instead dealing with multiple spatial dimensions, then we have

$$\eta \sim A \cos\left(\boldsymbol{k} \cdot \boldsymbol{x} - \omega t\right), \tag{6.5}$$

where $\boldsymbol{x} = (x, y, z)$ and $\boldsymbol{k} = (k_x, k_y, k_z)$ is the **wavevector**[3].

[3] More generally, we would write it as $\eta \sim \mathrm{Re}[A\mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{x}-\omega t)}]$, where $\mathrm{i} = \sqrt{-1}$ is the imaginary number, and A is in principle a complex number.

Note that, in general, we have

$$\omega = \mathcal{F}(\boldsymbol{k}, \ldots), \tag{6.6}$$

for some function $\mathcal{F}$, and this relation is known as the **dispersion relation**. The dispersion relation encodes the wave's frequency, wavelength, how it propagates, and is dependent on the physics supporting the existence of the wave. For example,

$$\omega = B_0 k, \quad \omega = -\frac{\beta}{k}, \quad \omega = \sqrt{gk}, \quad \omega = \frac{\hbar k^2}{2m}$$

are respectively the dispersion relations for (shear) Alfvén waves, Rossby waves, deep water waves, and de Broglie or matter waves in one space dimension.

In one space dimension, the **phase speed** and the **group velocity** are defined as

$$c_p = \frac{\omega}{k}, \qquad c_g = \frac{\partial \omega}{\partial k}. \tag{6.7}$$

A wave is said to be **non-dispersive** if the group velocity and the phase speed is equal. For example, the shear Alfvén wave example above is the only non-dispersive wave type out of the four examples.

In higher space dimensions, the phase speeds and the group velocity would be

$$c_{p,x} = \frac{\omega}{k_x}, \quad c_{p,y} = \frac{\omega}{k_y}, \quad c_{p,z} = \frac{\omega}{k_z}, \qquad \boldsymbol{c}_g = \nabla_{\boldsymbol{k}}\omega = \begin{pmatrix} \partial\omega/\partial k_x \\ \partial\omega/\partial k_y \\ \partial\omega/\partial k_z \end{pmatrix}. \tag{6.8}$$

The phase speed can be thought of as describing how individual waves propagate *in a particular direction*, tracking for example the peak (or the phase) of these waves. Note however the important distinction that the phase speeds are *not* components of an analogous **phase velocity**[4]: the phase of a wave propagates with speed $\omega/|\boldsymbol{k}|$ in the direction of the wavevector $\boldsymbol{k}$, but the individual phase speeds

[4] See Ch. 5 Appendix of Vallis [2006] for an explanation.

such as $c_{p,x}$ is not equal to $\omega/|\boldsymbol{k}|$. The group velocity describes how a group of waves superimposed on each other propagates, such as that described in the left panel of Fig. 6.4. The two individual waves interfere in this case to form a **wavepacket** where the *envelope* has a much lower frequency. At a later time, we note from the markers that the individual components themselves propagate to the left, while the wavepacket has propagated to the right. The group velocity is normally the one that is of physical interest, as this is the velocity that quantities carried by the waves (e.g. energy, momentum, information) travels at. Another example that highlights the difference between waves and wavepackets is that if you look at the sea and try to track the waves by looking at the crests, the crests and so the individual wave might be doing one thing, but the magnitude of the crests probably oscillates as well, indicating a collective behaviour as displayed by wavepackets.

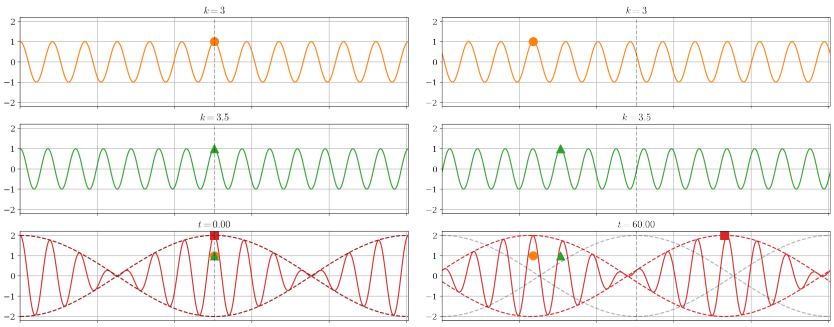


Figure 6.4: Example using one dimensional Rossby waves with $\omega = -\beta/k$. Showing how individual waves and how the interference can result in a wavepacket, and demonstrating the phase propagation and the group velocity are really different quantities. The circle and triangle marker tracks the crests of the individual waves, but the rectangular marker tracks the crest of the wavepacket. See `rossby_propagation.mp4` for an animation.

Changes in the phase and/or group speeds lead to various wave phenomena illustrated in Fig. 6.5, namely:

- **refraction**, which refers to how waves change direction as it propagates into a medium with slightly different properties (think of a car going from a road to mud at an angle, the mud has more ''grip' on the car, the car slows, and the car has to turn and change its direction of travel);

- **reflection**, where the wave could hit a boundary and gets reflected back, although some of it could also be **transmitted** depending on the medium of interest (e.g. a non-solid sponge layer);

- **diffraction**, leading to the spread of the wave as it propagates through a gap with width less than the wavelength (if the gap is big enough the wave just propagates through), arising because of secondary waves being generated by encountering an obstacle results in interference patterns.

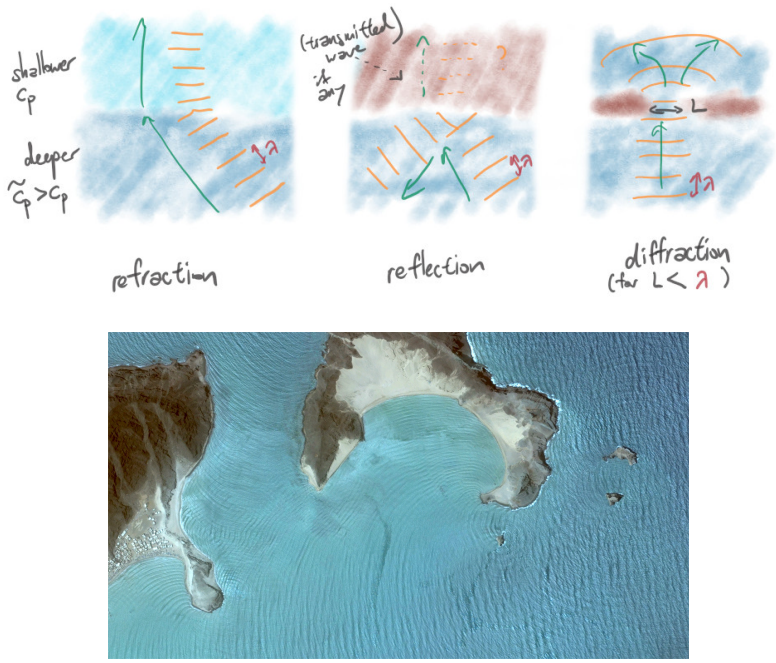


Figure 6.5: Schematic of refraction, reflection (and transmission), and diffraction, nominally using monochromatic (i.e. one choice of k) surface gravity wave as an example. The orange lines are phase lines (e.g. think wave crests). Bottom shows picture of (presumably non-monochromatic) waves over the Arabian sea, taken from `www.earthglance.com`, post 133835790223.

Fig. 6.5 also show surface gravity waves in the Arabian sea. The changes in water depth and presence of land obstacles leads to the various wave patterns, arising from the aforementioned phenomena associated with refraction, reflection and diffraction.

6.1.2 *Deriving dispersion relations*

The main takeaway from the definitions so far is that the dispersion relation effectively encodes almost everything that distinguishes a type of wave from another. Here I provide a quick example here on how to derive the dispersion relation from the governing equations, mostly as demonstration some of the mathematical gymnastics involved; it is supposed to be a quick demonstration, so I am going to use whatever I am familiar with with the main aim to keep this short. For this demonstration we work in a two-dimensional horizontal β-plane and start with the vorticity equation, i.e. (stated without proof)

$$\frac{\partial \zeta}{\partial t} + u\frac{\partial \zeta}{\partial x} + v\frac{\partial \zeta}{\partial y} + \beta v = 0. \tag{6.9}$$

Because of the assumption that $\nabla \cdot \boldsymbol{u} = 0$, we can define a *streamfunction* ψ such that $\boldsymbol{u} = \boldsymbol{e}_z \times \nabla\psi$, and additionally noting that $\zeta = -\nabla^2\psi$, where the Laplacian operator is defined around Eq. (3.14). The

equation of interest can then be written as

$$\frac{\partial \zeta}{\partial t} + u\frac{\partial \zeta}{\partial x} + \frac{\partial \psi}{\partial x}\frac{\partial \zeta}{\partial y} + \beta\frac{\partial \psi}{\partial x} = 0. \tag{6.10}$$

I am leaving the u untouched for the moment for reasons that should become apparent in a bit.

Now, convince yourself that $(u, v) = (U_0, 0)$, so that $\Psi_0 = -U_0 y$ (because $u = -\partial\psi/\partial y = U_0$) and $\zeta = 0$ is a solution of the equation (you get $0 = 0$). Then you consider a *perturbation* to this solution, i.e. you kick it a bit and see how the system responds (like hitting a drum and seeing what sound results). Mathematically, you take

$$\psi(x, y) = -U_0 y + \epsilon\tilde{\psi}(x, y), \qquad \zeta(x, y) = 0 + \epsilon\tilde{\zeta}(x, y), \tag{6.11}$$

where $\epsilon \ll 1$ is a small number, and substitute it into the equation. If you do this and let the dust settle after the algebra, what you are going to get are the following:

- a collection of terms not multiplied by ϵ, but these are zero because they only involve Ψ_0 and U_0, which we said is already a solution;
- some terms that are multiplied by one factor of ϵ;
- some terms that are multiplied by more than one factor of ϵ, but since $\epsilon \ll 1$, $\epsilon^2 \ll \epsilon \ll 1$, so we throw those terms away because they are really small.

This is what is known as **linearisation** (because I am only keeping terms that are linear in ϵ). If you do this for the above equations you get

$$\frac{\partial \tilde{\zeta}}{\partial t} + U_0\frac{\partial \tilde{\zeta}}{\partial x} + \beta\frac{\partial \tilde{\psi}}{\partial x} = 0. \tag{6.12}$$

The next step is to propose a **waveform**. This equation is what is known as a constant coefficient partial differential equation (the derivatives are hitting the things we want a solution for, and are multiplied here by 1, U_0 and β respectively), one possible solution is to consider[5]

$$\tilde{\psi} = A\mathrm{e}^{\mathrm{i}(k_x x + k_y y - \omega t)}, \tag{6.13}$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary number, (k_x, k_y) are the x and y wavenumbers, A is some (complex) number that is assumed to be fixed in time. The idea is that, since we have a linear equation, then we can add different solutions together to generate other solutions (the *principle of superposition*) to the equation, and it turns out the dispersion relation tell as which combinations of choices are the valid solutions.

[5] I am going to be using complex numbers, and the real part should be taken at the appropriate points.

Imaginary numbers essentially obey the same rules as numbers, so if we shove this into the linearised equations, we have

$$\frac{\partial}{\partial t} \to -\mathrm{i}\omega, \quad \frac{\partial}{\partial x} \to \mathrm{i}k_x, \quad \tilde{\zeta} = \nabla^2\tilde{\psi} = -(k_x^2 + k_y^2)\tilde{\psi}, \tag{6.14}$$

then you can convince yourself that the linearised equation becomes

$$\mathrm{i}\left[\omega(k_x^2 + k_y^2) - U_0 k_x (k_x^2 + k_y^2) + \beta k_x\right]\tilde{\psi} = 0. \tag{6.15}$$

Notice that the waveform basically survives and can be factored out accordingly. Now, i is not zero, so either $\tilde{\psi}$ or the stuff in the square brackets is zero, but $\tilde{\psi}$ cannot be zero because otherwise you have no perturbation in the first place, so it has to be things in the square bracket that is zero. Once you rearrange to make ω the subject of the equation you get

$$\omega = U_0 k_x - \frac{\beta k_x}{k_x^2 + k_y^2}. \tag{6.16}$$

This is the dispersion equation for β-plane *Rossby waves* that we will talk about in more detail later.

The handle you turn to get wave solutions may be summarised as follows: (1) get your equations; (2) find a *basic state* solution of the equations; (3) add a perturbation to the basic state and linearise; (4) propose an applicable waveform in the linearised equation for the perturbation; (5) do some maths, and get $\omega = \ldots$; (6) repeat for other systems until you get bored. Using complex numbers is slightly more preferable to using sines and cosines usually because the derivatives of the exponential functions e is slightly easier to deal with than derivatives of sines and cosines[6].

[6] And they are really the same things via the identity $\mathrm{e}^{\mathrm{i}\theta} = \cos(\theta) + \mathrm{i}\sin(\theta)$ for some θ.

6.1.3 *Gravity waves*

From now on we just state without proof the dispersion relations accordingly, and leave it to the reader to derive them if they so wish. We will also assume the amplitude of the waves is a constant in time, i.e. these waves are *neutral* and satisfy linear dynamics. The neutral aspect we re-examine in Ch. 6.2, but we do not touch on *nonlinear waves* and *wave turbulence* here.

Gravity waves are where the restoring force is buoyancy, and here in this section we are really talking about **surface gravity waves**, which are the ones you see on the surface of the ocean. Starting from what are known as the *Euler equations* (no viscosity, rotation, forcing), assuming $\rho = \rho_0 =$ constant and considering only one horizontal dimension setting for simplicity, it may be shown that the dispersion relation is

$$\omega = \pm\sqrt{gk\tanh(kH)}, \qquad \tanh(x) = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-x}} \tag{6.17}$$

where g has the usual meaning, k is the wavenumber, and H is the water depth. Note that we are allowed to have $\omega < 0$ (because e.g. $\cos(-\omega t)$ is still well-defined), but for simplicity we will talk about the positive branch. There are two limits that are of interest:

1. **deep water waves**, where $kH \gg 1$ so $\tanh(kH) \sim 1$, and that

$$\omega = \sqrt{gk}. \tag{6.18}$$

   The dispersion relation tells us that when the water depth is sufficiently large, the waves basically don't feel the bottom, and the wave characteristics only care about the magnitude of g, and shorter waves (i.e. those with $k \gg 1$) have a higher frequency. Upon working out the phase speeds, it is however the long waves (i.e. those with $k \ll 1$) that have the largest phase speeds.

2. **shallow water waves**, where $kH \ll 1$, and in this instance one can show that $\tanh(kH) \sim kH + O((kH)^3)$, so that

$$\omega = \sqrt{gk \times kH} = k\sqrt{gH}. \tag{6.19}$$

   The angular frequency in this case is directly proportional to the wavenumber. Shallow water waves propagate *slower* in shallow waters, since $c_p \sim \sqrt{H}$; this is partly responsible for breaking of water waves as the waves propagate towards the beach; see Chapter exercise. Notice also that shallow water waves are non-dispersive, i.e. $c_p = c_g$.

Fig. 6.6 shows ω, c_p and c_g (columns) of the deep, general and shallow water waves (rows), as a function of k and H, with slightly different choices of H depending on whether we are dealing with deep, shallow or shallow waves. If we instead talk about the negative branch, we flip the signs on the velocities, so we have a wave that propagates in the other direction but the properties are otherwise exactly the same.

A few features of note are that:

- the contours of the properties for deep water waves are vertical i.e. constant with depth, which is entirely consistent with the fact that the associated ω for deep water waves is depth-independent;
- the intermediate wave properties shows a transition between the deep and shallow regime, as it should;
- shallow water waves have c_p and c_g are exactly the same, and k-independent, as expected.

Technical note is that 'deep' and 'shallow' really refers to the size of kH rather than H itself. For example, it is easier to be 'deep' if

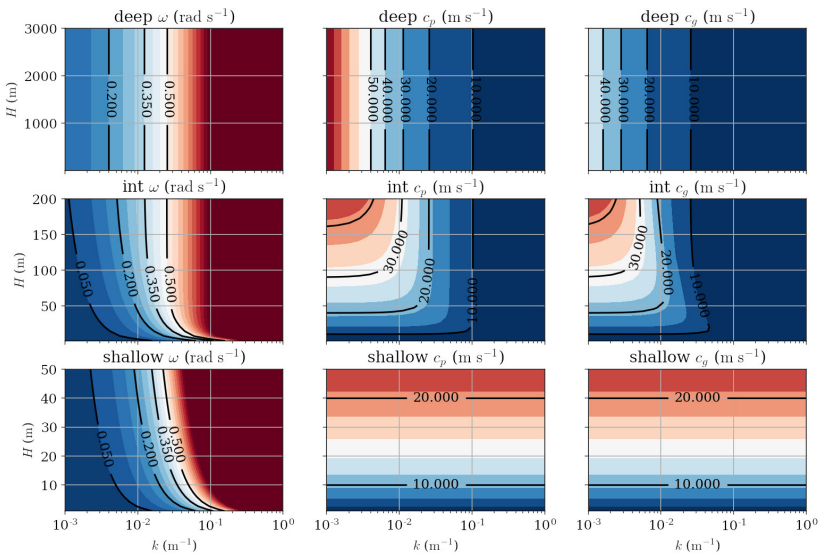


Figure 6.6: Water wave ω, c_p and c_g over (k,H) space, with k shown on a log axis. k chosen so wavelengths are roughly between 50 m to 5 km (recall $k = 2\pi/\lambda$). Also note the transitions from shallow to intermediate to deep are really to do with $kH \sim H/\lambda$. See `waves.ipynb`.

$k \gg 1$, i.e. short waves care less about the bathymetry than long waves.

6.1.4 *Inertia-gravity waves*

If we have rotation, then the wave characteristics are modified somewhat. Taking instead the two dimensional *shallow water equations* (partly because we now have the Coriolis effect), and taking the f-plane (cf. β-plane but with $f = f_0$ only) approximation, the dispersion relation for **inertia-gravity waves** (or sometimes **Poincaré waves**) is given by

$$\omega = \pm\sqrt{f_0^2 + gH(k_x^2 + k_y^2)}. \tag{6.20}$$

The restoring forces here are now buoyancy and rotation effects[7]. The dispersion relations bears resemblance to the dispersion relation for shallow gravity waves given by Eq. (6.19), because we employed the shallow water system in deriving this dispersion relation. Indeed, if $f_0 \ll gH(k_x^2 + k_y^2)$, i.e. if rotation is 'weak', then we essentially recover the shallow water gravity waves.

On the other hand, if $f_0 \gg gH(k_x^2 + k_y^2)$, i.e. if rotation is 'strong', then we end up with $\omega = f_0$, which are called **inertial waves** or **inertial oscillations** (mentioned in passing in Ch. 3.2.2), where the restoring of the wave is from the Coriolis effect. Pure inertial waves themselves are rather special as their phase propagation not in the same direction as the group propagation; we revisit this later when talking about *internal waves*. Inertial waves can arise generally when a

[7] Again I am maintaining that Coriolis effect is not really a force.

system is rotating, and occasionally been invoked as a mechanism in planetary sciences leading to *enhanced dissipation* through the presences of *wave attractors* (e.g. Fig. 6.7), arising because of a combination of domain geometry and wave propagation characteristics. We won't say too much about inertial waves here because waves in the ocean generically are more like gravity waves that are modified somewhat by rotation, i.e. buoyancy effects tend to be dominant.

The wave characteristics are mostly essentially shifts from that of Fig. 6.6 for ocean relevant parameters (but note that ω is now bounded below by f_0) so we omit the analogous diagram. One thing to note again is that whether rotation is 'strong' or 'weak' is relative to $gH(k_x^2 + k_y^2)$, so the scale of the wave matters: small-scale waves with large k_x and/or k_y care more about buoyancy and less about rotation, and vice-versa. From the dispersion relation Eq. (6.20), note that we can construct the length-scale (convince yourself this is a length scale)

$$L_d = \frac{\sqrt{gH}}{f_0}, \tag{6.21}$$

which is the shallow water version of the **Rossby deformation radius**, the horizontal length-scale that roughly delineates the regime where buoyancy is more important ($L < L_d$) and where rotational effects are more important ($L > L_d$). Note that this length-scale depends on the depth, i.e. in regions with larger depths require larger horizontal length-scales for buoyancy to not matter (cf. aspect ratio H/L). We give estimates of the deformation radius in the next subsection when we talk about the stratified analogue.

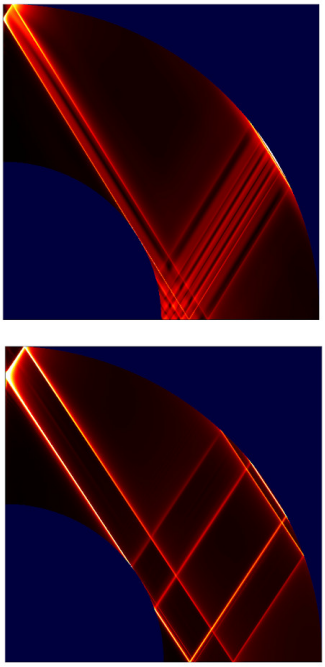


Figure 6.7: Inertial wave attractors in a homogeneous planetary interior at different tidal forcings. From Ogilvie (2009, *Mon. Not. Royal. Astro. Soc*).

6.1.5 *Internal waves*

Instead of waves with a signal on the ocean surface we could also have waves that have a signal in the interior of the ocean, restored accordingly by the Coriolis effect and/or buoyancy effects[8]. These would be *internal inertia-gravity waves*, though I am just going to refer to them as **internal waves** here.

A useful quantity to define before we go through the dispersion relation is the **Brunt–Väisälä frequency** or the **buoyancy frequency**

$$N = \sqrt{-\frac{g}{\rho_0}\frac{\partial \rho}{\partial z}}. \tag{6.22}$$

Normally it is N^2 that is talked about[9], and if $N^2 > 0$, $\partial\rho/\partial z < 0$, i.e. we have a stable stratification. By the analogous argument, $N^2 < 0$ implies unstable stratification, and we should expect overturns in the vertical. N^2 or N is perhaps a more useful quantity to plot when we are dealing with dynamics. Fig. 6.8 shows σ_0 (potential density

[8] If you consider the atmosphere and ocean as one fluid system, then in a sense you can think of surface gravity waves as waves on a buoyancy interface that we happen to call the ocean surface.

[9] Because then we don't have to deal with complex numbers arising from taking a negative square root, though the fact N is complex implies we have instability; see chapter exercise.

referenced to sea surface, see Ch. 2), N and N^2 at two different locations, and notice that where N and N^2 are the largest effectively corresponds to where the pycnocline is (because remember the pycnocline is defined as the region where the density changes the largest). On the other hand, N^2 and N are small near the ocean surface, because mixing is more intense near the surface boundary layer and thus erodes vertical density gradients, and in the interior, because the density stratification is small (after we have removed the pressure contributions).

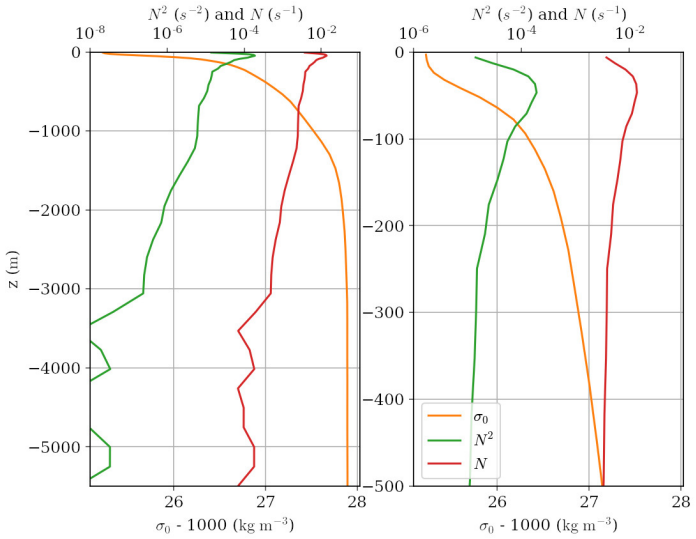


Figure 6.8: σ_0 (orange) and the associated N^2 and N (green and red) at two different locations. $N^2 \ll 1$ means weakly stratified (weak density gradients), whilst $N^2 < 0$ shows unstable stratification (none in this case). See `plot_eos.ipynb`.

One thing to note before we proceed. In the ocean, because we have a strong stratification, and partly because of the aspect ratio, vertical motion is expected to be suppressed relative to the horizontal motion. With this in mind, the vertical wavenumber k_z is expected to be much *larger* than the horizontal wavenumbers, i.e. vertical length-scale is small, at least relative to the horiziontal length-scale. So, working in a two-dimensional vertical slice (x, z) for simplicity, the dispersion for internal waves is given by

$$\omega = \pm\sqrt{\frac{f_0^2 k_z^2 + N^2 k_x^2}{k_x^2 + k_z^2}} \approx \pm\sqrt{f_0^2 + \frac{N^2 k_x^2}{k_z^2}}, \tag{6.23}$$

where the latter approximation comes from taking $k_z^2 \gg k_x^2$. In the ocean, we have N/f_0 typically around 10 to 100, so it is buoyancy that has the more significant effect (of course subject to the choice of k_x/k_z), so internal waves in the ocean are again really internal gravity waves that are influenced by rotation, rather than the other way round.

Some things we can get immediately from the dispersion relation:

- if you take $f_0 = 0$, or if you start from the unapproximated form of Eq. (6.23) and take $k_z = 0$ (so no vertical variation), then you get $\omega \sim N$, so internal gravity waves oscillate at the buoyancy frequency (hence the naming in the first place);
- if $k_x \ll 1$, i.e. long horizontal waves, rotation is more important, which we would expect (e.g. the Rossby number associated with the larger horizontal length scale should be smaller);
- from the unapproximated form of Eq. (6.23), noting that $1 \geq k_{x,z}^2/(k_x^2 + k_z^2) \geq 0$, and because $N/f_0 > 0$ in the ocean, we have $N \geq \omega \geq f_0$, i.e. the frequency of internal waves are *bounded* between the buoyancy frequency N and what is really the inertial frequency f_0;
- no such upper bound exists for surface gravity waves as seen in Eq. (6.20) (there is a lower bound by the inertial frequency), and we can also conclude that internal waves have a significantly *lower* frequency than the analogous surface waves, since internal waves can maximally have $\omega \sim N$ (with $N \approx 10^{-2}$ s^{-1} for the ocean maybe), but for surface waves we have $\omega \sim \sqrt{gH} \gg 1$ for $H \gg 1$ ($H = 100$ m leads to $\omega = 10^2 \gg 10^{-2}$).

If we take the approximated form of Eq. (6.23), we can derive and numerically plot out the characteristics associated with waves now that we have the dispersion relation. Fig. 6.9 shows ω, the phase speeds $c_{p,x}$, $c_{p,z}$, and the components of the group velocity $c_{g,x}$ and $c_{g,z}$. The diagram actually shows the $\log_{10}$ of the quantity, and the numbers show the exponents, i.e. the number a in $c_{p,x} = 10^a$ (recalling that $\log_{10} c_{p,x} = \log_{10} 10^a = a \log_{10} 10 = a$), because you basically can't see anything overly useful if it was on a linear scale. The plot for ω (top-left) does show that the frequency is bounded between f_0 and N (lighter shades correspond to values closer to zero), but I've drawn on some artificial boundaries to roughly denote the value of f_0 and N. The thing to really note is that these waves have phase and group speeds that are much *slower* compared to the surface waves: remember numbers shown in Fig. 6.9 are e.g. $10^{-1} = 0.1$ m s^{-1}, compared to typically 10^2 in Fig. 6.6. Again, if we take the negative branch, we get an extra minus sign on the velocities, the wave propagates in the other direction but the functional dependence is otherwise essentially unchanged.

The slightly more interesting aspect about internal waves is numerically demonstrated in Fig. 6.10, where we plot the angle the phase *velocity* makes with the horizontal x-axis, the angle the group

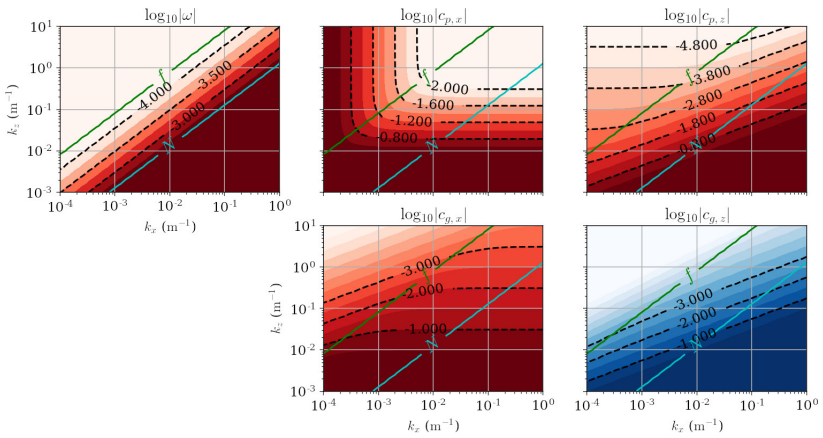


Figure 6.9: Internal waves (with the $k_z \gg k_x$ approximation) ω, $c_{p,x}$, $c_{p,y}$, $c_{g,x}$ and $c_{p,y}$ as a log-log plot in (k_x, k_z) space, with $f = 5 \times 10^{-5}$ and $N = 3 \times 10^{-3}$ (oceanic relevant values). The contours denote the exponent x of $|10^x|$ and the colour shading denotes the sign (more blue = more negative *actual* values rather than exponents, more red = more positive *actual* values rather than exponents); since k_x and k_z is chosen to be positive, everything except $c_{g,z}$ is positive. Contours of f and N plotted with an offset plotted to show the boundary beyond which everything is either gravity waves or inertial oscillations. See `waves.ipynb`.

velocity $\boldsymbol{c}_g$ makes with the horizontal x-axis, and the cosine of the difference between those two angles (all given in radians). The main result is the last one, which is essentially showing zeros[10]. If $\cos\theta = 0$, then this means $\theta = \pm\pi/2 = \pm 90°$, i.e. the group velocity is perpendicular to the phase *velocity*. Since the phase *velocity* is perpendicular to the phase lines, this implies the group velocity is *along* phase lines, as shown in Fig. 6.11. Indeed, since the phases propagate in the direction of the wavevector $\boldsymbol{k}$, it can be shown analytically that $\boldsymbol{k}/|\boldsymbol{k}| \cdot \boldsymbol{c}_g = 0$.

[10] Numerically these are not quite zeros because 'zeros' on a computer is usually only up the the machine accuracy. The numbers shown are on the order of 10^{-16}, so these are machine zeros.

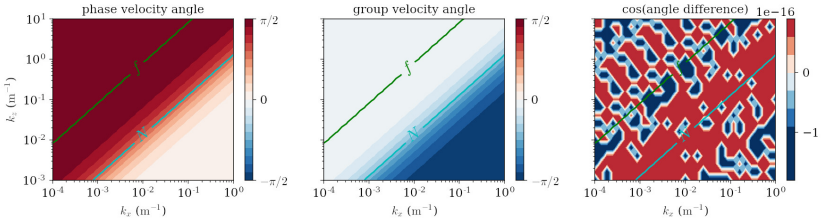


Figure 6.10: Internal waves (with the $k_z \gg k_x$ approximation) phase velocity angles and group velocity $\boldsymbol{c}_g$ angles (in radians, relative to the horizontal, and note $\pi/2 = 90°$). The final panel shows $\cos\theta$, where θ is the angle between the two velocities (and is zero up to rounding errors). Contours of f and N plotted with an offset plotted as in Fig. 6.9. See `waves.ipynb`.

The continuous analog of the Rossby deformation radius is defined as

$$L_d = \frac{NH}{f_0}, \tag{6.24}$$

where we could make the definition more general and use f instead of f_0 in the definition. In this case however you should think of H as the vertical length-scale associated with the dynamics, rather than the depth of the ocean, arising from the $1/k_z$ factor in Eq. (6.23). Again, as in the shallow water case, dynamics with horizontal length-scales below the deformation radius cares more about buoyancy effects. In the shallow water case however the only vertical length-scale is the fluid depth, because the fluid was assumed to have constant density,

so the fluid in that system has to move as columns and in a depth-independent sense[11]. In the continuous case where the density is not constant, i.e. $N \neq 0$, there are multiple possibilities of the vertical length-scale, and it is really the one associated with the dynamics that matters.

In the atmosphere, choosing a scale height H of around 10 km, then $L_{d,\text{atmos}} = O(1000\text{ km})$. On the other hand, if we do take the ocean fluid depth, the deformation radius is much smaller, with $L_{d,\text{ocean}} = O(50\text{ km})$, because the relevant H in the ocean is much smaller. The deformation radius is roughly the characteristic horizontal length-scale of *baroclinic eddies*. In the atmosphere *baroclinic instability* leads to what are the *synoptic structures* or weather systems of cyclones and anti-cyclones, while in the ocean they lead to geostrophic or mesoscale eddies. The dynamical process is essentially the same, but the resulting objects have different length-scales because the systems have different characteristics. Note also that the Rossby deformation radius is smaller in the higher latitudes, and is formally infinity at the equator.

[11] The vertical motion is *slaved* to the horizontal motion in such a model.

Figure 6.11: Phase lines and group velocity of internal waves.

6.1.6 *Kelvin waves*

There are two types of waves that don't quite fit as nicely in the above, but are important that we will make use of later. We start with **Kelvin waves**[12], which is a inertia-gravity wave that you only get in the presence of *boundaries*. The boundary however could be land itself, leading to **coastal Kelvin waves**, or could be a *wave guide* such as where f changes sign, i.e. the equator, giving rise to **equatorial Kelvin waves**. In the shallow water model the dispersion relation of Kelvin waves is given by

$$\omega = k\sqrt{gH}, \tag{6.25}$$

so that Kelvin waves are non-dispersive, propagating at the speed of surface gravity waves, but they actually need $f \neq 0$ even if f doesn't show up in the dispersion relation. These waves are *fast* (relative to internal waves certainly), propagating along the coast and/or the equator quickly.

Unlike the case for surface gravity waves where we mostly discuss the positive branch for convenience, here we *have to take the positive branch*! Briefly, if we take the usual (x,y) co-ordinate system to represent the zonal-meridional direction, taking $f > 0$ so we are in the northern hemisphere, and suppose we have the equator at $y = 0$, with $y > 0$ the region of interest (or you can think of $y = 0$ being a wall and $y < 0$ is 'land', it doesn't really matter). Then the

[12] After the British mathematical physicist William Thomson (1824-1907), 1st Baron Kelvin, who was knighted and made a Lord by Queen Victoria. The measurement of temperature K is named after Lord Kelvin, as well as the *Kelvin–Helmholtz instability* that we will encounter later. Most of his groundbreaking work carried out at the University of Glasgow in Scotland.

displacement waveform looks something like

$$\eta(x, y, t) \sim \mathrm{e}^{-fk/\omega y} \cos(kx - \omega t), \tag{6.26}$$

where the factor of $\mathrm{e}^{-fk/\omega y}$ arises because of the presence of the boundary at $y = 0$. Now, the maths tells us that $\omega_\pm = \pm k\sqrt{gH}$, but we cannot take the negative branch because then

$$\eta \sim \mathrm{e}^{-fk/\omega_- y} = \mathrm{e}^{f/\sqrt{gH}y} \to \infty \quad \text{as} \quad y \to \pm\infty,$$

i.e. the displacement gets exponentially larger as we move away from the boundary, which is unphysical. On the other hand,

$$\eta \sim \mathrm{e}^{-fk/\omega_+ y} = \mathrm{e}^{-f/\sqrt{gH}y} = \mathrm{e}^{-y/L_d} \to 0 \quad \text{as} \quad y \to \pm\infty,$$

so the solution is at least physical. Thus the physics tells us that we have to take the positive branch. We find that (1) these waves are *trapped* to the boundary, decaying to zero as we move away from the boundary, (2) the decay scale is the (shallow water) Rossby deformation radius L_d, and (3) in this instance, for the northern hemisphere case, the wave propagates to the *east* since the phase speed $c_p = \omega/k = \sqrt{gH} > 0$ is positive. It can in fact be shown that Kelvin waves propagate with the boundary to the *right* in the northern hemisphere, i.e. in an anti-clockwise sense around a basin, and to the *left* or in the clockwise sense in the southern hemisphere[13]. These will be seen again when we talk a bit about tides in Ch. 6.3.

[13] The way I remember it is that Kelvin waves propagate in a cyclonic fashion (so in the same sense as f).

6.1.7 *Rossby waves*

The other important type of wave that I will refer to quite a bit are **Rossby waves**[14], which are large-scale planetary waves, and requires *gradients* in the Coriolis effect to exist. I actually explicitly derived the dispersion relation for the β-plane in the lead up to Eq. (6.16). Taking no background flow ($U_0 = 0$), the dispersion relation is given by

$$\omega = -\frac{\beta k_x}{k_x^2 + k_y^2}. \tag{6.27}$$

[14] Same Rossby as the Rossby number in Ch. 3.2.2.

In this case there is no choice of branch to take, Rossby waves propagate to the *west* in *both* hemispheres. Another thing to note is that longer waves propagate *faster*, and in general these planetary waves are low frequency long waves. The various wave characteristics can be computed numerically and Fig. 6.12 shows ω, the phase speeds $c_{p,x}$, $c_{p,z}$, and the components of the group velocity $c_{g,x}$ and $c_{g,z}$. The diagram again shows the $\log_{10}$ of the quantity, and the numbers show the exponents, i.e. the number a in $c_{p,x} = 10^a$ (recalling that $\log_{10} c_{p,x} = \log_{10} 10^a = a \log_{10} 10 = a$). While for $k_x, k_y \geq 0$ the

phase speed is negative, the zonal group velocity depends on the choice of wavenumbers (bottom left panel). These observations are consistent with what we saw in Fig. 6.4, which takes $k_y = 0$, and while the individual waves have phase propagating to the left, the group velocity of the wavepacket propagates to the right.

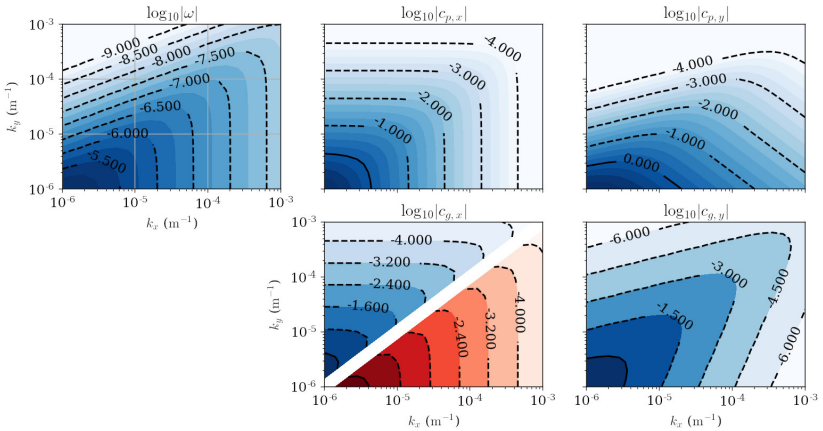


Figure 6.12: Rossby waves ω, $c_{p,x}$, $c_{p,y}$, $c_{g,x}$ and $c_{p,y}$ as a log-log plot in (k_x, k_y) space, with magnitude also as logs. The contours denote the exponent x of $|10^x|$ and the colour shading denotes the sign (more blue = more negative *actual* values, more red = more positive *actual* values); since k_x and k_y is chosen to be positive, everything except $c_{g,x}$ is negative. Choice of k_x and k_y correspond to wavelengths roughly between 6 km to 6000 km (Rossby waves are usually seen as planetary-scale waves). See `waves.ipynb`.

But why do Rossby waves propagate to the west[15]? It's certainly true from the maths, but here we give a kinematic argument using fluid parcels, similar to that used in Fig. 2.1 as an interpretation for static instability with stable and unstable density stratification. This bit will serve as the building blocks for interpreting *shear instabilities* in the next section.

Recall that in the parcel arguments given in Fig. 2.1, we assume *temperature* of the fluid parcel is conserved as the parcel is moved around (assuming mixing is weak), then we consider the buoyancy forces and/or implied motion on the parcel. We make analogous assumptions here for Rossby waves, except we assume that it is *vorticity* that is being conserved as the wave does its thing. The overall schematic is given in Fig. 6.13, where we take the northern hemisphere set up without loss of generality. The argument then is as follows:

1. The Coriolis parameter is largest at the poles and zero at the equator, so on this particular β-plane, we have a background gradient in the Coriolis parameter, which we will denote by $\pm\Delta Q$ respectively (capital Q because it is the background arising from the system)[16].

2. You take a waveform denoted by the black line, and for simplicity assume that this waveform itself carries vorticity equal to the average of both sides (in this case zero).

[15] Or, more generally, Rossby waves propagate in a *retrograde* fashion, counter to $\boldsymbol{\Omega}$

[16] We are going to be using q and Q rather than ζ and Z because we could in fact use *potential vorticity* such as that defined in Eq. (4.5) for this argument, and potential vorticity $q = (f + \zeta)/H$ includes relative vorticity ζ as a special case.

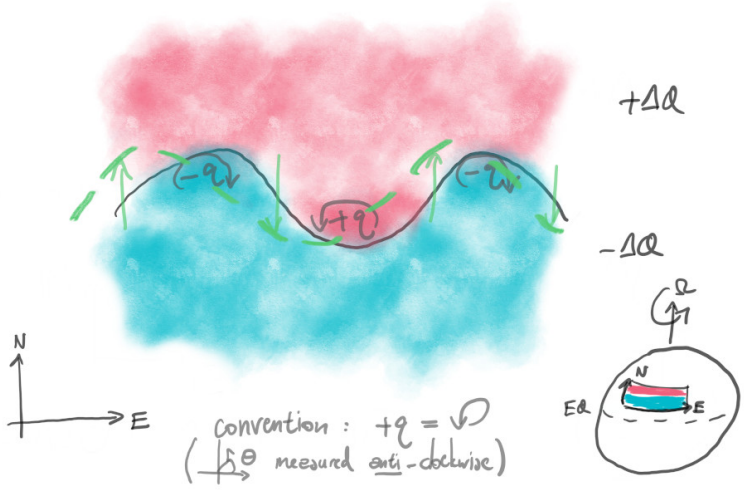


Figure 6.13: Rossby wave propagation schematic.

3. Now, assuming the wave itself carries zero vorticity, if the wave intrudes into the northern region preserving its vorticity signature, relatively speaking the wave will be carrying negative vorticity anomalies $-q$, since $0 < +\Delta Q$. Similarly, if the wave intrudes into the southern region, the wave will be a positive vorticity anomaly $+q$ because $0 > -\Delta Q$. We thus mark on the vorticity anomalies accordingly.

4. Now, recall that vorticity is related to the velocity via a curl and, by convention, $+q$ is anti-clockwise spin, and $-q$ is clockwise spin, so what this means is that vorticity anomalies *induce* a velocity. We mark on the sense of spin associated with the $\pm q$ anomalies.

5. Convince yourself that, at the nodes of the wave, we should have a north/south velocity induced by the vorticity anomalies, as marked on by the green arrows. The new position of the wave after some time from this self-advection is marked on as green dashed lines above and below the nodes of the wave.

6. The argument you then make is that we assumed wave-like solutions that do not change in amplitude, so you want to try and connect the green lines above and below the nodes in such a way that the waveform itself maintains the same shape and amplitude. From this, you argue that the only way you can satisfy these constraints is if the waveform itself is shifted to the left, i.e. the wave propagates to *west*.

   And that's it! Note that this argument works for the southern

hemisphere because even if we shift into the southern hemisphere, we still have the $-\Delta Q$ and $+\Delta Q$ configuration (f is more negative to the south and less negative to the north)[17]. Like in the wind-driven gyre theory in Ch. 5, vorticity arguments is one way, and there are other ways of arguing that lead to the same conclusions. I am using vorticity because to me it is the 'cleanest' way I know, and you can actually use the same argument for *shear instabilities* as well as other waves! For other types of waves, it might not be (potential) vorticity that is conserved by the fluid parcels, but something else, but then one can continue by considering how this 'something else' generates vorticity anomalies, and proceed accordingly[18].

[17] Because it is β that matters, and $\beta > 0$.

[18] See discussion about the instability mechanism later for more.

6.2 *Instabilities*

A slightly less trivial concept to bear in mind with waves is that, if waves are just propagating through a region in the medium without *breaking* or *dissipating*, then while the region feels a perturbation as the wave is travelling through, there is no irreversible change in the medium itself (cf. the concept of irreversible work related to in-situ temperature/density in Ch. 2). You can imagine our friendly neighbourhood pig walking around some park but doesn't leave any traces behind (e.g. footprints, remnants of its lunch etc.), and then after it walks home, hypothetically speaking you might examine the park and you wouldn't be able to tell if the pig was ever there in the first place. So in order for the small fluctuations to have any irreversible effect on the background state, i.e. forcing of the mean state by tracer mixing (e.g. momentum, temperature, salinity etc.), which is the kind of things we are interested in, some sort of *irreversible breaking* needs to happen[19]. In order for waves to break, intuitively we would expect we need the waves to *steepen* and/or grow in magnitude, much like the case of waves coming in towards the beach, steepening and eventually crashing and leading to mixing (e.g. Fig. 6.14).

[19] There are cases we one can formalise these statements, e.g. *non-acceleration theorem*, such as that of Charney and Drazin [1961], phrased in terms of conservation and forcing/dissipation of *wave action*.

Figure 6.14: Schematic of mixing by (irreversible) wave breaking leading to e.g. diapycnal mixing.

By assumption at the moment waves are 'small' perturbations on top of some basic state and/or equilibrium (e.g. the $\tilde{\psi}$ in $-U_0 y + \epsilon\tilde{\psi}$ in Eq. (6.11), where $\epsilon \ll 1$). One could argue unless they grow in amplitude (however 'amplitude' is measured) by some mechanism,

their overall effect should remain small. But the proposed waveform that looks like $\eta = A\cos(kx - \omega t)$ is assumed to not grow, because A is fixed in time by the neutral assumption.

Here we examine **instabilities** and the associated mechanisms that can lead to growth in amplitude of a disturbance, and describe some of the qualitative aspects such as how they are arise, and what kind of effects they have on the state that generated them in the first place. We do not pursue the quantitative aspect here (see e.g. Vallis [2006] for atmosphere and/ocean focused ones, or Drazin and Reid [1981] for more general fluid instabilities).

The concept is instability is best illustrated by considering the simple case of a spherical pig on some frictionless landscape acted upon only by gravity, as in Fig. 6.15:

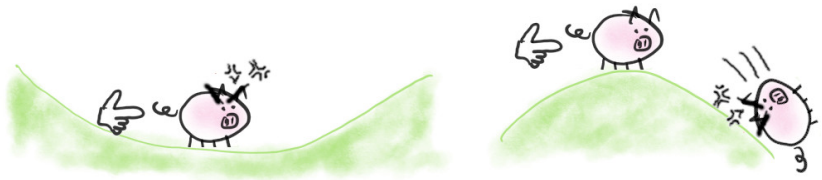


Figure 6.15: Pig being prodded (probably don't try this in real life).

If the pig is in a valley, and you prod it, it might wobble around the bottom of the valley but otherwise stay close to where it started from. If there was friction, then this would damp the oscillations over time, and the pig returns to its original starting position. However, if it was on top of a mound, and you prod it, then it could roll down the hill and not return to where it started from, and in that sense the amplitude of the perturbation keeps growing until the small amplitude assumption is no longer valid. The above three descriptions then encapsulate the state being

- **neutral**, where the perturbations do not grow or decay over time/space,
- **stable**, where the perturbations decay over time/space,
- **unstable**, where the perturbations grow over time/space.

Waves described above are assumed to be neutral. Here we are going to be mostly focused on things that go unstable, since these are the things that will change the background state and lead to phenomena such as large vertical transports, slumping of isopycnals, and irreversible mixing. Instabilities you can think of as the way the system gets rid of its excess energy. In an unstable state (e.g. gravitational energy possessed by the pig on top of the hill), there is

energy that is available for the system to get rid of, and the instability taps that energy reservoir to fuel itself (e.g. the pig rolling down the hill). When that reservoir is either depleted or no longer available to be tapped — mostly because the background state has been modified somewhat — then there is no more instability (e.g. the pig reaching the bottom of a valley, at a local minimum of gravitational potential energy).

Without showing the mathematical details, if one were to want to study the instabilities *quantitatively* (e.g. deriving *criteria for in/stability*, quantifying *growth rates*, *propagation*, *length-scales* etc.), then one approach is to do what we did in Ch. 6.1.2: find a basic state, perturb, be wise and linearise, propose the waveform, and solve the resulting dispersion relation. The main change here is that ω and/or k itself could be complex, which leads to temporal/spatial growth/decay in the waveform depending on the set up. The dispersion relations shown above by construction only gives you real values of ω assuming $N^2 \geq 0$; if $N^2 \leq 0$ we actually get instability (e.g. see Eq. (6.23) and one of the chapter exercise). We are not going to touch on the quantitative details here; see Vallis [2006] or Drazin and Reid [1981] for more.

6.2.1 *Static instabilities (i.e. density related mostly)*

Sometimes these are known as *convective* instabilities, and here I am referring to instabilities that are normally regarded as arising because of unstable density/temperature/salinity configurations in the *vertical*, and ones that do not rely on the existence of a background flow (as compared with *shear* instabilities). However, one should bear in mind that the instability will generically lead to perturbations in the velocity as well as density field, since these are coupled to each other.

We have actually alluded to convective / static instabilities before in Ch. 2, for example in Fig. 2.1, associated with a density configuration. With $N^2 \geq 0$, we have a stable density gradient, and perturbations (which you can think of as the water parcel being moved) lead to internal gravity waves as described in Eq. (6.23), with characteristics dependent on N and the wavenumber of the perturbation. On the other hand, if $N^2 < 0$, then a perturbation of this background state leads what is known as the **Rayleigh–Taylor instability**[20]. The instability is normally very quick, as the system really wants to go back to being a statically stable state, and it does this via plumes of fluid that lead to a vertical transport, growing quickly to lead to nonlinear dynamics (where the $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$ becomes important), rapid generation of small-scales and mixing of material,

[20] After the British scientists John William Strutt, 3rd Baron Rayleigh (1842-1919) and Sir Geoffrey Ingram Taylor (1886-1975), who are both well-known for their numerous contributions in the field of fluid dynamics.

until a stable density gradient is re-established and quenching the instability. Ample videos and/or pictures exist for what these instabilities look like[21].

If instead the background state is taken to be zero initially, but a heating is applied to the boundaries, then depending on the physical characteristics of the system the system could be unstable to **Rayleigh–Bénard convection**[22]. If you heat the bottom and cool at the top (as in the atmosphere, or you heat a pot of water from the bottom, as in Fig. 2.4), then if you apply a hard enough[23] heating, then there can be an unstable gradient leading to vertical overturns and convective cells, with a certain growth rate and spacing of the cells can be computed from a linear instability analysis (e.g. when you heat a pot of water in a pan, just before the water starts boiling, you might see hexagon-like structures, which can actually be predicted). However, in the case where heating is at the top (as in the ocean, as in Fig. 2.4), then you don't get these kind of things happening. From an energy perspective, here you start with no energy sin the system, but you are putting potential energy in the system via forcing the system, and the question is whether you put it in in such a way that this energy can be utilised by the dynamics.

The point here is that, in line with the discussion in Ch. 5.2.3, the aforementioned static instabilities in the ocean operate in very isolated places and also not all the time, because the ocean is by and large stably stratified, with $N^2 \geq 0$. They do operate, but only occasionally and at very specific locations, such as those mentioned in Ch. 5.2.3. These contribute to the downwelling of water at the surface boundary layers (but only over limited vertical extents), as well as deep water formation at those isolated places where there is significant buoyancy loss.

Before we move on to *shear* type instabilities, we touch on **double diffusive instabilities**. These kind of instabilities rely on a difference between the tracer diffusivities, which in the ocean's case is temperature and salinity diffusivities[24]. When we talked about Fig. 2.1 we ignored the contribution of salinity to density, and the background states we have are be buoyantly-(un)stable and temperature-(un)stable. It is perfectly possible to construct cases where you can buoyantly stable with $N^2 \geq 0$ — which fundamentally has to require one of temperature or salinity gradients to be in a stable configuration — but have either temperature or salinity taking the unstable configuration.

The former case is harder to get in the ocean (because temperature largely controls the density over most of the ocean), but the latter configuration is perfectly possible and is illustrated in Fig. 6.17. Here, the temperature is in a stable configuration, but the warmer water is

[21] `www2.eng.cam.ac.uk/~msd38/gallery.html` for example.

[22] Same Rayleigh above, and the French physicist Henri Claude Bénard (1874-1939). The case with surface tension included (which may be important for example at the ocean surface) is sometimes called *Bénard–Marangoni convection*.

[23] Again 'hard enough' is relative, and depends on a ratio. See the *Rayleigh number* on Wikipedia for example.

[24] In general it is usually temperature and something else, e.g. Lithium in the Sun, as a diffusivity on the composition gradient. Double diffusive convection is sometimes referred to as *semi-convection* in the astrophysics literature.

more salty (sort of like top part of the ocean, cf. Fig. 5.8 and 5.9). If you can somehow move the warm water parcel down to the colder fresher region, the parcel of water is stable from a temperature point of view, but unstable in the salinity gradient. Here we have to rely on some sort of mixing/diffusion, unlike the case of Rayleigh–Taylor instability: since temperature diffusion κ_T is faster than salinity diffusion κ_S, the heat could mix out with the surroundings, so the water parcel may equalise in temperature, but *not* in salinity, so the resulting water parcel is denser and continues sinking (and converse case of moving a water parcel up is argued in an analogous way).

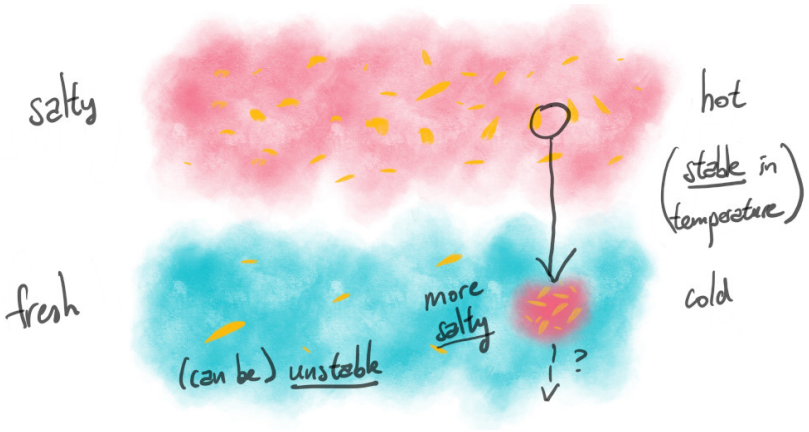


Figure 6.16: Salt fingering schematic: stable in temperature gradient but unstable in salinity gradient.

Instead of the vigourous overturning motion associated with Rayleigh–Taylor instabilities, you get what are **salt finger instabilities**. These salt fingers operate on a slower time-scale compared to Rayleigh–Taylor instabilities, because they fundamentally depend on diffusive effects, but result in well formed plumes such as those seen in the lower panels of Fig. 6.17, with a characteristic growth rate and spacing that can be calculated from a linear instability analysis. Its subsequent nonlinear evolution could be simulated accordingly, with one such result shown in Fig. 6.17. There is some evidence that salt fingering instabilities operate and lead to the characteristic *staircase* profiles that are sometimes seen in the vertical buoyancy profile, particularly in the ice-forming regions. The interesting thing to note is that, slightly ironically, while we have $\kappa_T \gg \kappa_S$, because of the presence of salt fingers, even if salinity *diffuses* less, it can actually *mix* more, because of dynamical effects.

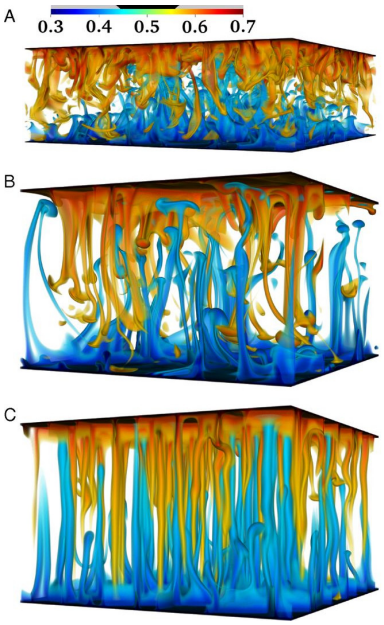


Figure 6.17: Simulation in in temperature stable but salt unstable regimes. From Yang, Verzicco & Lohse (2016), *Proc. Nat. Acad. USA*, modified from their Fig. 1.

6.2.2 *Shear instabilities (i.e. flow related mostly)*

Shear instabilities I'm going to to mean instabilities that require a background flow, and in particular a 'shear' in the flow, i.e. spatial

gradients. You will of course note that, for large-scale flow, flow and density are related through thermal wind shear relation, but here we are going to assume the background state is stable to static and double-diffusive type instabilities for sake of argument. You also probably guessed it from my emphasis of shear and gradients, I am going to use vorticity to describe these things.

Fig. 6.18 shows the result of a simulation a (horizontal) shear flow which is unstable to a (horizontal) shear instability, plotting the vorticity $\zeta = \boldsymbol{e}_z \cdot (\nabla \times \boldsymbol{u})$ associated with this *shear layer* profile. This output from a simulation you can consider as like Fig. 6.14 but following a wave crest, so we follow the instability as it develops. The initial perturbation takes its sweet time growing, but eventually, when it does grow to large enough amplitude, it starts acting back on the initial shear layer via the velocity associated with the instability, causing a roll-up of the shear layer, leading to mixing of material across the shear layer, and sheds a vortex. Though not shown, the initial shear layer broadens, and the initial gradients in the velocity profile are eroded somewhat by the (fundamentally nonlinear) feedback of the instability, reducing the 'unstable-ness' associated with the initial shear flow, but the velocity shear is not removed completely. From an energetic point of view, there is energy that is available to fuel the instability, but once the *primary* instability takes hold and modifies the background state, the basic state may have been modified such that there is probably still an energy reservoir there, but this may or may not be accessible to fuel further *secondary* instabilities. The experimental set up for Fig. 6.18 was chosen so that only one of these vortical eddies are formed, but in practice in the real world a train of these may form. These vortical eddies each have an induced velocity, and the mutual interactions allow them to move around, merge with each other, and can fuel further static and/or shear secondary instabilities.

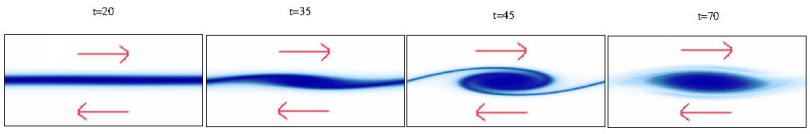


Figure 6.18: Roll-up of flow arising from a shear instability (horizontal, no stratification here). Adapted from Mak et al. [2017], Fig. 1. An animation of this diagram can be found in `omega_tanh.avi` (left column).

I'm just going to call the above a 'shear instability', and I was fairly deliberate in not specifying what scales of motion I am talking about here, since in principle these instabilities can occur for length-scales across all ranges. In practice they are fastest growing for small-scales (because the corresponding shear in the velocity is then more significant). Perhaps the most famous example of shear instabilities

is the **Kelvin–Helmholtz instability**[25], though that name is often reserved the naming for where the flow shear is in the vertical and the instability includes the effect of density stratification (the above case is purely horizontal). Kelvin–Helmholtz instabilities are normally seen in the atmosphere with cloud billows rolling up (because the flow entrains the cloud as it rolls up into the vortices), and tend to only have a lifetime of around 15 mins or so before it dissipates. The experiment shown in Fig. 6.18 is for a horizontal setup, but there are substantial commonalities between the features that arise in the horizontal and the vertical settings[26].

Baroclinic instability specifically is for larger-scale motion with a vertical shear but a stratification profile that is statically stable. By thermal wind shear relation Eq. (5.2), a vertical shear is related to tilting isopycnals, such as that displayed in Fig. 6.19. Since baroclinic instability is really a shear instability, the end results are similar in that the background state is susceptible to an instability, which results in roll ups and shedding of eddies as in Fig. 6.18, but with a particular growth rate and length-scale (normally with eddies around the Rossby deformation radius).

A snapshot of baroclinic instability in action in a ocean-relevant numerical simulation is shown in Fig. 6.20 (but see also Fig. 4.9 for the gyre case). One can think of the vertical flow state as being in a stressed state, and the system wants to relax by generating baroclinic mesoscale eddies so that the system can release some of its energy, which then leads to a more relaxed state that has a reduced vertical velocity shear. A reduction in the vertical shear in the velocity leads to a *slumping* of tilting isopycnals by the thermal wind shear relation Eq. (5.2), thus baroclinic mesosacle eddies can be thought as leading to an eddy induced overturning such as that seen in Fig. 6.19, and argued to be important in setting the Southern Ocean overturning circulation in Ch. 5.1.1. There are several types of baroclinic instability following the paradigms of Eady [1949], Charney [1947] and Phillips [1956] (see Vallis [2006], Ch. 6), which all have slightly different flavours to them, but I will distinguish these after the discussion of the mechanism for shear instability.

Baroclinic instability is one type of shear instability that operates in the ocean, but is by no means the only one (e.g. *symmetric instabilities*, *inertial instabilities*, *Holmboe instabilities*, etc.), and we are not going to review all of them. The main point is that, generically, the process of the instabilities being generated by the mean flow leads to a broadening of the background velocity shear, so as to relax the system somewhat and to release some of the energy in the reservoir. In that sense one could think of these instabilities are leading to enhanced viscosity, through the enhanced broadening of

[25] Same Kelvin as above, and after the German physicist and physician Hermann Ludwig Ferdinand von Helmholtz (1821-1894). Helmholtz is perhaps better known for his fundamental contributions to foundations of thermodynamics, underlying some of the concepts used in Ch. 2. The original Kelvin–Helmholtz instability analysis formulates the shear buoyancy jump as being isolated in an infinitely thin *vortex sheet*, where analytical solutions are available.

[26] With a vertical setting and stratification is involved, a lot of *secondary instabilities* can be triggered. See e.g. Mashayek and Peltier [2012a,b] for a comprehensive list.

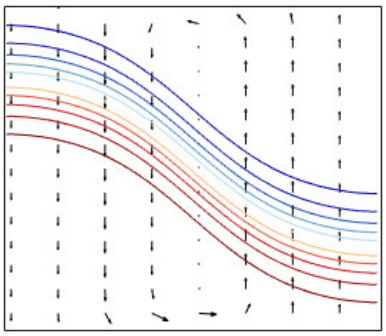


Figure 6.19: Eddy induced circulation of baroclinic instability, acting to flatten tilting isopycals. Numerical experiment after Gent et al. [1995].

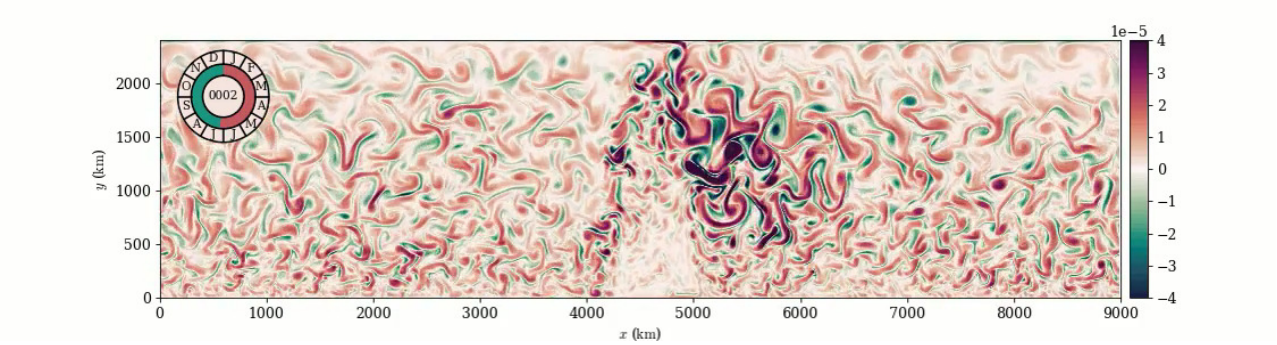


Figure 6.20: Simulation in a wind-forced zonally-periodic β-plane channel, with a ridge in the center of the domain, showing vertical component of (relative) vorticity. After set up of Munday et al. [2015].

the background velocity shear, and/or diffusion, since the eddies *stirs* the system, which leads to enhanced mixing by generating smaller length-scales so that diffusion can efficiently act on. These processes we already argued to be important in the ocean in Ch. 3.4 and Ch. 5. One thing to note before we move on is that while eddies are generally seen to lead to enhanced diffusion, that is of course not all they can do...[27]

So how do these instabilities become unstable, and can we use a kinematic argument based on considering how a fluid parcel behaves in the background state? You could try this with a shear flow say in Fig. 3.20. I personally can't make head or tail out of it, because the only thing I see is a fluid parcel being sheared and/or twisted around, which doesn't really tell me whether there is something (ideally a displacement) growing[28]. Below I'll walk through a way that at least makes sense to me pictorially, taking a vorticity picture, as advertised at the beginning of this section. I should just emphasise this is only one way of looking at the problem (cf. the gyre explanation in vorticity arguments, but equally ok in momentum and pressure), there are presumably others that are equally good, but this one works for me.

The central ingredients we are going to use are related to how Rossby waves propagate (see Fig. 6.13 and the surrounding discussion), namely:

- assumption that fluid parcels conserve (potential) vorticity q as they move around with the waveform;
- the vorticity anomalies denoted $\pm q$ has an associated a velocity that can lead to self-advection;
- in the presence of multiple waves and depending on phase differences of the waves, these can lead to constructive/destructive interference accordingly, i.e. instability or stability depending on the phase shift.

[27] See Fig. 4.9 for a hint of an anti-diffusive behaviour, where the presence of eddies seems to energise the model Western Boundary Current. This *backscatter* phenomenon is a matter of current research; see Vallis [2006] Ch.9 and Ch.12 for an introduction to the topic maybe.

[28] There is a explanation by Orr [1907], which never seemed very intuitive to me.

The only slight change in the thought process for the moment is to think of the background velocity $\boldsymbol{U} = U(y)\boldsymbol{e}_x$ as the supplier for the *background (potential) vorticity* profile, when previously when we were talking about Rossby waves we used the *background planetary rotation* profile instead. I will argue these are semantic differences, and the formal argument itself can be continued as normal.

First lets set the scene. The setting is as in Fig. 6.21, which is as follows:

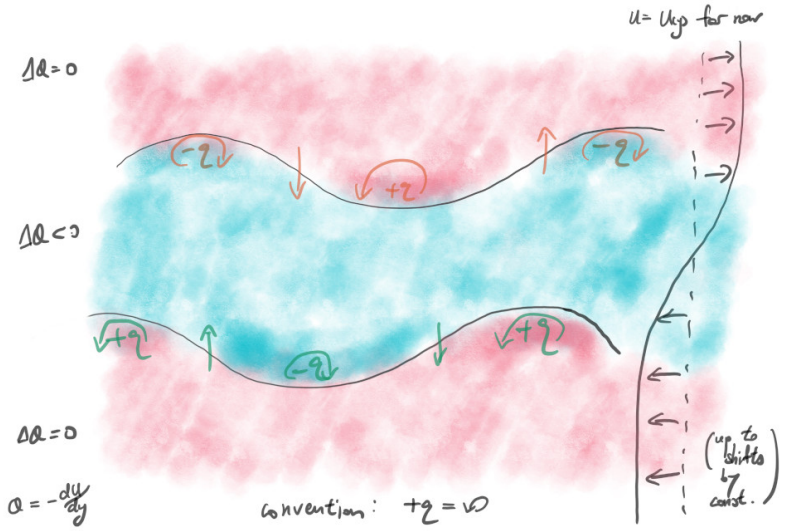


Figure 6.21: Schematic of the problem, with two waves drawn on, each preserving (potential) vorticity anomalies $\pm q$ accordingly.

1. We put down a basic background flow, which is $\boldsymbol{U} = U(y)\boldsymbol{e}_x$, as drawn on with the black line on the right.

2. The background flow then implies what the basic state vorticity should be via $Q = -\mathrm{d}U/\mathrm{d}y$. For this particular (though rather typical) shear flow, we have essentially $Q = 0$ outside of a certain region (because the flow is close to uniform), and in the middle we have a region with positive velocity shear, so in the middle we have a layer of negative background vorticity $-\Delta Q$. These background vorticity signs are marked on in black on the right.

3. We draw on the waveforms and work out what are the associated vorticity anomalies $\pm q$. The top waveform situation is as in Fig. 6.13, since we have a region of zero background vorticity 'above' a region of negative background vorticity. Note in this case we have an 'up' displacement η correlated with a negative vorticity anomaly $-q$, and vice versa, i.e. $\pm\eta \sim \mp q$. For the bottom waveform the situation is reversed. Convince yourself that $\pm\eta \sim \pm q$ in this case.

4. The associated vorticity anomalies lead to an induced velocity, with anti-clockwise flow taken to be associated with positive vorticity anomalies in line with convention. In the absence of the other wave, these vorticity anomalies of the individual waves induces a self-advection.

Note that, by themselves in the absence of the other wave, the 'top' wave propagates to the left (as in Fig. 6.13), while the 'bottom' wave propagates to the *right* (use the same arguments surrounding Fig. 6.13). These waves are generically propagating *against* the background flow $U(y)\boldsymbol{e}_x$, and we call them Rossby waves also here because it turns out one could almost think of β being replaced by ΔQ in the dispersion relation Eq. (6.16) here. One of the key ingredients for our description on shear instability mechanism are these **counter-propagating Rossby waves**.

Now we look at how the pair of these counter-propagating Rossby waves interact, as depicted in Fig.6.22:

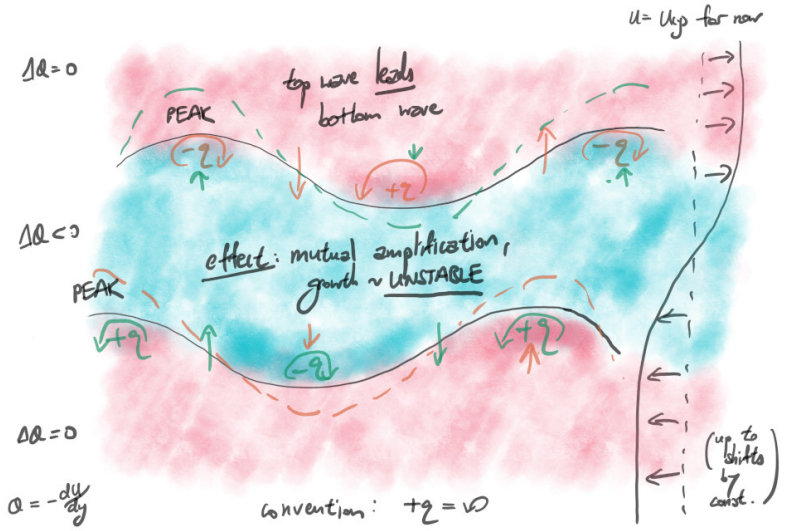


Figure 6.22: Counter-propagating Rossby waves in an unstable configuration.

1. The vorticity anomalies of the 'top' wave induces a velocity (all marked in orange) that decays with distance from the vorticity anomaly. The induced velocity not only leads to counter-propagation of the wave, but because of the particular phase shift chosen, the 'top' wave has a 'down' velocity at where the 'bottom' wave is a trough, and an 'up' velocity at where the 'bottom' wave is a peak. This is a *magnifying* effect, leading to an increase in the 'bottom' wave displacement.

2. Similarly, the vorticity anomalies of the 'bottom' wave induces a velocity (all marked in green), and because of the phase shift chosen, the 'bottom' wave also magnifies the displacement of the 'top' wave, leading to larger displacements. So this particular set up where the top wave *leads* the bottom wave by a quarter of a wavelength (so a phase shift magnitude of $\pi/2$), we have a *constructively interfering* set up, where both waves are mutually amplifying.

3. What it also turns out is that, if the phase shift is not quite a quarter of a wavelength, the waves mutually try and drag or push the configuration back into the quarter of a wavelength configuration, by helping/hindering the propagation of the other way. This leads to a phase locking set up, where the waves are forced into a mutually amplifying configuration, which leads to an overall increase in the displacements in the system, and thus can be thought of as an *instability*.

As an exercise, try convincing yourself that if either of the waves are shifted by half a wavelength (so we still have a quarter wavelength phase shift but in the other direction), then the waves mutually act to *damp* the other wave, so it is a destructively interfering case, and we have instead *stability*.

The key ingredients we need for this **Counter-propagating Rossby Wave** mechanism[29] are then:

[29] Though more accurately described as a *Wave Interaction Theory*; see justification coming up. ref

1. at least a pair of counter-propagating waves relative to the background flow, because if we get pro-propagating waves they can't remain in the phase-locked position since they would be swept away by the flow;

2. wave displacement that somehow leads to vorticity anomalies (either by vorticity conservation directly, or something else), which in turn induces a velocity that can lead to *action-at-a-distance*;

3. a phase-shift between the waves so that the action-at-a-distance can lead to constructive interference of each of the waves.

A number of comments then following the above discussion of the instability mechanism:

- The picture presented above uses the background flow as the source of the (potential) vorticity, but in reality there could be multiple contributions to this background (potential) vorticity profile (e.g. from tilting density configurations).

- Although I formally demonstrated this for a *horizontal* shear flow, the same picture carries over for a *vertical* shear flow even when

stratification is involved, i.e. baroclinic instabilities, by considering instead potential vorticity (this is Bretherton [1966]; see also Hoskins et al. [1985]).

- In the absence of a flow, the ocean and/or atmosphere is singled signed in the (potential) vorticity gradient (encapsulated essentially by β), and this does not support the counter-propagation we need to get an instability. Related to this point is that since we need at least a pair of counter-propagating waves (although by itself that is not sufficient for instability, because they don't have to phase-lock), this could serve as an interpretation for the *Rayleigh inflection point theorem* and/or the *Charney–Stern condition*, which states that a *necessary* condition for instability is that the (potential) vorticity profile has to change sign somewhere in the domain. refs

- From a baroclinic instability point of view, recall I mentioned above there are the Eady [1949], Charney [1947] and Phillips [1956] flavours without explicitly saying what they are. They can respectively be thought of as the instability arising from two counter-propagating Rossby waves at

  Eady: the vertical boundaries, where potential vorticity gradients are from the buoyancy discontinuities at the vertical boundaries (the original set up of Eady [1949] has zero interior potential vorticity by construction)

  Charney: one of the vertical boundaries and one in the interior (Charney [1947] has a interior potential vorticity via including the β-effect)

  Phillips: two in the interior (the set up of Phillips [1956] is a two-layer *quasi-geostrophic* system with different uniform flow in each of the two layers, leading to vertical differences in the potential vorticity profile; cf. Bretherton [1966]).

  These instabilities have slightly different characteristics depending on where we are in the ocean, with different growth rates, spatial scales, and eventual nonlinear effects.

- Notice that in the second key ingredient we need, I was deliberately vague and said "*wave displacement that somehow leads to vorticity anomalies*". (Potential) Vorticity being conserved (in purely horizontal systems and some specific vertical and stratified systems) is one way, but it turns out we can be more general. Two examples are pure internal gravity waves (displacement conserving buoyancy but generates vorticity by baroclinic torque) and Alfvén waves (displacement conserving magnetic potential but generating vorticity by the Lorentz force). See for example

Harnik et al. [2008], Rabinovich et al. [2011], and Heifetz et al. [2015]. Note we still need the instability generating pair to be counter-propagating waves (these aforementioned wave modes have left and right-ward propagating branches).

- As in the above in classifying the different types of baroclinic instability, we can characterise the types on shear instabilities we have by the *kind* of waves that partake in the wave interaction. For example, the standard shear instability is through two Rossby waves (or vorticity waves), the Kelvin–Helmholtz instability would be Rossby-gravity waves, the *Holmboe* instability would be something like a Rossby wave with a gravity wave, the *Taylor–Caulfield* instability would be two gravity waves, and so on. See Carpenter et al. [2011] for a review.

- Presumably a similar picture would hold for other types of shear instabilities, such as *inertial* and *symmetric* instabilities, although I am not aware of works on this. Perhaps this picture might be useful in explanation some characteristics associated with instabilities (e.g. reduction of baroclinic instability over slopes Isachsen

- One could use this picture to rationalise some static instabilities (e.g. Rayleigh–Taylor instabilities; see Rabinovich et al. [2011] and chapter exercise)

- The theoretical works mentioned above shows the case mostly for idealised profiles. The picture is slightly less clean cut in more 'realistic' profiles but a similar picture does hold (e.g. Fig. 6.23 for the case with $U(y) = \tanh(y)$).

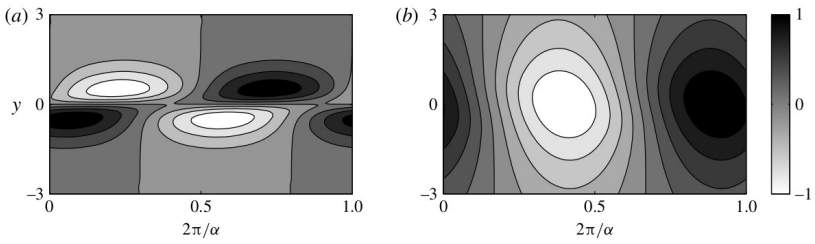


Figure 6.23: Vorticity (left) and pressure (right) eigenfunctions for the most unstable mode from a linear instability analysis for the profile of $U(y) = \tanh(y)$. From Mak et al. [2016], Fig. 6.

6.3 *Tides*

So far we have talked about some characteristics of waves, and presented an interpretation of shear instabilities in terms of waves constructively interfering with each other. Bringing it back to how

these small-scale dynamics contribute and shape the global MOC, the static / convective instabilities contributes to the downwelling of water from the surface, the baroclinic instabilities mediate the momentum transfer from the surface to the bottom, and shear as well as static instabilities presumably operate in overflows when dense water flows down bathymetric features, which can have consequences for the supply of bottom water into the ocean. All this mostly concerns downward transfer of 'stuff'. What is perhaps not yet so clear is how do instabilities operate for the upwelling of water important for closing the MOC. The answer is probably *tides*, specifically *internal tides*, and the shear instabilities and the secondary instabilities associated with steepening of these internal tides, to be viewed here as *internal waves*.

6.3.1 *Surface tides*

Instead of going straight to internal tides it is probably worthwhile going through some of the more easily observed features associated with tides at the surface first, with the understanding most of these aspects carry over to the case of internal tides.

Tides is perhaps one of the most notable features we know about the ocean, where over most places on Earth there is a twice daily increase and decrease of the sea surface elevation, which is particularly notable near coastal regions (because there is relatively speaking a larger observed change since the water depth is shallower). Fig. 6.24 shows a particular example of this in the Scottish town of Tobermory on the Isle of Mull. For Tobermory, having its origins as a fishing village, tides are important because it is only during high tides when the water levels are sufficiently high that fishing boats can leave and/or come back to port. Tobermory is also famous for having colourful houses (you can see these also in Portree at the Isle of Skye), which actually serves as useful markers for visually assessing differences in water levels. Go to the photographer Michael Marten's website `www.michaelmarten.com` for better pictures and more dramatic examples at other locations.

We can of course assess as well as predict water levels more quantitatively, and Fig. 6.25 shows in this case a typical signal in the total water depth at a location over seven days at Tobermory. When the water level is high we have **high** or **flood tide**, and when water level is low we have **low** or **ebb tide**. The time-varying signal oscillated with well-defined frequencies, the dominant one being the 12 hour **semi-diurnal** signal, although there is a hint of the 24 hour diurnal[30] signal. There amplitude of the signal in the water depth is around a meter or so, although there are variations over the length of

[30] Diurnal means day-night cycle.

Figure 6.24: High (or flood) and low (or ebb) tide at Tobermory, Isle of Mull, Scotland, using the pastel pink and red house as references. Modified images from www.thechaoticscot.com (left) and personal collection (right).

time shown, which indicates the presence of other wave frequencies, and the signal we see is actually a combination of those individual waves.

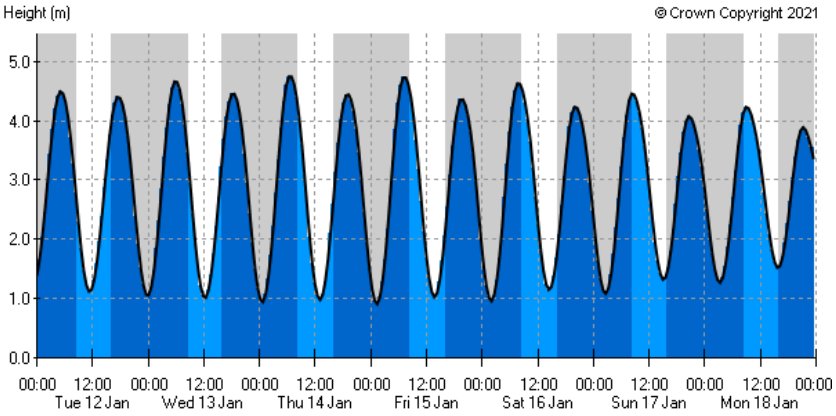


Figure 6.25: Predicted (!) tidal charts at Tobermory, Isle of Mull, Scotland between over a period of seven days. Taken from www.ukho.gov.uk.

The longer term variations are seen better in Fig. 6.26, showing daily maximum and minimums of SSH (relative to the sea floor again) as well as their difference over a typical six month period, also at Tobermory. There are clearly longer term fluctuations, and in particular the signal is suggestive that there are multiple frequencies present. The difference between the maximum and minimum oscillates with a dominant period of around 14 days, although the bigger maximums and minimums seem to occur every 28 days. Where the difference between highest and lowest is maximum is called a **spring tide**[31], and where the difference in tide heights is smallest is called the **neap tide**. The fact the periods of the oscillations mentioned here are multiples of each other is not a coincidence, as we will see in a bit.

One final thing to note is that the signals arise from the fact that we are measuring the disturbances, and these disturbances are probably related to waves. Since waves propagate, tidal signals can

[31] This has nothing to do with the season, but more to do with the bouncy springs, 'springing' into action.

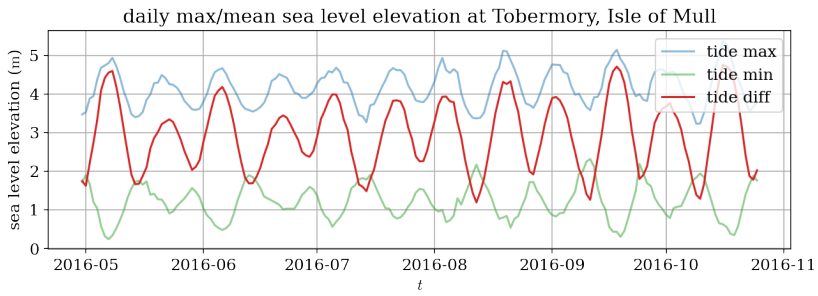


Figure 6.26: Daily maximum and minimum sea surface elevation (blue and green) and their difference (red) over a six month period. Data from BODC, see `tobermory_tides.ipynb`.

be expected to propagate too, and to see this we need to go beyond Tobermory and consider a larger region. Fig. 6.27 shows the phase lines associated with a particular tidal disturbance, and the numbers marked on are hours of the day. How you should see this is that the phase lines are denoting the peaks (or troughs, or nodes, whatever fixed reference you choose) of the wave as it propagates. In this case these waves propagate in an anti-clockwise fashion, with the land on the right. We don't distinguish whether the wave propagates at the phase or group velocity in this case (it should really be phase velocity), because these waves are actually coastal Kelvin waves, which are non-dispersive (Ch. 6.1.6). In general, the disturbances travel cyclonically, so anti-clockwise in the northern hemisphere with land to the right of the wave, and clockwise in the southern hemisphere with land to the left of the wave.

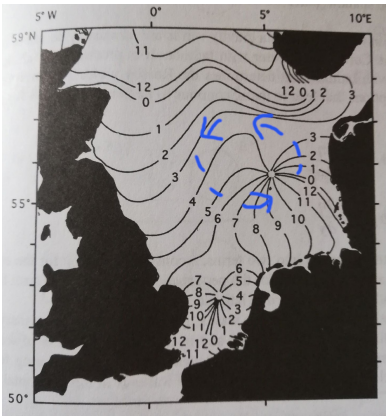


Figure 6.27: Phase lines (denoting peak of SSH, in hours in GMT) of the semi-diurnal tidal signal. Notice the anti-clockwise propagation (with boundary to the right). Diagram modified from Knauss (1997), Fig. 10.5.

6.3.2 *Tidal forcing*

Before we proceed I am going to be a bit more strict and define what I mean when I am referring to **tides**, which is going to be the deformation of a sea surface arising *only* from gravitational forces arising from other astronomical bodies[32]. Note that with this definition we are going ignore the fact that the Earth is rotating, and not bother with fictitious forces because we are not going to be in an accelerating frame. With this view, tides arise entirely from the fact that there is a *differential* in the gravitational force[33], and any additional effects arising from the presence of rotation modifies the outcome but by itself is not responsible for tides. There are multiple explanations that invoke the centrifugal and centripetal effects, but to me these confuse the picture and are not really necessary, so I am going to throw those out. The theory of **dynamic tides** involves wave propagation and geometric considerations, is much more complicated, but is the one we actually need for *predicting* SSH arising from tides. We are only going to focus on the theory

[32] You can have tidal effects on the atmosphere and solid Earth too.

[33] This point of view is consistent with for example that used in General Relativity (e.g., Schutz [2009]) which is the extension of Newtonian mechanics.

of **equilibrium tides** here, to illustrate some of the main aspects of tides we need for descriptive purposes (see e.g. Talley et al. [2011] Ch. 8.6 for more).

Consider for the moment we have a moon and Earth system as in Fig. 6.28, and assume the Earth is not rotating about its own rotation axis. Then there are three geographic points of principal interest: the point on Earth closest to the moon, the **sublunar point** denoted S; the center of Earth, denoted C; and the **antipodal point**, the point furthest away from the moon, denoted A. Lets assume point S and A are both on the equator for simplicity. Then, recalling that we have the equation for gravitational attraction from Eq. (3.2) — the main point being $F \sim 1/r^2$ — convince yourself that the gravitational attraction at points A, C, S (denoted F_A, F_C, and F_S) because of the presence of a moon are such that

$$|F_A| < |F_C| < |F_S|,$$

because point S is closest to the moon while point A is furthest away from the moon.

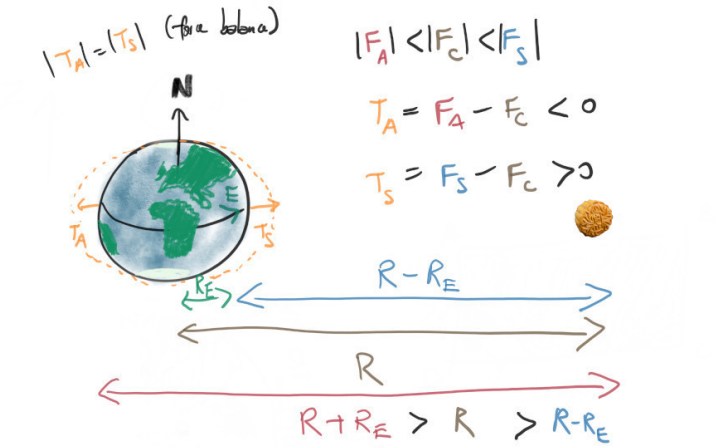


Figure 6.28: Schematic of tidal forcing by an astronomical body. The implied changes to the sea surface is massively over-exaggerated for illustrative reasons.

We want the forces at the surface of Earth relative in this case to the centre of Earth (i.e. I am taking point C to be the 'zero' point of the co-ordinate system), so we define the differences in the forces at point A and S relative to the center of Earth C as

$$T_A = F_A - F_C, \qquad T_S = F_S - F_C. \tag{6.28}$$

Here we have $T_A < 0$ (since $|F_A| < |F_C|$) and $T_S > 0$ (since $|F_S| > |F_C|$), and from geometric considerations these two forces are equal in magnitude, and are the **tide generating forces** at these two points. In principle you can work out the tidal generating force

at every other point on the surface of the Earth and draw them as arrows, and what you find is that these forces are converging water onto the sublunar and antipodal points, and diverging water in this case from the points at $\pm 90^\circ$ away from the sublunar point (rather than at the poles, as could be interpreted from Fig. 6.28). This leads to the bulges (i.e. the high tides) at the subpolar and antipodal points, and depressions (i.e. the low tides) at the poles. The main point is that you get two bulges, but these bulges in this definition of tides have absolutely nothing to do with rotational effects. Rotational and geometric effects however are absolutely crucial in the dynamic theory of tides because these affect how the associated inertia-gravity waves and Kelvin waves propagate around the system. If we are in a non-inertial frame, the sensible 'zero' point should now be the *barycentre* (the center of mass of the combined Earth moon system), and we could proceed as before, but now we have to deal with fictitious forces such as centrifugal forces.

It we rotate the Earth but assume instantaneous equilibrium response, the location of the bulge moves around. Since there are two bulges, and it takes the Earth one day to rotate around itself, we have two high tides and two low tides at a fixed location per day in this equilibrium tide picture[34], i.e. the semi-diurnal cycle. Since the moon also rotates around the Earth, there is not only the semi-diurnal cycle and the dirunal cycle, there is a longer lunar signal too. The picture of course holds if we have the Sun instead of the moon, and we can also have the Sun and the moon together, leading to a (linearly) combined effect on Earth.

[34] When dynamics are involved there are places that occasionally only experience one tidal cycle a day, e.g. the Gulf of Mexico.

It turns out the dominant contribution to tides on Earth is from the moon, particularly the principal lunar semi-diurnal mode, denoted[35] M_2. While the Sun is much heavier, the moon is much closer, and the relevant tidal generating force from the M_2 mode is roughly twice as strong as the strongest solar mode. Table 6.1 shows a few of these tidal modes.

[35] The symbols M_2, K_1 etc. are known as *Darwin symbols*, after Sir George Darwin (1845-1912), son of Charles Darwin of the origin of species fame.

symbol	period (in solar hrs)	rel. amp (to M_2)	name
M_2	12.42	1	principal lunar (semi-diurnal)
K_1	23.93	0.58	luni-solar (diurnal)
S_2	12.00	0.47	principal solar (semi-diurnal)
O_1	25.82	0.42	principal lunar (diurnal)
N_2	12.66	0.19	larger lunar elliptic (semi-diurnal)
⋮	⋮	⋮	⋮
Mf	327.85 ($\approx$ 14 days)	0.09	lunar fortnightly
Mm	661.30 ($\approx$ 28 days)	0.05	lunar monthly
SSa	4382.86	0.04	solar semi-annual

Table 6.1: Some sample tidal forcings sorted by relative amplitude to the M_2 tide (which is the largest forcing for Earth). Subset of Table 6.2 given in Wunsch (2015). The last few entries are weak and long term but they are there.

So we roughly explained why we have the twice-daily high and low tides, but not why we have the roughly twice monthly spring and neap tides. We can explain this as well within the equilibrium tide picture, but now we need to introduce the Sun. Fig. 6.29 shows the case where we have an alignment of the Sun, moon and Earth (the moon being 0° or 0 radians on the Earth-Sun axis), and where the moon is in *quadrature* (the moon being 90°, $\pi/2$ radians, or quarter of a wavelength out of alignment relative to the Earth-Sun axis). Assuming linearity, we can add the contributions from the moon and the Sun together. When the moon is on the Earth-Sun axis, the signals are in phase and lead to maximum displacement, while when the moon is in quadrature to the Earth-Sun axis, the contributions are mis-aligned / out of phase, and do not add to each other completely. The former and latter scenario is of course the spring and neap tide scenario, so the frequency is going to depend on how long it takes for the moon to rotate around Earth. Since the lunar month is about 28 days, so the fact we have springs and neaps every 14 days or so is entirely consistent.

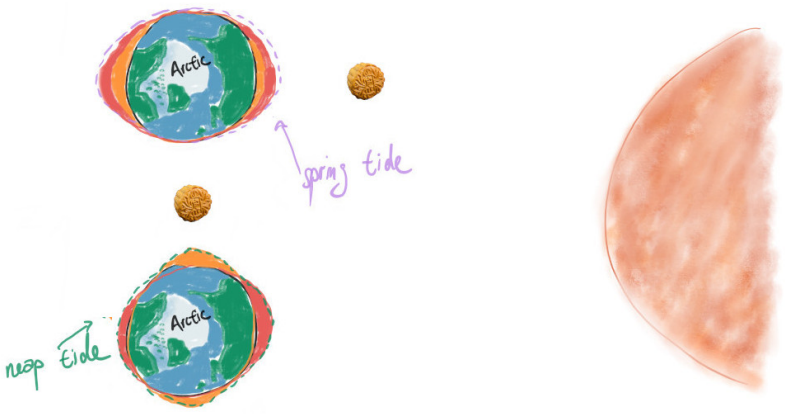


Figure 6.29: Schematic of spring and nap tides. Notice here we are looking down onto the north pole here, with the moon rotating in the plane slicing through the Earth's equator.

6.3.3 *Modes and internal tides*

The ocean depth is tiny relative to the distances involved in Fig. 6.28, and from that point of view we can assume the tide generating forces acting on the ocean to be essentially uniform over the ocean depth. As the Earth rotates around its own axis, the moon rotate around the Earth, and the Earth rotate around the Sun, this leads to an essentially depth-independent sloshing of water over the ocean, generating surface waves and leading to perturbations in the SSH. But at the same time, since there are bathymetric features

in the ocean, this sloshing over bathymetric generates internal waves (Ch. 6.1.5) because the water has to be move upwards by the sloshing motion, leading to a perturbation in the isopycnals and exciting waves, schematically shown in Fig. 6.30. Similar phenomena occur in the atmosphere when we have air moving over topographic features, leading to an excitation of internal waves, the only difference being the medium has a different density characteristic (but otherwise described similarly).

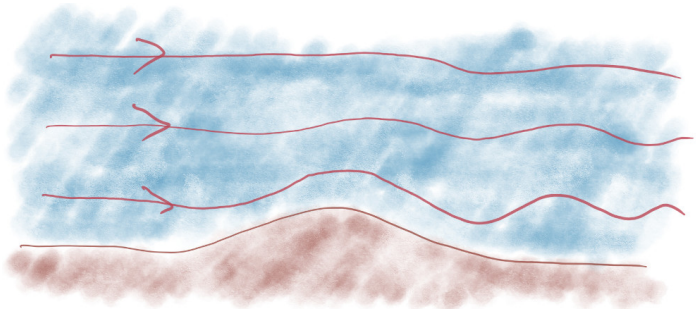


Figure 6.30: Flow over topography (e.g. tidal motion) leading to wave generation.

The essentially depth-independent tidal forcing from the tidal forcing leads to what is commonly called **barotropic tides**[36], which can be denoted by the $k_z = 0$ wavenumber if we take the bottom and top of the ocean as our boundaries for the wave. The situation is that sketched out in the top-left panel of Fig. 6.31. As the depth-independent flow moves over topography, it leads to motions with some vertical variation called **baroclinic tides** (they are really internal waves that can be attributed to tidal forcing), such as the top-right and bottom-left panel, characterised by higher vertical wavenumbers (remember the wavenumber $k_z \sim 1/L_z$, where L_z could be seen as a wavelength, or as a characteristic vertical length-scale of the motion). Associated with this higher wavenumber motions are shears in the associated flow, which itself could be susceptible to shear instability, or partake in some sort of nonlinear wave-interaction phenomenon (which we will not talk about here), but either way the end result is usually motions at even larger vertical wavenumbers, i.e. even smaller vertical and/or vertical length-scales. This generation of small-scale motion leads to mixing usually through a primary shear instability, which triggers further shear and/or static secondary instabilities, contributing to turbulent diffusivities that lead to enhanced diapycnal transfers, as needed in Ch. 5.2.4 to close the global MOC through upwelling. The scale transfer from the large-scale $k_z = 0$ motions to small-scale $k_z \gg 1$ motions (a *downward cascade*) is arrested by viscosity and/or diffusion when the appropriate Reynolds or Péclet numbers (Ch. 3.4.2) become around unity (bottom-right panel of Fig. 6.31).

[36] Although again I don't like the word 'barotropic' being used to mean 'depth-independent', because that's not what 'barotropic' means.

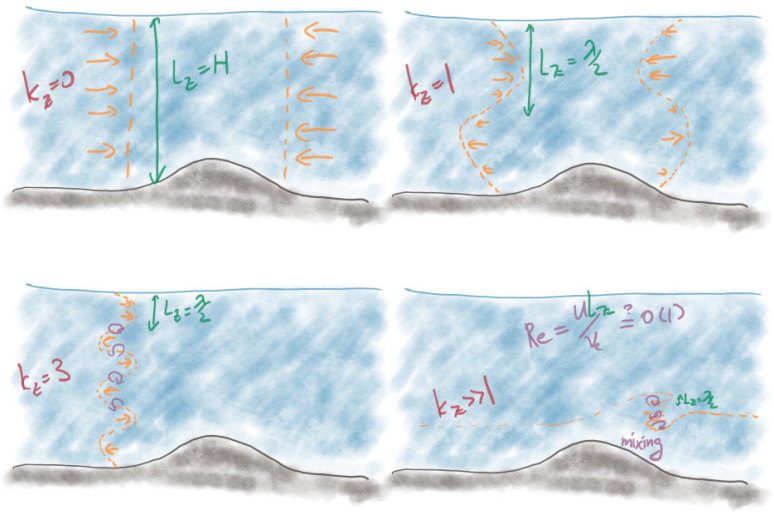


Figure 6.31: Motion associated to theoretical $k_z = 0$, $k_z = 1$, $k_z = 3$ and a more realistic case with nominal $k_z \gg 1$ (with boundary as ocean surface). Note associated with k_z is a length-scale.

To close, we note that while we might understand reasonably well how the individual baroclinic tides (really internal waves) behave, the collective behaviour is less well understood. There does appear to be some *statistical* features we could observe (and even derive), the notable one being the **Garrett–Munk spectrum** (ref), where the energy spectrum $E(k_z)$, i.e., the energy distribution in terms of wavenumbers instead of in space, is (skipping a ton of important proportionality factors with the squiggle $\sim$)

$$E(\omega, k_z) \sim \omega^{-1}(\omega^2 - f^2)^{-1/2} k_z^{-2}, \tag{6.29}$$

where ω is the internal wave frequency and f is the Coriolis parameter. This relation seems to be fairly *universal*, and has been found in numerous observations. An example is given in Fig. 6.32, showing the vertical displacement spectrum, which itself scales like the above energy spectrum with some more factors, and particularly highlighting the k_z^{-2} dependence within the region where internal waves operate (remember for internal waves we have $f < \omega < N$).

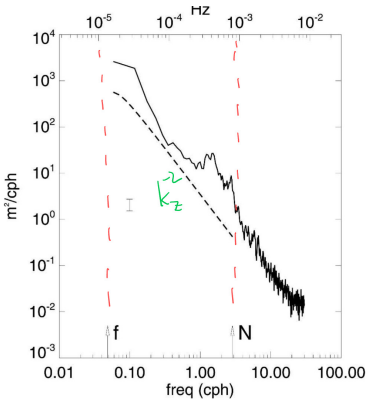


Figure 6.32: Observation of ocean displacement as a power spectrum, with the Garrett–Munk wavenumber (so frequency) spectrum dependence put in. Note also the frequency boundaries marked on since internal gravity waves should satisfy $f < \omega < N$. From Stevens *et al.* (2005), *J. Phys. Oceanogr.*, modified from their Fig. 3.

Summary and further reading

In Ch. 5 we highlighted some features of the global MOC and argued that *dynamical processes* are crucial for the maintenance of the global MOC, such as through convective motion leading to formation of deep water, baroclinic instability leading to transfer of momentum via eddy form stress, and bottom boundary instabilities leading to the diapycnal mixing that closes of the MOC. In this chapter we highlighted a few related aspects, introducing the concept of

waves and instabilities. An attempt at parcel arguments for the static and shear instabilities was presented. Waves were argued to be the building blocks for shear instabilities, and the principal quantity for the mechanistic description utilises isolated vorticity anomalies, inducing a broader velocity, which in certain configurations can lead to a pair of waves constructively interfering. Baroclinic instabilities shed eddies that lead to the eddy form stress. Tidal forcing leads to baroclinic tides, which can be susceptible to primary shear instabilities that, in turn, be susceptible to further secondary shear and/or static instabilities, eventually leading to mixing in the deeper parts of the ocean.

As mentioned in Ch. 5.2.4, it is generally accepted that dissipation or baroclinic tides arising from tidal motion accounts for around half of the energy required to 'stir' the ocean to get the upwelling we need to close the global MOC. A representation of the resulting diapycnal mixing arising from the breaking of internal tides however is still an ongoing area of research, because part of the problem here is these baroclinic tides, which are really small-scale internal waves, propagate and do not necessarily break and/or mix near where they are generated (e.g. there might be *attractors* such as that shown in Fig. 6.33 that waves preferentially break at). Without the breaking and/or mixing there is no feedback onto the background state (cf. non-acceleration theorems), but if we don't know where they break then we don't know where the forcing by the small-scale acts.

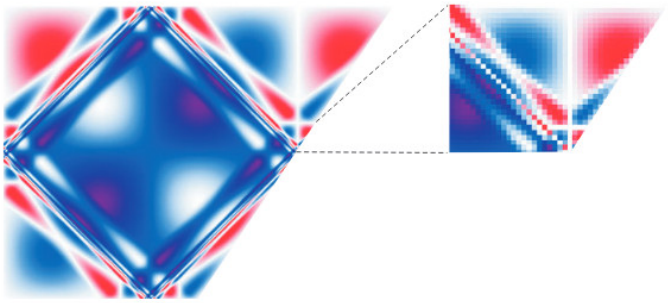


Figure 6.33: Internal gravity wave attractors in a uniformly stratified channel. Image from Maas (2005), *Int. J. Bifurcat. Chaos*, reworking of Maas *et al.* (1997), *Nature*.

The content here only very briefly touches on the broad dynamical aspects, and the devil really is in the details, which I have ignored and/or omitted on purpose, since this document was advertised to focus on the qualitative rather than quantitative aspects. We have for example not talked about the important topics such as derivation of the various dispersion relations quoted here (except for Rossby waves in Ch. 6.1.2 since it was one of the cleaner / easier examples), wave propagation and critical layer theory, stability criteria (needing

a linear stability analysis), theory of dynamic tides (necessarily involving wave propagation, geometric considerations, resonances, etc.), turbulent cascades, how do we use the Garrett–Munk spectrum and/or derive it (ref), sub-grid representations of these processes, and many others. The interested reader is instead referred to refs and references within for some of these topics.

Chapter exercises

1. If I have two waves are described by

$$\cos(kx - \omega t), \qquad \sin(kx - \omega t),$$

what is the magnitude of the phase shift between them (give this ideally in radians, but degrees is ok)? What is the phase difference of the cosine wave relative to the sine wave (i.e. what is the sign, with the sine wave as the reference)?

2. Compute the phase and group velocity for the one-dimensional Rossby wave $\omega = -\beta/k$, and use this to argue that the propagation directions of waves and wavepacket shown in Fig. 6.4 are consistent.

3. Show that $L_d = \sqrt{gH}/f_0$ really is a length-scale.

4. Which 'density' (cf. Ch. 2) should we be thinking about when we are talking about the buoyancy frequency N in Eq. (6.22)? Why?

5. For a pure internal gravity wave we might have a waveform that looks something like[37] $\psi \sim e^{-iNt}$. Making use of $i^2 = -1$, what happens to the waveform $N^2 < 0$, i.e. $N = N_r + iN_i$ as time increases, where $N_{r,i}$ are real quantities (so these are respectively the real and imaginary parts of N)? What if $N^2 > 0$?

[37] I am omitting the wavenumber and definition of ψ because it doesn't really matter what they are for this exercise.

6. Show mathematically that, for internal waves given by the dispersion relation Eq. (6.23), the group velocity is along the phase lines, noting the direction perpendicular to phase lines is given by the wavevector $\boldsymbol{k}$.

7. compute + meaning of Rossby deformation radius/wavenumber

8. Look up the phenomenon of *brinicles* (underwater finger of death) and describe by text and/or pictorially how these function as an instability (there is something in the Blue Planet episodes narrated Sir David Attenborough).

9. Carry out a similar vorticity argument to the Rossby wave propagation mechanism given in Ch. 6.1.7 but for a pure internal gravity waves with $N^2 > 0$ (assume a single interface for simplicity). Here it is not vorticity that is conserved but buoyancy, so you have to consider how changes in the buoyancy from the displacement of the waves lead to vorticity anomalies, and how this then induces an advection and in turn a propagation of the waves (there are two cases to consider, giving you the right and left propagating waves). See Harnik et al. [2008] to get started.

10. As above, but now argue along the same lines of Ch. 6.2.2 how pure internal gravity waves could phase-lock and lead to mutual amplification, and it has to be the counter-propagating pair that does the job. See Harnik et al. [2008] and/or Rabinovich et al. [2011] to get started.

11. Sticking with the internal gravity wave case on a single interface, but consider the case $N^2 < 0$, try and repeat the arguments two questions ago to get an interpretation of the Rayleigh–Taylor instability in terms of vorticity arguments. See Rabinovich et al. [2011] for pointers.

12. Suppose for whatever reason we had a uniform gravitational permeating through space (this is of course not possible but bear with me), do we still get tides, and why?

13. (After Lamb 1932) Suppose you had two moons of the same mass instead and they are 180° (or π radians) apart, at the same distance away from the center of Earth. What would happen to the tidal amplitudes and frequencies due to the moon? Justify your answer. (Hint: draw out Fig. 6.28 again.)

14. As above, but what would happen to the lunar tidal frequency if the second moon was only 90° (or $\pi/2$ radians) apart instead? Ignore the fact this means the barycenter has to shift, and assume rotation of the two moons is still around the centre of Earth.

15. Certain exoplanets are closer to the parent start and are said to be *tidally locked*. Look up and describe pictorially what this actually means.

16. (More open ended) Does the Earth induce a tide on the Sun? If so, estimate the magnitude of deviation we might expect to get.

17. tidal resonances

18. wave breaking at the beach

7 *Observations*

So far we have focused on theories and made use of data as if it was just something we can get easily when we want it. Here we talk a bit about how the data is actually acquired, some of the instruments and principles involved, and how these fit in with the theories in the preceding chapters. There is some argument the following discussion could actually have been placed at the beginning of this document, to frame the problem from the beginning and then justify the need for the theories. I can see how that would work, but I didn't do it mostly because I am a theorist and personally don't do observations, so it was my choice to frame the narrative this way. From a practical point of view (at least to me), ocean observations are inherently costly and difficult to do (see for example Ch. 7.2.2 later), and so we almost never do observations for the sake doing it, and from that point of view I am of the opinion that the basic theory should come first: you probably want to know what you are looking for before you dump a big chunk of money to go look for it.

I also don't claim to understand all the intricate details to do with observations and with the equipment. The sketches below are biased towards the equipment and applications I find particularly interesting, and supported by physical principles that I think I essentially understand. The reader should be warned the narrative below is almost certainly going to make observations and equipment design/constructrion sound substantially *easier* than it actually is. The underlying principles are maybe not that difficult to understand, but the devil is absolutely in the details, such as calibration of instruments, controlling the uncertainties, conversion of raw data to processed data, observation campaign design and associated logistical aspects (e.g. permits, visas, rights etc.), and so forth. The following is only meant as a sketch of the ideas and provide some interplay between theory, observation, as well as numerical modelling of the ocean. The reader is referred to Wunsch [2015] and Talley et al. [2011] for example for excellent expositions on the theory and observations play together in the context of physical oceanography.

7.1 *Prelude: some things to bear in mind*

7.1.1 *What might we actually want?*

A few examples of the things we want to observe were highlighted by the previous chapters (in no particular order):

- T and S, and other tracer properties to understand watermass properties (Ch. 5.2.2);
- T and S to get (the various variants of) ρ for the MOC (Ch. 5.2), to obtain the thermal wind (Ch. 5.1.3), to get information regarding the dynamics (Ch. 6), etc.;
- velocities $\boldsymbol{u}_g$ and $\boldsymbol{u}$ to monitor the western boundary strengths (Ch. 5), to understand/rationalise the watermass properties (Ch. 5.2.2), infer for the gyre circulation (Ch. 5) and MOC (Ch. 5.2), to get information regarding the dynamics (Ch. 6), etc.;
- SSH to get (various) sea level (Ch. 3.1.1), for tides (Ch. 6.3), to infer for $\boldsymbol{u}_g$ (see Ch. 7.5.1 later), etc.;
- g to get the SSH of interest (see Ch. 7.3 later),
- ...

We want to do observations partly to utilise the theories and/or models to tell us something we don't know or might not be obvious to directly observe (e.g. the geostrophic flow $\boldsymbol{u}_g$), but also to constrain, test and improve the existing theories. Usually the kind of observations we want to do and the data we want to obtain depends on the scientific questions in mind; especially in oceanography, often it is too expensive to observe things for the sake of it (see for example Ch. 7.2.2 later). In that regard, it is important to consider, given the scientific question(s):

1. what do we actually want?
2. what data do we need?
3. if we can't get that data directly, can we make do with something else?
4. how good (e.g., reliability, accuracy) does the data need to be?
5. how much data might we want (e.g. for computing statistics)?
6. how best to get the data (e.g., practicality, cost etc.)?

Answers to the above questions often lead to refinements of the observational system design, equipment design, better theories/models, better ways to do data analysis, cheaper and/or more practical methods.

7.1.2 *Data, errors and uncertainties*

There are various other important aspects to do with the *data* that is not elaborated in detail here but the reader should bear in mind. As an illustrative example, lets take the problem of measuring chlorophyll-*a*, which is important for ecological problems:

- Is the data actually of the quantity of interest, or is it based on a **proxy**, i.e. we 'measure' something by measuring something else instead? Chlorophyll-*a* is sometimes measured as the 'greenness' from an optical satellite observation, and converted to a concentration by a model, in contrast to taking direct samples and measuring a concentration. The former case provides substantially much better spatial and temporal coverage and is over longer time-scales less costly, but presumably subject to much larger errors and uncertainties.

- Is the data **raw** or **processed**, and if it is processed, *how much* has it been processed? With both satellite and water samples, there is the raw signal that, by itself, may not be that useful directly. Some calibration, error control etc. is required, which means some touching up of the data. The satellite data would get further fed into a model to convert into the quantity we are interested in. In some instances both of the data might be put through further processing for ease of use, for example via averaging and/or interpolation onto a regular geographical grid (**gridded data**), as oppose to satellite tracks or isolated spatial data points where the sample took place. Errors and uncertainties are invariably introduced at these steps.

- What is the error margin and what is the uncertainty? The use of proxies and the degree of post-processing done on the data would lead to variations in the errors and uncertainties.

Depending on the question, the choices we make might differ. In this example, if one was only interested in the broad distribution, then chlorophyll-*a* from a visual proxy is probably fine. However if one was questions that require detailed in-depth quantification, then the choice of employing a proxy might not be appropriate.

7.1.3 *Difficulties with ocean observations*

In principle we might know what data we want, but that doesn't mean it is easy to get! Ocean observations are difficult and/or costly for a variety of reasons:

- Unlike the atmosphere, the ocean is *opaque* (Ch. 2), meaning that electromagnetic radiation (Fig. 2.5) gets scattered (Ch. 6.1) very easily. Beyond the obvious problem of not being able to visually see into the ocean, it also means we are more constrained in what we can use with *remote sensing* techniques (see later in Ch. 7.3), compared to the atmosphere say. This also means equipment such as floats and drones that go underwater has to come back up to the sea surface periodically to receive further instructions and/or passing of data by satellite links. This is extra work done against buoyancy and potentially exposes the equipment to rough conditions near the ocean surface, affecting the cost, the equipment's longevity, and poses challenges for operating the equipment.

- Equipment that go underwater potentially has to work in regions with *very high mechanical pressures* (Ch. 3.1.2) induced by the environment. From one of the Ch. 3 exercises we get that pressure increases by 1 atmospheric pressure with around every 10 m of seawater, meaning that if we go down to 1000 m depth then the equipment needs to be able to withstand a 100 atmospheric pressure[1]. Fig. 7.1 shows the before and after of a stryofoam cup that goes down with the CTD (see Ch. 7.2.1) subject to the natural pressure increasing with depth. The related equipment does need to be designed and constructed to operate in a wide range of conditions, driving up costs.

- Sea water is chemically **corrosive**, which is a particular problem for things made with iron content. As mentioned in Ch. 2, since there are salts dissolved in sea water, sea water conducts electricity, leading to enhanced oxidation of iron into iron oxide, which then subsequently forms rust, leading to structural damage. Fig. 7.2 shows a marooned ship that is particularly badly hit by rust, with notable structural damages, such as holes on the haul.

- Besides sea water, marine organisms can and will latch on to surfaces if given the chance, leading to **biofouling**. A common example of this in everyday life would be mold on a wall. Fig. 7.3 shows a current meter with mussels covering the equipment. Physically the organisms latching onto equipment leads to added weight, which can affect the movement of the equipment and/or

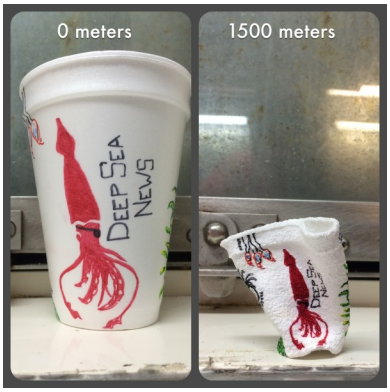


Figure 7.1: Styrofoam cup before and after being lowered to 1500 m depth in the ocean. There is apparently a tradition of making these crushed cups as a souvenir when going on research cruises. Taken from www.deepseanews.com.

[1] Taking $\rho = 10^3$ kg m^{-3}, a 1 m by 1 m by 1000 m cuboid has mass 10^6 kg. That is roughly equivalent to the total mass of 50 double decker buses, each with a mass of 20 metric tonnes.

Figure 7.2: Amorgos Shipwreck of Olympia. From taosailing.com.

the sensors. These organisms can excrete chemicals that attack the equipment and/or sensors, so extra work needs to be done to combat biofouling particularly for equipment that operate nearer to the surface of the ocean (e.g. moorings and ships)[2].

- Everything mentioned above ultimately comes down to **cost**. Iron-based material are used because it is cheaper. If there is money to burn we can always just buy more equipment or get it cleaned out frequently. In reality this is never the case and money usually acts as one of the leading order constraints on what can be done. To throw some numbers out there (in units of USD, but on the understanding that they are constantly evolving and probably out of date), a seaglider costs on the order of 100k each, Argo floats are around 20k each, a satellite probably[3] (!!!) start at around 100m, and renting a ship for a research vessel is around 30k *per day*, and scientific staff salaries are extras on top! It is true in the times gone by it was possible for wealthy people to carry out self-financed expeditions (e.g. Nansen, von Hulmboldt, Prince Albert I of Monaco), but it is generally true that nowadays these activities are co-ordinated by government backed entities, at the national and international level[4]. We will see some of these joint efforts later on in the chapter (Ch. 7.4).

[2] AUVs for example move up and down over large enough extents to kill off most of the nasties that do latch on. However, this doesn't stop them being smacked by things like seals, sharks, dolphins and octopus.

Figure 7.3: Zebra mussels on a current meter. Image from NOAA.

Figure 7.4: A moldy sea cucumber (?) feeding off of a couch? From personal collection.

[3] Probably underestimating this one!

[4] Elon Musk being a notable exception.

7.2 *In-situ observations*

With the above in mind, lets talk a little bit about the equipment. I have chosen to split the discussion into **in-situ** observations and **remote sensing**, the former to mean we actually go into the water to measure things, and the latter to mean we use the reflection of signals to infer for various quantities. This is not to say remote sensing things are any less accurate; the magnitude of uncertainties you can get with remote sensing and particularly satellites is actually fantastic and really does boggle the mind (in a good way). The approaches are just different, and each have their own advantages and disadvantages. I will start with the in-situ equipment and will be biased towards equipment that I am aware of, and of those I am aware of.

7.2.1 *Equipment*

Tide gagues

Tides have been observed for a long time, although there is a bias to be near coastal settlements and ports, for various practical reasons

some of which have already been mentioned in Ch. 6.3. Tide gauges measure tend to measure sea level *relative to the bottom bathymetry* (see Fig. 3.4); this may or may not be what you want depending on the application (from a dynamics point of view you might care about the sea level relative to the geoid instead). The way tide gagues work are theme and variations of something like Fig. 7.5. Assuming you know the depth of the bottom bathymetry, the most basic way is to construct a ruler or some physical measuring device, stick it in the water, and read off the water level as appropriate[5]. A slightly more sophisticated way would be to put something buoyant into the water, and encase it in a shell with water levels marked on the shell (so like a thermometer), from which you can read off the sea level as the buoy moves up and down with the rising and falling seas. More modern constructions have the buoyant float attached to some mechanical or electronic device that feeds the sea level data back at the station, which is then recorded.

By construction there is a bias towards data near coastal regions, and the method doesn't adapt that easily to the deep ocean, partly because it's harder to construct the analogous ruler and to maintain the stations. Nowadays, satellites provide much better coverage of sea level spatially (Ch. 7.3), but tide gauge data predates the satellites and goes back quite a long time. See Ch. 7.5.2 for a reconstruction of global sea level over the past century or so using the combined data.

[5] You still see these at ports for example to give a quick indication of the water levels.

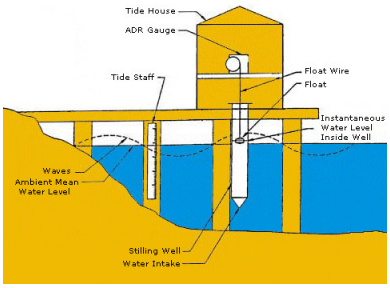


Figure 7.5: Old style tide gauge station schematic. From NOAA.

Reverse thermometers

These are like regular thermometers and measure temperature, by knowing how much a substance expands and contracts with changes in temperature (e.g. you know the coefficient of expansion α of mercury, from which you get the density; Ch. 2.3), and from that construct a measuring device. Unlike standard thermometers, however, the reverse types are constructed with loops such that when the thermometer flips over, the measuring substance gets trapped and can't flow back into the chamber. The way you then use it is you attach it to a rope or something similar, dip the thermometer into the surroundings (at whatever depth you decide), let it equilibrate and settle, then flip the thermometer, and you take it out and read it. Since the liquid doesn't flow back into the chamber, you get a reading of the temperature at the appropriate depths.

One thing to be careful about is the difference between in-situ temperature (Ch. 2.3.3) vs. potential temperature (Ch. 2.3.4), the former being higher than the latter. Depending on the depths of interest, it may be worthwhile having a pair of reverse thermometers, one as is, and one in some pressurised casing to isolate the thermometer

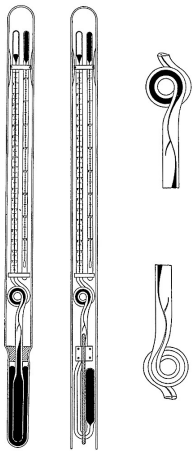


Figure 7.6: Schematic of unprotected and protected reversing thermometer. From *The Oceans Their Physics, Chemistry, and General Biology* by Sverdrup, Johnson & Fleming (1942).

from the surrounding water pressures. The pair then gives in-situ temperature and potential temperature (normally referenced to sea level). From the differences one can infer for the pressure, and in turn use pressure to get depth assuming hydrostatic balance Eq. (3.1b).

Nansen bottles

Nansen[6] bottles take water samples and work on a conceptually principle as the reverse thermometers, where the bottles don't take in water until they are flipped. So what you do is to attach the bottles to a rope or something similar, lower the bottles in normal configuration into the water, and since the bottles are denser than the surroundings even when empty, they sink. When the bottles reach the depths required as dictated by their position along a rope of known length, you somehow get them to flip, after which it takes in water, seals itself when the bottle is full, and you pull the bottles back up. Then the water can be analysed at a station or on board of ship as appropriate.

It used to be that you pair the Nansen bottles with reverse thermometers, and you chain them along a long rope (say 10 of these), and you probably lower them gradually into the ocean (otherwise they can smash into each other). You wait a while for the equipment to fall accordingly to the depths you care about (but with the understanding that they will be moved around a bit by the currents), and let the thermometers equilibrate. Then the traditional way of doing it is to send down a mechanical weight called a *messenger* (see Fig. 7.7), and this weight hits the highest bottle, which flips the bottle and thermometer. When flipped, this releases another messenger attached to the bottle, and it falls down the rope and hits the next bottle, and so on[7]. After sufficient time has passed, you pull it all back out, then you have some data. This is quite time consuming as the whole process can take a few hours depending on how long the chain is. Additionally, water of course is very heavy, so extra equipment is required to pull the whole chain back out up to the ocean surface, which places an extra constraint for example on the ships that could be used with this kind of equipment.

[6] Fridtjof Nansen (1861–1930) was a famous Norweigian explorer, scientist and diplomat, and one of the revolutionaries in equipment design for Arctic and Antarctic exploration. He was awarded the Nobel Peace prize in 1922 for his work on behalf of refugees displaced particularly by the First World War. The Nansen bottles were subsequently refined by Shale Niskin (1926-1988), whose design is the one that is usually employed now.

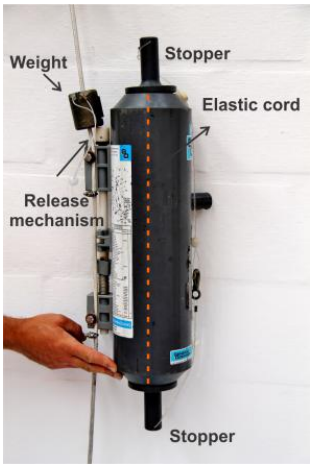


Figure 7.7: A picture of a Nansen/Niskin bottle. From Flanders Marine Institute (VLIZ).

[7] Like falling dominoes, and conceptually similar to the principle of mathematical induction if you what that is.

Current meters

The mechanical current meters measure speed and direction (and therefore a velocity) of the water currents, by means of a device like in Fig. 7.8. The faster the current, the more the turbine rotates, and we can get a speed by means of electrical induction or otherwise (e.g. the higher the rotation rate, the larger the induced current). To get the direction, the black rudder causes the device to pointed in

the direction of the flow, and with a compass for example, gives an indication of the current's direction.

An immediate disadvantage of this model is that these current meters only really give the horizontal velocity. Another is that these have moving parts, so it is subject to fatigue and may need replacing; while these devices are presumably not that expensive, the annoyance comes from the actual going out and replacing the devices. The acoustic version of these current meters (Ch. 7.3) are usually preferred. A more subtle and generic problem with measuring currents is the wish to separate the total 'noisy' signal into something that we can attribute to tidal flow, geostrophic flow, waves, instabilities, and so forth. This is not a trivial task, and it is probably a reasonably uncontroversial statement to say we still do not have a complete understanding on how to do the desired decomposition in theory and in practice[8].

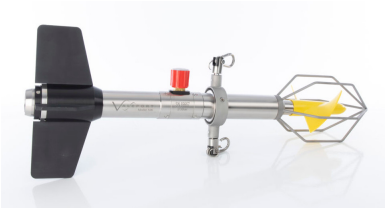


Figure 7.8: A mechanical ocean current meter. Image from www.valeport.co.uk.

[8] Tides we might be ok with, but it is not immediately obvious to me how we would do it for the others. A way to do this is through decomposition of the signal into frequencies, by a Fourier transform, wavelet decomposition, low/high pass filter, or otherwise.

XBTs

The mechanical bathythermographs[9] are fairly big torpedo like devices that you lower into the water and it measures temperature and depth, the latter via some sort of model for the object falling through water, up to a depth of around 300 m. Ships sometimes have these attached to the vessel if they use sonar; see later (Ch. 7.3). The advantage of using these is that you can take observations while the ship is moving. The disadvantage is that you normally want these back, and them being big and heavy can be very dangerous to retrieve[10].

[9] Coincidentally designed by Spilhaus of the Spilhaus projection in Fig. 5.1.

[10] When these were still regularly used people on ships take turns to retrieve these, because it seems to be universal that no one likes doing the retrieval...

The mechanical ones are not really used anymore, because there is the **eXpendable BathyThermograph** (XBT), which does basically the same job but is quicker and easier to use, and can go down to depths of up to 1000 m. These things are shown in Fig. 7.9 and is just a small bathythermograph attached to some copper wires, so as it falls and gets dragged by the moving ship it records the temperature and pressure of the surroundings, and passes the signals back through the wire. Eventually the wire runs out and snaps, and the devices are considered lost. These are fairly cheap devices, easy to use, and does not require special dedicated equipment on the ship, but the accuracy is perhaps not as good.

Figure 7.9: XBT being launched (left) and schematic of XBT (right). From NOAA (left) and NASA (right).

CTDs

The **Conductivity, Temperature and Depth** (CTD) sensors essentially replace the reverse thermometers, where the sensors pick up conductivity (to get salinity; see Ch. 2), temperature and pressure (and therefore depth). The standard set up of what are normally referred

to as CTDs is given in Fig. 7.10 in a *rosette* configuration, where the CTD sensor is in the middle, with water sampler bottles surrounding it; these are essentially updated versions of the Nansen bottle plus reverse thermometers. You chain it and lower the whole rosette into the water, as and as it is lowered, the sensors do their observations, and at set depths water enters certain bottles and sealed accordingly. When the sampling is done, the whole rosette is pulled back out and retrieved onto the ship. The advantages and disadvantages are similar to the Nansen bottle and reverse thermometer combo. It is probably uncontroversial to say that CTDs are the standard workhorses of the modern day ocean observation.

Figure 7.10: CTD profiler (in the center of rosette for this set up) and Nansen bottles surrounding it. Image from OCEAN-HK.

Floats and drifters

A more flexible and lighter version of a CTD rosette would be a float with a CTD sensor attached to it (e.g. top of Fig. 7.12). The float could move up and down the water column by something that alters the buoyancy of the float. This could be done for example by the means of a pump that moves oil between an external *bladder* (see for example Fig. 7.14): when oil is pumped into the external bladder, the object has the same mass but occupies a larger area, thus decreasing its density and becoming more buoyant, and vice-versa (this uses substantially less energy than taking in and ejecting water like a submarine would). With this, the float can move up and down the water column, but essentially gets carried around by the flow. With the CTD sensors, the float can then give you profiles of the salinity, temperature at various locations and depths. The communication of data and locating the float is usually by remote means, such as satellite communication, but requires the float to be at the surface of the ocean and be exposed to the rougher surface ocean conditions.

Drifters on the other hand usually refer to the things that float on the surface, though not all of these are in the ocean intentionally. By knowing how the drifters' (as well as floats up to a point) location change over time one could get a sense of the currents that are carrying them, from a *Lagrangian perspective*. Some interesting cases are as follows:

- Message in a bottle. You might think the standard cliché of someone stranded on a deserted island sending off bottles[11] wouldn't have a place in observational oceanography. Well one case of use was in the early 20th century to probe the circulation of the North Sea in Europe. The British marine biologist George Parker Bidder threw in about a 1,000 bottles into the North Sea to probe the circulation, each with a (hand-written?) message in English, Dutch and German asking anyone who picks these up to send a

Figure 7.11: An Argo float being thrown off a ship (top) and some rubber ducks at the Ken-ducky derby (there is a caption mismatch). Image from NOAA (top) and Cassie Marshall (bottom).

[11] Inspiring the classic song by *The Police* fronted by the legendary Sting.

post-card back to some address in the UK indicating where and when they found it, for a reward of a shilling. Most of these were recovered quickly but one was found on an island in Germany as recently as 2015.

There was another story that I can't seem to find any more, of some bottles being thrown into the ocean, with messages within it saying something like "*please note the date you found this, write it on the piece of paper in the bottle, put it back into the bottle, seal it and throw it out again*", as a way to get a feel for the ocean circulation patterns.

- Parsnips. Besides being nutritious and tasty, these root vegetables have also helped our quest to understand the current and its dispersion characteristics. To quote Richardson and Stommel [1948],

  "We have observed the relative motion of two floating pieces of parsnip, and have repeated the observation for many such pairs at different initial separations...in the sea we used floats of parsnip because it is easily visible, and because it is almost completely immersed so as not to catch the wind which, moreover, was slight."

  Here they were interested in quantifying the effective diffusivity (Ch. 3.4) related to how particles have a tendency to spread away from each other, formalised by Taylor [1921]. Recall that while we know the molecular diffusivity, it is the effective or eddy diffusivity that dominates the contribution, so it is of interest to find some way to quantify this. This kind of *Lagrangian particle* approach is regularly used in numerical models as a way to quantify eddy induced diffusion (e.g., Abernathey et al. [2013]).

Figure 7.12: A Russian variety of parsnips. Image from www.adaptiveseeds.com.

- Floating seaweed. To quote Langmuir [1938][12],

  "On August 1927, when about 600 miles from New York on an Atlantic crossing to England I noticed there were large quantities of floating seaweed, most of which was arranged in parallel lines with a somewhat irregular spacing ranging from 100 to 200 m. These lines, parallel to the wind direction, which I shall call streaks, often had lengths as great as 500 m."

  What he is describing here led to his formulation of wind and wave induced circulation patterns that are now known as **Langmuir circulations**. These are known to have an important consequence for the ocean mixed layer, thus mediating exchanges of

[12] Irving Langmuir (1881-1957) was an American chemist, physicist and engineer who made fundamental contributions to atomic theory and plasma physics (e.g. Langmuir waves). He won the Nobel prize in chemistry for his contributions to surface chemistry.

momentum, chemicals and other tracers between the atmosphere and ocean, with impacts for biology via increasing supply of nutrients in upwelling zones, as well moving plankton around.

- The case of friendly floatees. In 1992 a ship from Hong Kong carrying plastic toys (duckies, beavers, turtles and frogs) bound for Tacoma, Washington encountered a storm, and one of the containers with these toys in fell overboard and released about 30,000 of these guys into the ocean. Over the next 20 years of so they keep being found all over the world[13], and have helped test and refine our understanding of how the global ocean circulation functions. Something similar happened to Lego pieces.

Floats and drifters are normally regarded as disposable, because the cost of retrieval generically outweigh the cost of dedicated ship time needed to retrieve these.

[13] And these are now prized collector items and can fetch quite a bit on the market, since these toys are no longer manufactured, and because of the media attention surrounding them.

AUVs

The **Autonomous Underwater Vehicles** (AUVs) are basically drones with sensors (e.g. CTD, current meters etc.) attached that you can program to undertake certain tracks underwater for taking observations. These are used for example in inaccessible regions to boats for either physical and/or human reasons (e.g. under ice, pirates, political tensions). The AUVs receive further instructions and/or pass data back via satellite links when they surface, although in certain cases when that cannot happen (e.g. under ice), the data can be stored on the onboard system. These you probably do want to retrieve where possible, since these are usually deployed as part of a research cruise (and thus places a requirement on the ships that could be used), or because there is valuable data on board you want to recover.

Figure 7.13: Boaty McBoatface, a Autosub LR (top) and a Kongsberg sea glider with Bastien Queste (now at Gothenburg) for scale (bottom). From UK National Oceanography Center (top) and Sergey Piontkovksi (bottom).

There are many different types and designs of AUVs depending on the intended purpose, such as depths of operation, coastal vs. open ocean applications, the length of observation. The sensors they have usually consist of at least the CTD, but the end choice depends on a balance between scientific question and logistical aspects, such as battery life, weight and space for the sensor. Fig. 7.13 shows two examples of these AUVs. The top one is the (in)famous *Boaty McBoatface*[14] long range autosub, which moves horizontally by mechanical means (notice the propeller at the back) and vertically by buoyancy changes. Boaty is currently housed at the National Oceanography Center in the UK, and can operate down to depths of 4000 m. It has participated in various missions in the Antarctic (to look at underwater mixing), North Sea (for carbon sequestration),

[14] So NERC in the UK at some point commissioned a new research vessel with the intention to replace the *RRS James Ross Clark*, so in their infinite wisdom decided to poll the internet for names. *Boaty McBoatface* came out the clear winner, but the eventual ship was named *RRS David Attenborough*, after the famous British broadcaster and natural historian who is particularly well-known for his documentaries of the natural world.

and notably under-ice missions in the Arctic and the Weddell sea.

The bottom one in Fig. 7.13 is a **seaglider**. Unlike Boaty, which is mostly mechanically driven, seagliders are kind of like floats and move around via adjusting its buoyancy. A schematic of the Kongsberg type seaglider shown in Fig. 7.13 is given in Fig. 7.14. Oil gets pumped in and out of the external bladder, thus changing the glider's volume and in turn density, causing the glider to rise or sink until it is neutrally buoyant. Unlike floats, however, wings and rudders are attached to the external part of the glider, so depending on the angle of attack as the glider rises and falls, there is associated horizontal motion because there is a preferred direction with smaller resistance against the water[15]. The different angles of attack are made possible by an unevenly distributed battery than can shift and rotate around, via shifting the center of mass (adjusting the *pitch*, so pointing up or down) and by turning the uneven center of mass (adjusting the *roll*, so the glider rolls to the left or right). With these, the glider can 'fly' a section and take measurements along the way.

[15] Hence the naming, because gliders in the sky work by a similar principle.

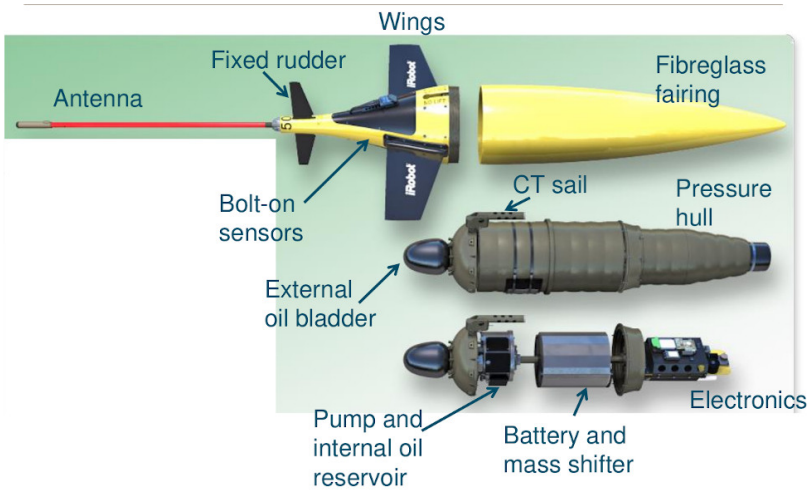


Figure 7.14: Components of a seaglider. Slide from 2018 ATSC winter school on gliders at University of East Anglia, UK.

Since there are minimal moving parts, and energy is only consumed from the battery in pumping/sucking out the oil, moving the centre of mass, and by the sensors, seagliders can engage in longer continuous observations than some of the mechanically driven AUVs. For models designed for open ocean applications, a continuous operation period of three to six months is not unheard of. The coastal models require a much bigger battery for similar performance partly because by design they don't dive as deep (and usually don't move as far horizontally), so more dives are required, which uses more energy.

7.2.2 *Moorings and ships*

The aforementioned equipment are usually deployed from a coastal station or a ship (which is a floating station), or attached to a mooring. A mooring is an anchored device normally for long term continuous monitoring, and an example of this is Station Papa in the North-East part of the Pacific (see Fig. 2.7*a*), shown in Fig. 7.15. The surface buoy is connected to the anchor by a wire, and along the wire there are various sensors appropriately placed to measure temperature, conductivity, currents, pH, and so forth. The buoy at the top of the mooring allows communication of data back to the land stations for data processing.

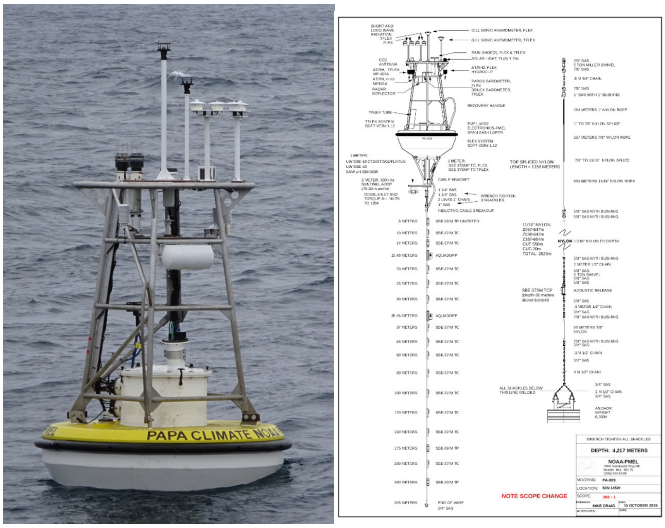


Figure 7.15: Station Papa (left) and schematic (right) in 2015; see also Lec. 5. From PMEL NOAA.

These moorings provide long term continuous monitoring and are an integral part of the global ocean observing system. One thing to bear in mind though is that these do get moved around by the wind and the currents, and since they are anchored down but the line is not rigid (and you really don't want it rigid for structural integrity reasons), there are going to be errors and biases associated with shifts in the location and depth the data is actually coming from. These moorings also need replacing every so often, so there is the associated maintenance, replacement, ship time and personnel cost; station Papa seems to be have been replaced/maintained every year.

Ships have been the standard workhorse of ocean observation for a long time. Ships can be broadly separated largely into two classes, those that are designed for ocean observation (e.g. **research vessels**), and those that are not. The dedicated research vessels are essentially floating stations, possibly equipped with laboratories, equipment for deploying the aforementioned gadgets (e.g. winches to lower and lift the CTDs), designed for longer periods at sea, and other

Figure 7.16: The German RV Sonne when it was docked in Hong Kong. Image from personal collection.

specialist equipment (e.g. ice breaking capabilities for venturing into sea ice regions). Fig. 7.16 shows an example of a research vessel, the German RV Sonne. Notice the size of the ship, and that they are equipped with winches (in the lower picture). These dedicated ships are very expensive, with an operating cost of tens of thousands of US dollars *per day*, not including cost associated with scientist salaries, extra equipment and so forth. Nevertheless, sometimes ships are necessary, and they still provide the 'gold standard' in terms of ocean observation.

Ships of opportunity refer to the non-dedicated ships, such as merchant ships and fishing boats. Since these ships travel around the oceans anyway, they sometimes participate in helping the ocean observation effort. Some merchant ships may be given XBTs, equipped with bioacoustic sensors, devices to measure atmospheric conditions and so forth, which can be used by the members on board to do observations. Fisherman may get provided sensors on their fishing boats or nets that can take observations, and these data are reported back in return for monetary compensation. Ships of opportunity is substantially cheaper than dedicated research cruises in the grand scheme of things, and observations are taken along the same tracks repeatedly, giving data at an increased temporal resolution. The disadvantage is that there are only so many things you can do on these ships, the data quality suffers somewhat, and observations are biased towards larger shipping lanes by the nature of how ships of opportunity operate.

Figure 7.17: The MV Black Marlin. Not really a ship of opportunity, but I just really wanted to put "a shipping ship shipping shipping ships" in. Image from Wikipedia, user Kees Torn.

7·3 *Remote sensing*

One of the easy (and perhaps slightly cheap) criticisms of in-situ observations is the limited spatial coverage. Nowadays this is com-

plemented by observations done via **remote sensing**, but requires more theory and substantially more accurate instruments. I am going to talk about a few of the uses I somewhat understand here.

Remote sensing refers to 'sensing' by the signals the environment or objects emits or reflects. A trivial example is us being able to see things with our eyes: light hits an object and is reflected into our eyes, triggering a reaction that sends a signal to our brain, so that our brain knows there are objects at some place, even though we haven't interacted with it directly so to speak. Other examples include:

- seismology, probing the interior structure of the Earth by the wave signals we observe;
- Magnetic Resonance Imaging (MRI), often used in medical purposes;
- metal detectors at air ports, detecting the reflection of emitted waves from metallic objects;
- X-ray imaging, to see the interior of a body without opening up the body (Fig. 7.18);
- ultrasound, same as above, but using sound waves (Fig. 7.19).

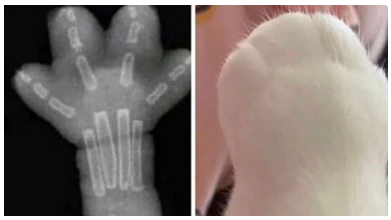


Figure 7.18: Probably a kitten paw X-ray? (Bones may appear detached in very young animals in X-rays; see `veteriankey.com`.) Image taken from `cheezburger.com`, source Twitter account appears suspended.

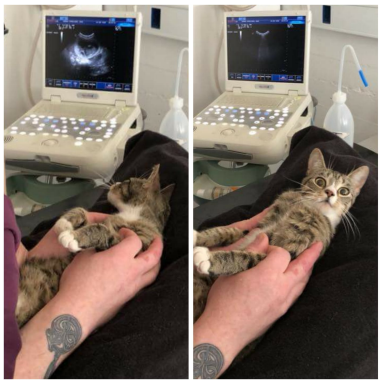


Figure 7.19: Ulla shocked (?) at finding out she is pregnant from the ultrasound. From Dyrenes Venner Greenland's facebook page.

The main principle for remote sensing are relate to wave dynamics (Ch. 6.1). We might shoot some waves out (sound waves or electromagnetic waves), and we want to know how these signals are modified when it hits an object and is reflected back, and from that infer about the object and/or the medium of interest. The advantage of remote sensing is usually we get a reasonably good coverage and is usually non-invasive. The disadvantage is on the accuracy of the sensors (just because you can do something in principle doesn't mean it is accurate enough to be of any use!) and the possibly high initial cost.

7.3.1 *Acoustics*

Recall that the ocean is opaque, so electromagnetic waves (Ch. 2 and Fig. 2.5) such as visible light, microwaves and ultraviolet waves get scattered very efficiently by seawater. So while above ground we might be able to use things like radar for purposes such as imaging and communications, this is basically out of the question in seawater. However, there are a few things that we can do with **acoustic** waves (i.e. *sound* waves).

Sonar

SOund NAvigation and Ranging, i.e. **Sonar** is basically echolocation, and the idea is illustrated in Fig. 7.20, for bats, dolphins and submarines. You send a ping out, the wave hits the object, and reflects back to source. From the time elapsed t and by knowing the speed of sound c in the medium of interest, assuming the object is not moving or hasn't changed its position that much since the elapsed time, you can get the distance L between the source and object as

$$c = \frac{2L}{t} \qquad \Rightarrow \qquad L = \frac{ct}{2}. \tag{7.1}$$

(Convince yourself you need a factor of 2 because of how L and t are defined.)

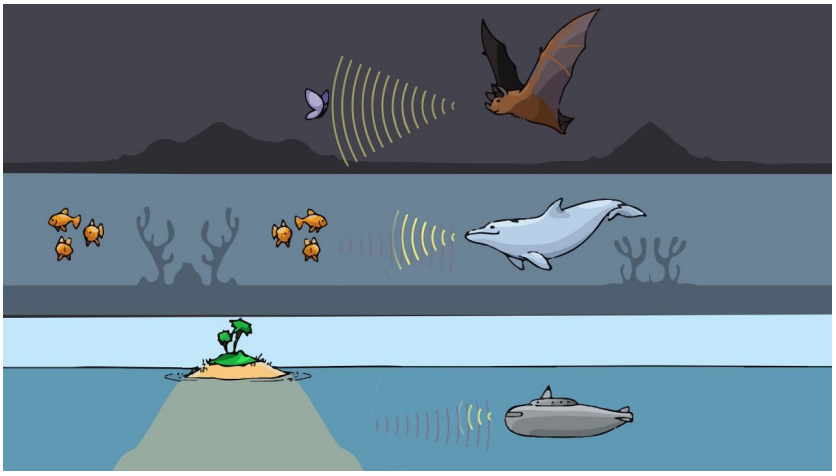


Figure 7.20: Schematic of Sonar. Image from the Smithsonian National Museum of Natural History.

While the human brain can't do the relevant calculations that fast, computers can, and when programmed properly, the computers can adjust automatically for changes in the sound speed[16] and the relevant distances. If there are fixed sound sources in the ocean (e.g. from known moorings for example), then similar principles could be used for navigation purposes.

[16] Changes in south speed in water is most via changes in density, which in most of the ocean is controlled by changes in the temperature, hence why there are might be bathythermographs attached to submarines, as a way to sense the temperature around the region to calibrate the sonar.

Doppler shifts + ADCPs

The above discussion assumes the objects are not moving very much relative to each other, but what if they are? The resulting phenomenon is known as **Doppler shifts**, and results in the *frequency* of the wave changing from the *observer's point of view*. This is akin to the observation that when you have an ambulance sounding its siren travelling past you the observer, the siren *drops* in pitch. The schematic of this phenomenon is given in Fig. 7.21. Assuming our

neighbourhood friendly pig and the source of waves is not moving at all, then the incoming waves are reflected as usual, with no change in the frequency. Now, assuming the pig is trotting along to the right at steady speed, and the sources of waves are themselves not moving at all for simplicity. Then, from the left emitter's point of view, since the pig is moving away, the same emitted wave takes longer to reach the pig before it gets reflected, so there is a *dilation* in the wavelength. Given $\gamma = c/\lambda$, and the wave speed is not changing, if λ increases, the frequency γ has to *decrease*, i.e. a drop in the frequency, and what is known as a **red shift**. For the right hand side emitter, the opposite is true: the pig is moving towards the emitter, so the wave gets 'crushed' and the wavelength decreases, so the frequency increases, leading to a **blue shift**.

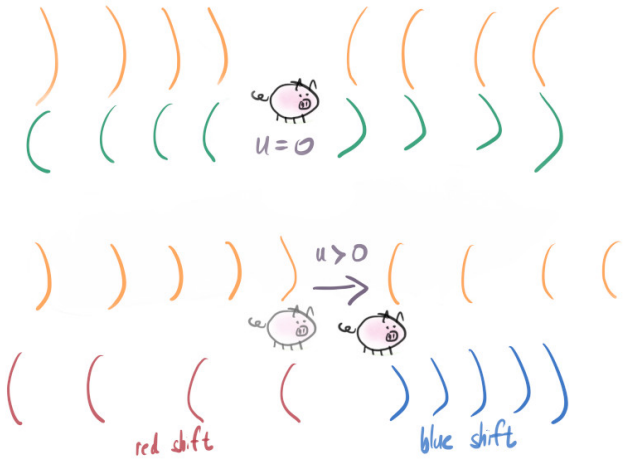


Figure 7.21: Schematic of Doppler shifting. Red/blue shift (decrease/increase in frequency) if target is moving away/towards observer. The red and blue shift comes from the fact red light is on the lower frequency / higher wavelength end of the visible spectrum, while blue is on the higher frequency / lower wavelength end; see Fig. 2.5. Compare the naming convention with Infra-Red and Ultraviolet. Principle applies generically for waves (e.g. similar principles are used to detect universe expansion).

Doppler shifts (for objects moving at non-relativistic speeds, i.e. substantially slower than the speed of light) are given by the formula

$$\gamma = \left(\frac{c \pm v_r}{c \pm v_s}\right)\gamma_0 \approx \left(1 + \frac{\Delta v}{c}\right)\gamma_0, \tag{7.2}$$

where γ is the Doppler-shifted frequency, γ_0 is the original frequency, c is the wave speed, $v_{r,s}$ are the speeds of the receiver and source of wave respectively. The latter approximation comes from assuming the magnitude of c is much bigger than $v_{r,s}$, and Δv is defined as velocity differences between the receiver and source, and is positive if source and receiver are moving towards each other (which makes sense because then the Doppler factor on the right is positive, leading to an increase in the observed frequency γ). Notice we need *relative motion* for Doppler shifts, i.e. $\Delta v \neq 0$; if both receiver and source are moving by uniformly at the same velocity, then there is no Doppler

shift. Notice also that nowhere in this paragraph is the discussion restricted to sound waves: the phenomenon applies generically to waves, with c to be interpreted accordingly depending on the waves in question (e.g. c would be speed of light for electromangetic waves).

For the ambulance example, assuming it is moving at constant speed towards and past you, the incoming observed siren you hear from the as it is coming towards you is *already* blue shifted (i.e. *higher* than what the driver of the ambulance would hear). As the amublance goes pass, *you* hear the pitch drop because of the red shifting. The ambulance driver doesn't hear any difference throughout, because from their point of view, they are not moving relative to the siren.

This kind of phenomenon can be exploited to measure velocities of ocean currents, via a **Acoustic Doppler Current Profiler** (ADCP); a mounted version of an ADCP is shown in Fig. 7.22. If the water contains some particles, there is something for the sound to reflect off of, so assuming the particle movement is dominated by advection of the currents and is moving relative to the ADCP, there will be a shift in the frequency of the reflected sound wave, and with precise enough measurements we can get a relative speed in a certain direction. With multiples sound beams at different directions and frequencies, it is possible to get a sense of the current in the surrounding regions via the received signals via a formula such as Eq. 7.2.

Figure 7.22: An ADCP mounted on a frame. Image from Wikipedia, user DopplerMusic.

The big advantage of ADCPs over mechanical current meters is the wider coverage, and the lack of moving parts (which is usually a good thing for equipment longevity, particularly in the ocean where they are susceptible to being attacked by seawater). The disadvantage is that they require things to reflect off of, so it doesn't work that well in clean water. The lack of moving parts in this case could be a disadvantage, because the surfaces are subject to biofouling, which will severely affect the performance of the equipment. There are concerns of noise pollution and its impact on the surrounding marine ecology, though these smaller ADCPs are generally not very noisy (certainly not compared to ship traffic). ADCPs are now generally preferred over the mechanical current meters.

7.3.2 *Satellites*

Satellites are objects that orbit another planet or star; in that sense the moon is a satellite of the Earth, and the planets are satellites of the star. In the following we are interested in *artificial* satellites that allows us to observe Earth in a variety of ways. Satellites make use of **electromagnetic waves** (see Fig. 2.5) at different frequencies

depending on the purpose. Note these waves travel at the speed of light $c_{\text{light}} \approx 3 \times 10^8$ m s^{-1}, and satisfy the usual wave phenomena described, such as scattering, reflection, Doppler shifts and so forth.

Different satellites orbit the Earth in different altitudes and different orbits depending on the purpose. The orbits are split as low, medium and high Earth orbits (L/M/HEO) depending on the altitude; see Fig. 7.23 for a to-scale diagram of the Earth, the LEO and the MEO. The LEO has an altitude of no more than about a third of the Earth's radius, so roughly less than 2000 km above the Earth's sea surface. At this altitude, the Earth's gravitational field strength is not so different to the Earth's surface. To stay in the orbit, these satellites need to be moving fast enough, and tend to circumnavigate the Earth is less than about 2 hours (i.e. the *orbital period* is around 2 hours). Thus, within the LEO, repeated observations can be taken quickly. A higher bandwidth of waves can be used for communication between satellites as well as with the surface of the Earth, and there is less latency/lag in the communication, because of the proximity to Earth. The drawback is that, being closer to Earth means there the angle of 'vision' of the satellites is smaller, the presence of space debris in this region can lead to satellites being damaged, and the satellite might need to do more work to stay in the intended orbit. The advantages and drawbacks are reversed if satellites are put in higher orbits.

Figure 7.23: To scale diagram of sample orbital ranges of Earth. Counting from the upper limit of the MEO, the moon is about ten lots of MEO away to the right of the diagram (not shown). Image taken from Wikipedia, user Rrakanishu.

Communication and navigation

Most of the satellites for ocean observation use the LEO but communication and navigation satellites tend to be in the higher orbits as well as being in the **geostationary orbit** at 35,785 km above the equator, the latter having the characteristic that satellites staying in this orbit essentially rotate with the Earth, so the receiving dishes on Earth can just keep pointing at the same location. At the orbits away from the geostationary one, there is some shift because of the mistmatch of the orbital period with the Earth's rotational period, but of course if we have a network of these satellites then as long as there is one satellite in the line of 'sight' of the receiving dishes on Earth, then communication is possible (and the signal will probably be of a higher quality).

Essentially satellites communicate with stations of Earth as well as between satellites by emitting some waves with some distinguishing characteristic, so that it can be identified by the receivers accordingly; much like flying a flag or wearing your favourite sports team's shift with its patterns and/or colours, for satellites this could be through different frequencies and type of signals. For communication between different Earth stations (say between stations), satellites act as a relay: while you might not be able to send a signal directly to another station because the Earth could be in the way, if there is a satellite that has a line of sight including your station and another station, then one could send a signal to the satellite for it to be passed to the other station.

If you have a network of communication satellites and stations you can use them for navigation purposes. If you have a satellite orbiting the Earth, and within its line of sight it can see multiple Earth stations, then by knowing the differences in travel time of waves between stations and satellite, one can infer for the distances accordingly (since we know these waves should be travelling at the speed of light), and from those distances **triangulate** for the satellite position. In two space dimensions, think of triangulation as drawing two dots denoting the station on a piece of paper, and drawing some circles with those dots as the centre of the circle. The travel time of the waves tells you the radius of the circles you should draw, and the satellite is going to where the two circles cross. In most cases you are going to get two crossing points, but one of them you can most likely rule one of them. In three space dimensions, you are going to be drawing spheres instead, but you can do the same problem with at least three stations (or more)[17]. The same principles apply to locating for example a boat using a combination of satellites and/or ground stations, and is essentially how the **Global Positioning System** (GPS) and other satnav systems work: it is in a nutshell an exercise in geometry, and you (or your computer) find the solution that is consistent with all the data provided by the satellites.

While this might sound simple in principle, remember the wave signals have to travel through all of the Earth's atmosphere, all the while being scattered and/or modified by the medium it is travelling through, and the signal being caught needs to be of a high enough quality to be of any use. For me, when framed that way, makes we wonder why it should even work at all, but the amazing things is that is actually does work[18].

[17] It turns for accuracy reasons you want at least four satellites, because you need to account for relativistic effects (i.e. to constraint the four coordinates of spacetime), as well as to re-calibrate the 'times' on the satellites at least daily. The theory of relativity tells you the clocks on Earth are *slower* (the satellite is not travelling that fast, and gravity is stronger on Earth, so spacetime curvature effects dominate the time dilation effects), otherwise you can get a drift of the positioning on the order of 10 km per day.

[18] I actually feel the same way about aeroplanes, even as a fluid dynamicist.

Altimetry (TOPEX/Poseidon + Jasons)

Satellite altimetry refers to the process of using satellites to get land and/or ocean elevation, and here we are particularly interested in getting the shape of the ocean surface, i.e. the SSH. If we have the SSH, we can use that to infer for the hydrostatic pressure and thus the geostrophic flow (Ch. 7.5.1), tidal signals, monitor sea level rise and fall (Ch. 7.5.2), observing basin-wide signals such as El-Niño Southern Oscillation (ENSO), large-scale waves (Rossby and Kelvin waves), and so on. The underlying principles are analogous to sonar and some of the navigation satellites mentioned above: you send a ping, measure how long it takes to receive the reflected signal, and knowing the speed (which is the speed of light since these invariably use electromagnetic waves), you can infer for the distance (again, with the same caveats above about satellites). Again, the principles sound easy, but in practice there are many details required to get good enough signals for the measurements to be of scientific use.

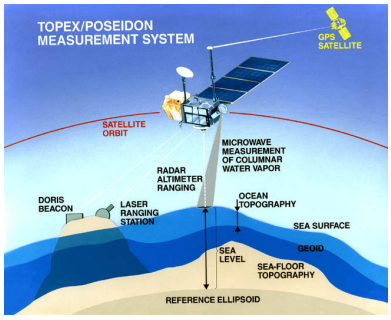


Figure 7.24: Brief schematic of TOPEX/Poseidon.

Besides the short-lived *Seasat* satellite, TOPEX/Poseidon[19] is the first really long-term satellite altimeter, running for around 13 years from 1992 to around 2006 before an instrument failure. The satellite is a joint venture between US and French space agencies, NASA and CNES, and the double naming is apparently because they can't agree on the name the satellite should have, so they had both. TOPEX/Poseidon operates about 1000 km above the Earth's surface, with accurate geolocation (otherwise the measurements do not correspond to the location) and repeating its observation track every 10 days or so. The altimeter uses radar to map out the SSH and land topography, with an accuracy of around a few centimeters[20].

TOPEX/Poseidon was succeeded by the Jason satellite series running on similar orbits and altitudes, and is again a joint venture between NASA and CNES; we are currently on the third one of these (Jason-3). This time they can agree on the name[21], and additional refers to the Greek mythology of Jason and the Argonauts (Jason the satellite looks down from the sky, while Argo the floats look by being in the ocean). While Jason-1 and Jason-2 were planned to run for three years, both of them ended running for over ten years, and we are currently five years into Jason-3. The Jasons are now part included as part of the Sentinel missions.

[19] TOPEX = ocean TOPgraphy EXperiment, and Poseidon = Premier Observatoire Spatial Étude Intensive Dynamique Ocean et Nivosphere, or Positioning Ocean Solid Earth Ice Dynamics Orbiting Navigator. The last French word is spelled wrong (?) on purpose for the acronym (presumably it refers to *lithosphere*, which is land).

[20] That is kind of insane if you think about it, because that is roughly an accuracy of 1 in 10^8 in physical measurement, so it's almost like trying to shoot the eye of a fly at the bottom of the Himalayas with a gun (assuming it can travel that far) from the top of the Himalayas (mm to 10km, which itself is a difference in size of 10^7).

[21] Based on the French JASO1 meetings (Journées Altiméériques Satellitaires pour l'Océanographie) and Joint Altimetry Satellite Oceanography Network.

GRACE

While satellite altimeters measure SSH relative to the ellipsoid, from a dynamical point of view we are also interested in knowing the sea level relative to the geoid (Fig. 3.3 and 3.4). To get the geoid we need very accurate measurements of the gravitational field (or at least the

anomalies), and this is where the **Gravity Recovery and Climate Experiment** (GRACE) satellites come in. By measuring the gravity field over long periods, we can also make observations on how the shifts in mass is occurring within the ocean, land and cryosphere (the ice), which can be used to infer for bottom ocean pressure, rise in sea level and ice melts (mass changes in ocean and ice), and moving tectonics.

GRACE is a joint mission between NASA and the German Aerospace Center DLR, and consists of two essentially identical satellites with a schematic given in Fig. 7.25. If I have to choose a favourite among satellites I would probably pick GRACE, because the way it works to me is one of those that make me think "*oh that's really smart, I wish I thought of that*". If you have a satellite orbiting around the Earth, as it encounters regions with anomalous gravitation field, it will get accelerated or decelerated because of the change in the gravitational attraction (since there is a change in the force, and $F = ma$ from Ch. 1.4.1). While it might be difficult to continuously pinpoint to work out the speed of a moving satellite several hundreds of kilometers above the Earth's surface, you could get around that by having two satellites, one chasing the other some distance behind[22], and you work out the changes in relative velocities between the two satellites by Doppler shifts (Fig. 7.26 for a schematic). The continuous change in received frequencies are recorded by both the satellites, which can then be cross-calibrated. One can link that to a change in velocity, from which you can get an acceleration, and from which you can infer for the change in the gravitational field strength at some particular location determined by the satellite position, made possible by accurate geolocation.

The principles sound easy but of course there are many things that make this hard to do in practice. Changes in gravitational attraction is already very small, so controlling the noise is crucial, otherwise they can swamp the results. A big problem is being able to disentangle the contributions to the acceleration because of gravity from other sources (e.g. drag, solar radiation pressure, etc.). Since gravitational attraction decays like $1/r^2$, a low orbit is ideal for a sensing point of view, but then one has to be aware of other contributions to satellite acceleration, so an altitude of about 500 km above the Earth's surface was chosen. To accurately measure the changes in relative velocity via Doppler shifts, K-band microwaves at 24 and 32 GHz (GHz $= 10^9$ Hz) with some minor shifts between the two satellites are used, which then gives four signals to calibrate against (two satellites each with two channels). Accurate geolocation is made possible by combined surface based geolocation as well as GPS. Note that GRACE is one of the few satellites that do remote sensing of the

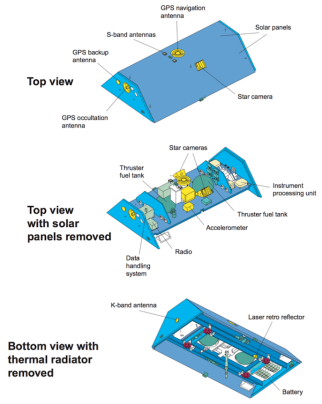


Figure 7.25: GRACE schematics. Modified image originally taken from NASA.

[22] Nicknamed "Tom" and "Jerry", after the classic American cartoon.

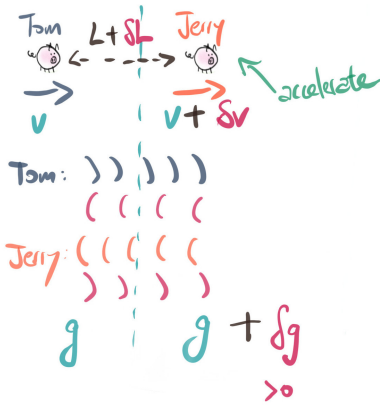


Figure 7.26: Rough schematic of idea behind GRACE represented by the neighbourhood friendly pig and their twin. As one satellite goes into a region of higher gravitational field, it accelerates and pulls away from the other, resulting in a red shift, which is then linked used to infer for the gravitational field anomaly resulting in δg.

Earth without directly remote sensing the Earth.

QuikScat

Going back to satellites that look at the reflected waves from Earth, the Quick Scatterometer (QuikSCAT) operated by NASA observes the ocean wind speed and direction by seeing how waves are scattered by the ocean surface, which is an important set of data for oceanography given the atmospheric winds are one of the primary drivers for ocean dynamics. Low powered microwaves are emitted by the instrument, and when it hits the ocean surface it gets scattered depending on the roughness of the ocean surface. In this instance the roughness depends on the characteristics of waves on the ocean surface, which at the smaller length-scales concerned are largely related to the wind forcing, and thus the scattered signal gives an indication of the wind above the ocean surface. Fig. 7.27 shows a processed output from QuikSCAT of Hurricane Katrina in 2005. QuikSCAT operated from 1999 for about 10 years equipment failures, after which it was operated in a reduced mode, until it was fully retired in 2018. Its operations have been since covered by instruments that work on similar principles (e.g. MetOP-A + B, Oceansat-2, HaiYang-2A etc.).

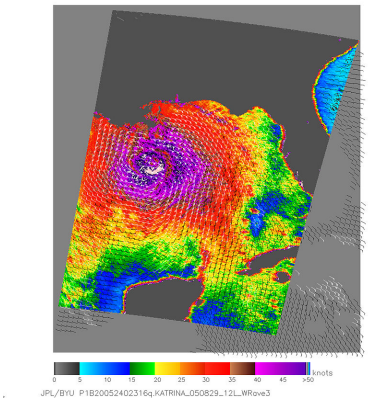


Figure 7.27: QuikSCAT image of Hurricane Katrina in 2005. Image from NASA / JPL.

7·4 *Some observational programs*

As mentioned, since ocean observation is expensive, and exacerbated by the spatial-temporal coverage desired, larger-scale programs are usually joint international efforts. Here are a few of the programs I know about, touching on the equipment used but focusing mostly on the science behind the programs.

RAPID

In place since 2004, the RAPID program aims to monitor the characteristics of the MOC via the AMOC (Ch. 5.2), by observing currents going in and out of the 26.5°N transect in the Atlantic. This is principally achieved by multiple moorings along this transect, appropriately spaced out, with more moorings over the western boundary current region; see Fig. 7.28. The northward and southward branch of the AMOC are monitored by observing the density distributions. With the measurements of density, it is then possible to infer for the transports via using thermal wind shear relation (Ch. 5.1.3).

The moorings of RAPID are regularly maintained (every 12-18 months), and RAPID provided a long term direct observation of the AMOC since 2004 and was supposed to be funded until 2020.

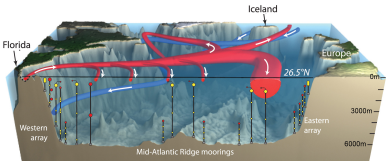


Figure 7.28: Schematic of the RAPID array at 26.5°N. From `rapid.ac.uk`

Fig. 7.29 shows an old time series of the data up to around 2008 of the various components of the AMOC transport, showing a strong surface flow northwards and a return flow at depth. While satellite altimetry can get some of the data, the moorings give the details about the flows at depth. It is from the RAPID data that there was some direct evidence to support a slowdown of the AMOC in recent years (more notable in the signal from 2009 to 2018; not shown here).

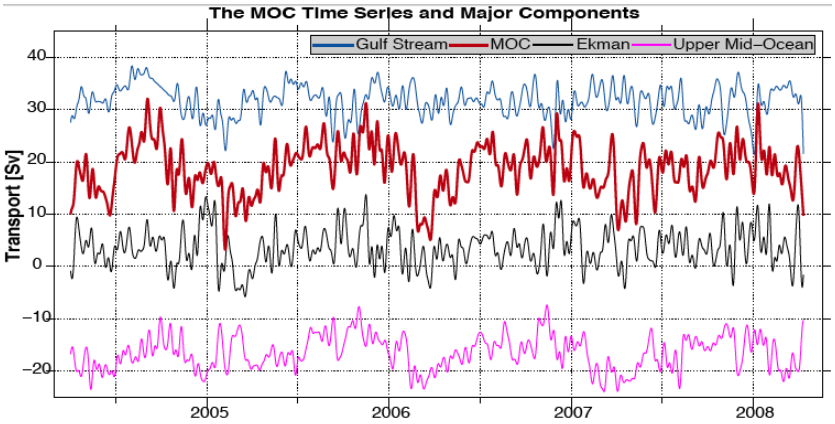


Figure 7.29: Time series from the observed data of RAPID. From rapid.ac.uk

OSNAP

Continuing with the AMOC, the OSNAP program (Overturning in the Subpolar North Atlantic) aims to understand the circulation characteristics in and out of the North Atlantic subpolar gyre at different depths over inter-annual time-scales, with some focus on quantifying the watermass transformation that contributes to the lower branch of the AMOC. Operating since 2014, the observation system consists of moorings (the OSNAP East and West lines), sub-surface floats, gliders and complemented by cruises. The overall OSNAP program appears to be funded until 2024.

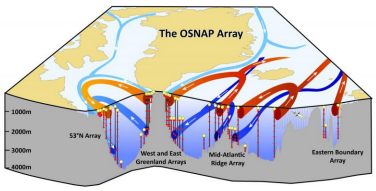


Figure 7.30: Schematic of the O-SNAP array in the North Atlantic subpolar gyre. From www.ukosnap.org.

GO-SHIP

The GO-SHIP[23] (Global Ocean Ship-based Hydrographic Investigations Program) is an international program to co-ordinate the collection and management of data from ship-based observations, to contribute to our overall understanding of the global ocean from a physical, chemical, biological and ecological point of view. A panel was established in 2007 to coordinate this global effort, which includes for example formalising some defined sections that should be repeatedly observed; see Fig. 7.31. The collected and collated data

[23] Inspired by the card game "Go Fish" maybe?

is intended to be publicly accessible, and has some set of submission criteria with regards to the data being submitted, and the time-scales being submitted.

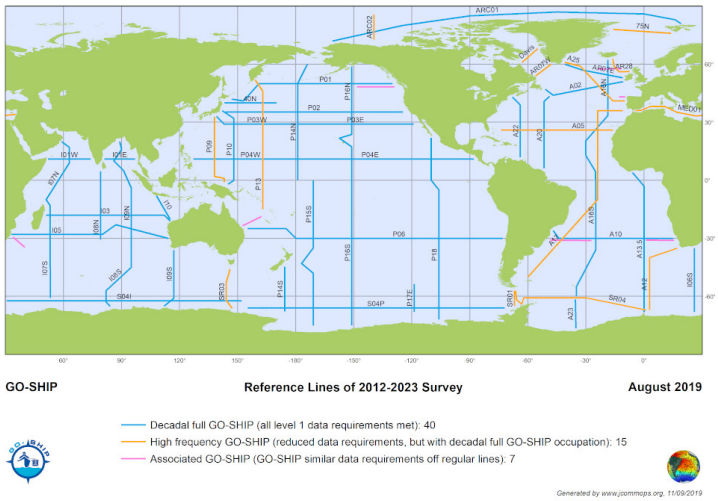


Figure 7.31: Reference sections as of 2019. From `www.go-ship.org`

Argo

Like GO-SHIP, the Argo program (no acronym here) is another international joint effort in observing the ocean, where the principal focus here are collecting and managing ocean data from in-situ floats. The naming of Argo was to complement the Jason satellites, after Jason and the Argonauts from Greek mythology[24]. Functioning since the early 2000s, to qualify being an Argo float the data collected has to be made publicly available, with some set criteria on the ocean properties being collected, and the operation cycle they run on. Fig. 7.32 shows the 10 day operation cycle of the standard Argo floats:

[24] Accordingly to one of the people responsible for the inception of the Argo program, this naming may or may not have happened over an unspecified amount of beers.

- Starting from the surface, the float location is recorded (by satellite communication), and descends to 1000 m drifting depth via changes in buoyancy. Measurements are *occasionally* taken during this step.
- The float drifts for about 10 days at the drifting level.
- The float sinks more (to 2000 m for the standard Argo floats), settles, and then raises to the surface. *Measurements are taken at this step*.
- Once reaching the surface, the location of the float is recorded and data is sent back remotely. The cycle is then repeated.

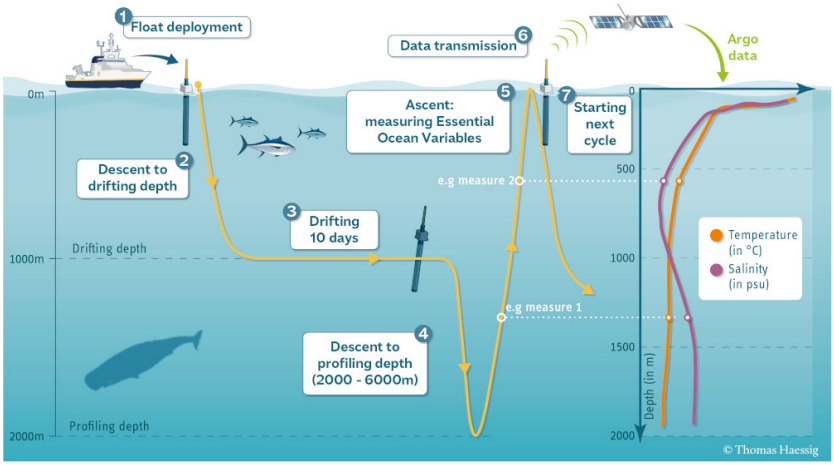


Figure 7.32: Argo float cycle schematic. From `argo.ucsd.edu`.

Each float is estimated to have a functioning period of about five years, and the plan is to have about 4000 of these continuously operating at any one time. Fig. 7.33 shows the current locations of the Argo floats that are feeding back data around the world. There are other varieties of Argo floats that are increasingly being used for operation (deep Argo and Argo-BGC).

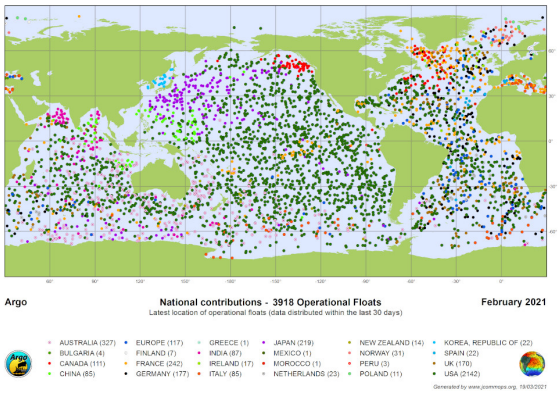


Figure 7.33: Argo locations as of Feb 2021. Note the dots are enhanced in size, so coverage is not as dense as it seems. From `argo.ucsd.edu`.

MEOP

At some point ecologists were interested in knowing where the tracks of high latitude marine mammals to study their living and hunting patterns, so they *tagged* these animals with a geolocating device, such as that in Fig. 7.34; these tags are either taken off the animals after a while, or they fall off when the animal moults in the case of seals, or grows out from the dorsals in the case of whales. So someone then had the bright idea that, if you are going to tag

Figure 7.34: Seal with tag. From Fabien Rouqet (Gothenburg University).

them anyway, why not put on extra sensors (e.g. CTD) on those tags, then you get extra observations as a bonus, particularly given in-situ observations in high latitudes are not very easy to do. The combination of several related national programs became MEOP (Marine Mammals Exploring the Oceans Pole to Pole), which was started around 2008.

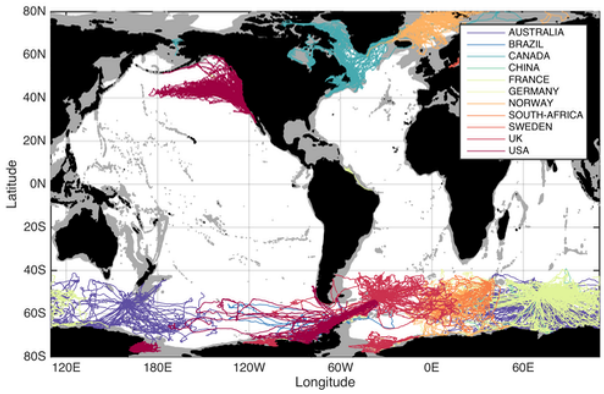


Figure 7.35: Map of sea tracks. From MEOP website www.meop.net.

While there is very little control on where the marine mammals go, one of the unique features particularly with seals is that they make repeated sections into regions under ice, which is previously inaccessible because instruments would generally not be able to come out of the ice covered region to send the data back. The MEOP observations gave unprecedented view of conditions under ice, advancing our understanding of the ocean conditions and its interaction with ice. A set of tracks falling under the MEOP program is given in Fig. 7.35. There are some animal welfare concerns and ethics relating to this program[25], highlighting another possible consideration we might need to make when carrying out observations.

[25] Although research seems to show the tagging seems to have very little biological impact; McMahon et al 2008, Journal of Experimental Marine Biology and Ecology

Copernicus

The Copernicus Programme[26], is managed by the European Commission as part of the European Union, with working partners all over the world. The programme aims to provide "*accurate, timely and easily accessible information to improve the management of the environment, understand and mitigate the effects of climate change and ensure civil security*" (cited from the European Space Agency website).

[26] Copernicus was the one who advocated for a helicocentric universe, when back then the accepted view is that the Earth is the center of universe, and was credited as the start of a process that lead to Newton formulating his theory of gravity a century after his work. His theory was probably formed independently of the Green astronomer Aristarchus of Samos who did something similar 1800 years earlier.

The Sentinel program with Copernicus consists of multiple missions, each mission consisting of multiple satellites that are dedicated to a specific subset of objectives. The missions that are currently approved as of today (April 25, 2021) are given in Table 7.1:

Sentinel 7-12 are in the pipeline. Fig. 7.36 shows one such output

mission	purpose	some sensors
1	for land and ocean services emergency response (e.g. floods)	
2	high res optical imaging for land	multi-spectral imaging
3	ocean + land monitoring	temperature radiometer ocean + land colour radar altimeter
4	atmospheric composition (not launched yet)	hyperspectral spectrometer infra-red sounder
5P 5	atmospheric composition (air pollution)	UV-VIS-NIR spectrometers
6	satellite altimetry for land and ocean services	radar altimeter microwave radiometer

Table 7.1: List of currently approved Sentinel missions.

from the Sentinel-3 mission, which is the concentration of Chlorophyll-*a* inferred from optical images (very roughly, you measure 'greenness' and use that as a proxy for Chlorophyll-*a*).

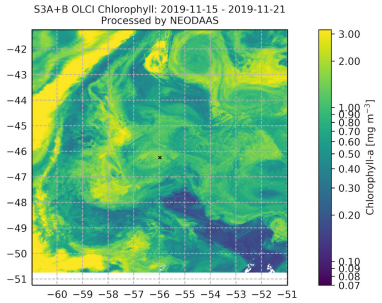


Figure 7.36: Inferred Chlorophyll-*a* from Sentinel 3A and 3B OLCI data. Image from UK NERC NEODAAS.

7.5 *Inference from observations*

In this last part, I try and sketch out how some of theory in the previous chapters and the content in this chapter can play together. This is only meant as an illustrative sketch, and again I have been biased with the examples I chose.

7.5.1 *Geostrophic flow revisited*

By geostrophic balance (Ch. 3.2.3), if we have the distribution of pressure, we can use that to infer for the geostrophic flow $\boldsymbol{u}_g$. By hydrostatic balance (Ch. 3.1.2), we can approximate pressure from knowing the SSH, which we can get from satellite altimetry. An example of this is already given in Fig. 3.8, which shows the mean dynamical topography (the time-mean global SSH). Given we don't really have such a wide spatial coverage of the ocean currents by in-situ observation, this is the method that is generally used to obtain the geostrophic flow. The method is entirely analogous to how things were (and still is to a certain extent) done in synoptic meteorology, where if we know the pressure patterns, we have a very fairly good approximation to the atmospheric winds on sufficiently large-scales.

There are a few complication to bear in mind. One is that we might want the sea surface height anomalies relative to the geoid (which is the dynamically relevant quantity) to compute the pres-

sures. Satellite altimetry gives the anomalies relative to the ellipsoid, so we may want to isolate these accordingly (you certainly need this for sea level change applications). The other is that there are multiple frequencies inherent in SSH data (e.g. tides), so we may need to isolate and filter the signals accordingly. The method should also be limited to the large-scale flows and away from the equator, since we formally need to remain within the small Rossby number regime to apply geostrophic balance.

In principle we can go a bit further with the inference. If we know the temperature and salinity over the globe, via in-situ observations for example, we can compute (potential or neutral) density gradients, then we could extend the observed surface geostrophic flow to below the ocean surface via thermal wind shear relation (Ch. 5.1.3). This for example is what might be done with mooring data taken from RAPID and/or OSNAP to infer for the flow in and out of the section. The problem with that approach is that since thermal wind shear relation Eq. 5.3 appears as an equation with derivatives, in the process of getting $\boldsymbol{u}_g$ there will be some integration constants that show up. If those constants are not chosen appropriately (constrained by observations or otherwise), then it is entirely plausible that the resulting flows led to violation of conservation laws such as conservation of salt or mass, since those constraints were not used in deriving the thermal wind shear relation. Another is that there might not be a unique choice for these constants[27]. The interested reader is referred to Ch. 9.1 of Wunsch [2015] for a related discussion.

[27] A differential equation need not be *well-posed*, i.e. have a unique solution, or even a solution.

Once we have the geostrophic flow there are several things we could do. One example is to obtain the **eddy kinetic energy** associated with the geostrophic flow, given by

$$K = \frac{1}{2}\rho\, \overline{\boldsymbol{u}_g' \boldsymbol{u}_g'}, \qquad \boldsymbol{u}_g = \overline{\boldsymbol{u}_g} + \boldsymbol{u}_g', \tag{7.3}$$

where $\overline{(\cdot)}$ denotes an average, a prime denotes the deviation from that average, such that (1) the average of the average doesn't do anything, (2) the average of the deviation is zero, and (3) the average is a linear operator, i.e.

$$\overline{\overline{A}} = \overline{A}, \qquad \overline{A'} = 0, \qquad \overline{a(A+B)} = a\left(\overline{A} + \overline{B}\right),$$

for some number a and functions A and B. Eddy kinetic energy (usually defined relative to a time average) provides a measure of the mesoscale eddy activity, and Fig. 7.37 shows the (depth-integrated) eddy kinetic energy obtained the geostrophic flow inferred from satellite altimetry data, showing for example large eddy activity in the western boundary currents (Ch. 5) and in the Southern Ocean (Ch. 6), as expected. The high latitudes and equator have been

blocked out because of ice (so no/limited SSH signals there) and the formal breakdown of geostrophic balance needed to infer for the geostrophic flow. These kind of data are useful for example as a target for what numerical ocean models should try and reproduce, or to understand how the dynamical features might be evolving over time.

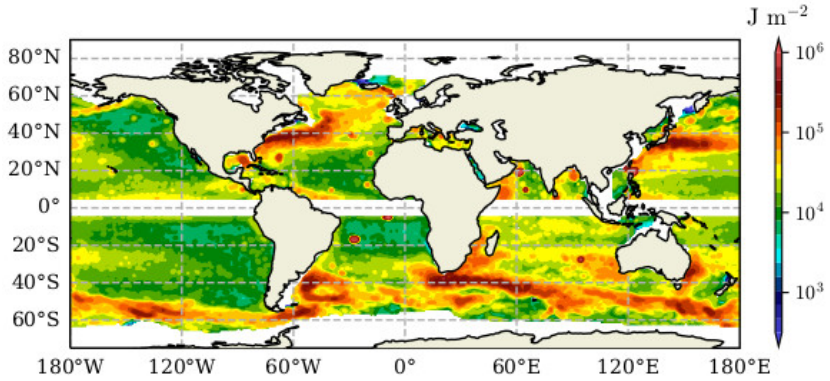


Figure 7.37: Depth-integrated eddy kinetic energy of geostrophic flow from satellite altimetry, extended down in depth using the first baroclinic mode. Data product provided by Xiaoming Zhai (University of East Anglia).

7.5.2 *Reanalyses and data assimilation*

In the previous example we have highlighted a possibility that, by knowing the surface data, we can extend it to depths in principle. Given the general sparse coverage in space and in time relating to observations, we may also want to 'fill out' the data in space and in time, so how do we might we go about that?

One thing we could envisage doing is to just interpolate the data: if we have some temperature and salinity distribution over two ship tracks, we might be tempted to say the data is just some value in between the two tracks where we have data. This is almost certainly going to be problematic, because that kind of procedure will generically not give an overall state that satisfies the governing equations, such as Eq. 1.3. When we are dealing with physical oceanography we have the added benefit in that we know what our governing equations are, and what kind of conservations we should have, so we should try and use that information as a *constraint* when inferring for the ocean state given some observational data. Thus, the general problem we want to solve is as something like:

Given the (fundamentally incomplete and noisy) information from ocean observations, we seek the ocean state (e.g. T, S, $\boldsymbol{u}_3$, SSH etc.) such that it exactly matches the observed data at the relevant locations, and satisfies the

equations of motions and constraints.

This class of problem is generally known as an **inverse problem**. Normally, you input something in some machine and it gives you an output (the *forward problem*, to make *predictions*), but here you have the output and the machine, and you seek the input instead (the *inverse problem*, to make *inferences*). However, one thing highlighted in the previous section is well-posedness of a problem, and in general inverse problems are not well-posed as we have insufficient information to make the ocean state we seek match exactly to the given observational data. Instead, we change the problem slightly to the following:

Given the (fundamentally incomplete and noisy) information from ocean observations, we seek the ocean state (e.g. T, S, $\boldsymbol{u}_3$, SSH etc.) such that it is close enough to the observed data at the relevant locations, and satisfies the equations of motions and constraints.

The demand for approximate matching over exact matching makes the problem easier to do in practice. While formally there is still no guarantee the problem is well-posed and has solutions, in practice such an approach tends to lead to *a* solution, though of course there is no guarantee of uniqueness. In this case 'something' is arguably better than 'nothing', and the validity of the approach is examined by the output it gives.

Usually this kind of procedure or problem is carried out in conjunction with a numerical ocean model, which is a numerical implementation of the equations governing ocean dynamics. The act of forcing the ocean model to incorporate incoming observational data is known as **data assimilation**. The 'filled out' ocean state with (past) constrained upon the observational data is generally known as **reanalyses**, or sometimes **state estimates**. Without going into too many details, there are multiple approaches, choices and assumptions one has to make when generating a reanalyses, such as:

- what data to use?
- what numerical model to use?
- which approach to take for assimilating observation data?
- what variables do we want reanalyses for?
- ...

Each different combination is generically going to result in different reanalyses products, and one should just be aware that there are strengths and weaknesses with different choices, which can

impact the choice of product to download from the data repositories depending on what question you are interested in.

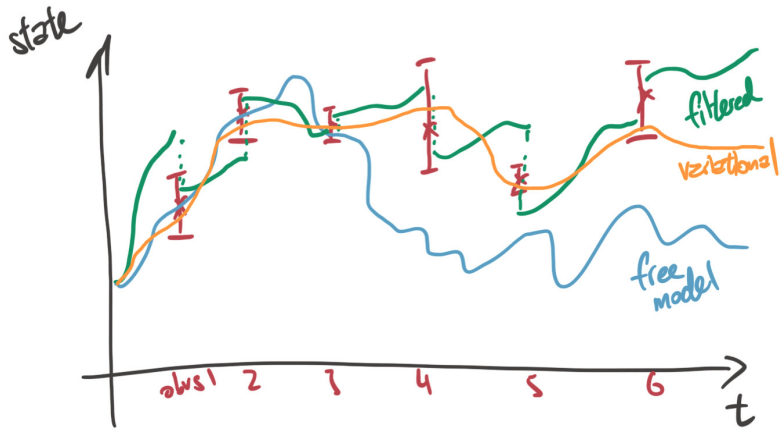


Figure 7.38: Schematic of solution trajectories with observations marked on (in red), with no data assimilation (blue), assimilation of data by filtering approach (green), and by a time-continuous variational approach (orange).

To illustrate the above point, Fig. 7.38 shows a schematic difference between products resulting from the way data is assimilated, assuming all cases start at the same point. The free running model with no data assimilation 'matches' (or really, is 'close enough' to) the observations for a little while, but afterwards it drifts and does whatever it wants with no regard for the observations; the free-running model is allowed and generically will do this. If we instead take a *filtering* approach, then what we do is we let the model run for a bit, but when observations come in, we utilise the information from both the model and observations to 'kick' the modelled state back towards the observed state, and then start again. If we take a particular *variational* approach where we additionally demand the *whole* whole trajectory to be continuous in time (cf. 4D VAR), then we try and force the trajectory to be 'close' to the observational data. The filtering approach is the standard approach when making predictions (e.g. in weather forecasting), but has the inherent problem that because the solution is being kicked away at every analysis step when we assimilate data, a shock is introduced, and extra things have been added to the reanalyses product that are inconsistent with how the state should have evolved if left to its own devices. On the other hand, the variational approach is in practice incredibly expensive to do partly because of the demand for time-continuous trajectories, but does not formally have problems to do with shocks. The product resulting from the latter variational apporach would then be much more appropriate if we are interested in analysing budgets (e.g. how salt and heat might have moved in and out of the basins over long times) than the product from the filtering approach (though it might

not be too bad for short times).

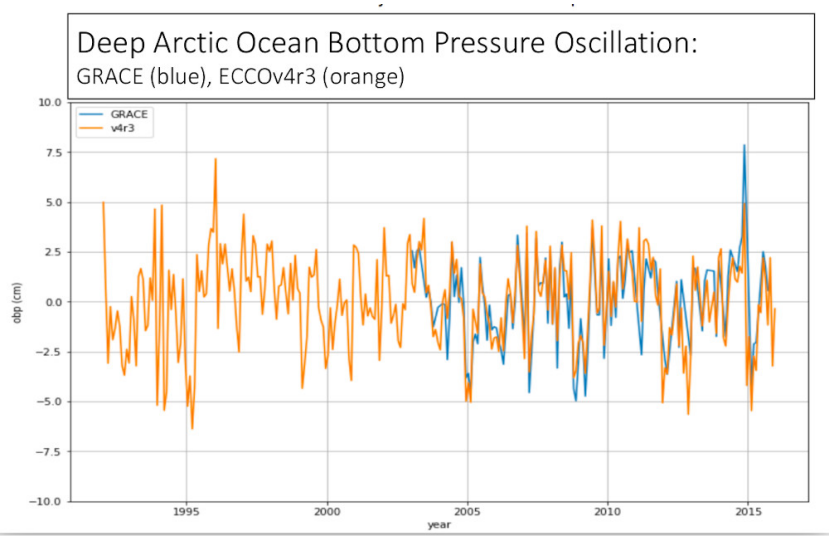


Figure 7.39: Bottom pressure in the Arctic Ocean as measured by GRACE (blue) and from the ECCO v4r3 state estimate. Figure from Ian Fenty (NASA JPL).

Usually gridded data that one can obtain from the data repositories have gone through something like the above process to fill in the gaps (with additional processing as required), so just bear in mind that depending on the scientific application, different products might want to be used. Fig. 7.39 shows a plot of the (appropriately averaged) bottom pressure in the Arctic Ocean, from both the GRACE satellite itself, and from the ECCO (Estimating the Circulation and Climate of the Ocean; e.g. Forget et al. [2015]) product, which does use the aforementioned 4D VAR approach, assimilating for example satellite data, mooring data, floats, sea ice, marine mammal tracks, among many others. There any many other examples where ECCO has been used (see `www.ecco-group.org/publications.htm` for a massive list). Other products using filtering methods exist (e.g. SODA, Carton et al. [2000]).

Another example from reanalyses products is shown in Fig. 7.40, which is the evolution of global mean sea level over time, with data from Church & White (2011), which contains tide gauge data as well as satellite data (from 1993). The trends[28] are plotted on and demonstrate the rise of global mean sea level. There are regional differences in sea level change that could be analysed accordingly from other data products.

[28] From linear least squares fitting, which happens to be one of the textbook examples of an inverse type problem that is well-posed and has a unique closed-form solution.

Summary and further reading

The aim here was a sketch of how some of the instruments observe the oceans, and I have chosen ones where the underlying principles

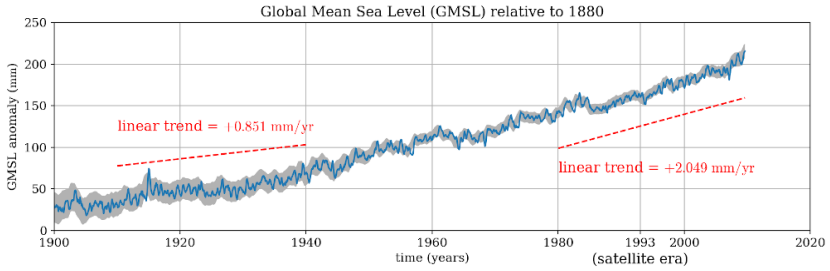


Figure 7.40: Global mean sea level over a hundred years or so. Data from Church & White (2011). See `tobermory_tides.ipynb` and `historical_sea_level_plot.ipynb`.

are by themselves not mega complex (in my opinion) and are interesting to myself. I have invariably made the observation aspect sound much *easier* than it actually is: it is in the execution of basic ideas in a highly controlled and intended way that I personally find awe inspiring, and really is a fine example of scientific and technological progress, together with human ingenuity.

There are many notable things I have not talked about here, and this was deliberate. These include:

- *Where to acquire data*. This was deliberate because it depends on what you need. The first good places to start would be to just search in Google. Some specific places include WOCE (ocean climatologies, `http://woceatlas.ucsd.edu/`), NASA PO.DAAC (ocean data, `podaac.jpl.nasa.gov`), NOAA NCEP (atmosphere and ocean, `psl.noaa.gov/data/gridded/data.ncep.reanalysis.html`), NERC NEODAAS (Earth observation, `www.neodaas.ac.uk/Home`).

- *How to analyse data*. This is outside of the scope here, and again depends on the specific problem you have. The codes to generate some of the more quantitative diagrams are available as Python Jupyter notebooks in the GitHub repository hosting this document. See also Appendix A of Wunsch [2015] for example.

- *Uncertainty quantification and error analysis*. Absolutely crucial, but outside of the scope here. For example, how much accuracy do we really need to get acceptable geostrophic flow speeds? If we have an error of something in the GRACE frequency measurements, how much error in the gravitational field strength anomaly does that translate to? If we use observational data with known uncertainties to get reanalyses products, what is the corresponding uncertainty in the product? What if we have unknown unknowns (as opposed to known unknowns in the previous part)?

- *Applications*. There are too many, and I made the lazy choice to put in topics that I already have code and/or data for. There

is for example the all important problem in quantifying ocean heat content evolution (e.g. Cheng et al. [2019], Zanna et al. [2019]), AMOC circulations (e.g. Pillar et al. [2018], Smith and Heimbach [2019], Kostov et al. [2019]), biogeochemistry in data assimilation (e.g. Verdy and Mazloff [2017]), data assimilation and observational system designs (e.g. Fujii et al. [2019]), and numerous others.

- *Best practices in running observation campaigns*. I don't do observations personally so I have nothing to add here...my understanding is that there are generally 'better' or 'accepted' ways of doing observations, but there probably isn't a 'right' way as such.

- *Best practices in designing observation campaigns*. As above, and have a catchy acronym? Joking aside, there are some interesting work recently using data assimilation infrastructure to inform how one might want to design future observations (e.g. Fujii et al. [2019], where best to put observations and what kind of observations to maxmise information gain, say).

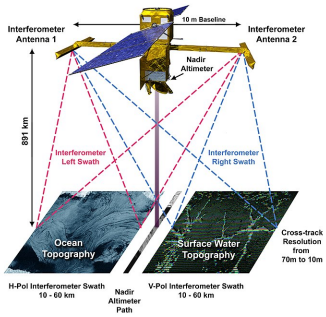


Figure 7.41: Schematic of SWOT. Image from NASA JPL.

Like theories, observational techniques are constantly evolving with the emerging theories, technologies and scientific questions. There is the upcoming SWOT (Surface Water and Ocean Topography) that will perform satellite altimetry over ocean as well as rivers, lakes and floodplains and unprecedented spatial scales (15 to 25 km), and questions exist on how best to use the data, what kind of new science could be done, how to manage to massive amount of data coming in, and so forth. Some older ideas have been revisited, such as *acoustic tomography* (e.g. Munk and Wunsch [1979]) to observe the ocean temperature[29]. Previously killed for its perceived impact on marine animals[30], acoustic tomography is making a reappearance (e.g. Wu et al. [2020], using sounds from earthquakes). Given 2020 is the UN decade of the ocean, it will be interesting to see how much of the content in this document becomes outdated.

Going back to right at the beginning of Ch. 1, physical oceanography in particular is to do with *what* the ocean looks like, and *why* does it look like the way it does. The two questions and the methodologies employed are not independent of each other, and it's not that one is more important and/or 'better' than the other: they are just different, and they are both important. The observations constrain the theory, and the theory rationalises the observations. Throughout this set of notes I have focused mostly on theory, but I hope I have highlighted in this last part on why observations are important, and how the theory can play together with observational data. The reader is referred to the excellent book of Wunsch [2015]

[29] Roughly, sound travel time depends on density, which in the ocean is largely governed by the temperature. By monitoring changes in the sound travel times and their propagation characteristics we can potentially infer for how the temperature profiles in the ocean changes over time.

[30] Research have found that noise levels from acoustic tomography has no biologically significant effects, and is small compared to noise from ship traffic.

for a much more thorough and careful narrative than the one I have attempted here.

Chapter exercises

1. We talked a bit about the ambulance and Doppler shift problem in Ch. 7.3.1. Supposing you as the observer is holding a glowing light bulb, and the ambulance driver is driver towards you (just assume it is driving straight in to you). Using the formula in that section, estimate how fast the ambulance driver has to be travelling relative to you the observer to see a noticeable increase in the light bulb *colour*. Fig. 2.5 may or may not be of use.

2. basic error analysis

3. some of Alex's internal wave stuff

Appendix: Some useful (?) maths

Vector calculus

Complex numbers

Linera algebra

Exercises

Bibliography

R. Abernathey, D. Ferreira, and A. Klocker. Diagnostics of isopycnal mixing in a circumpolar channel. *Ocean Modell.*, 72:1–16, 2013. DOI: 10.1016/j.ocemod.2013.07.004.

Y. Aksenov, E. P. Popova, A. Yool, A. J. G. Nurser, T. D. Williams, L. Bertino, and J. Bergh. On the future navigability of Arctic sea routes: High-resolution projections of the Arctic Ocean and sea ice. *Mar. Policy*, 75:300–317, 2017. DOI: 10.1016/j.marpol.2015.12.027.

H. S. Ashbaugh, T. M. Truskett, and P. G. Debenedetti. A simple molecular thermodynamic theory ofhydrophobic hydration. *J. Chem. Phys.*, 116(7):2907–2921, 2002. DOI: 10.1063/1.1436479116.

F. P. Bretherton. Baroclinic instability and the short wavelength cut-off in terms of potential vorticity. *Q. J. Roy. Met. Soc.*, 92:335–345, 1966.

H. L. Bryden. New Polynomials for thermal expansion, adiabatic temperature gradient and potential temperature of sea water. *Deep-Sea Res.*, 20:401–408, 1973. DOI: 10.1016/0011-7471(73)90063-6.

J. R. Carpenter, E. W. Tedford, E. Heifetz, and G. A. Lawrence. Instability in stratified shear flow: Review of a physical interpretation based on interacting waves. *Appl. Mech. Rev.*, 64: 060801, 2011.

J. A. Carton, G. Chepurin, X. Cao, and B. S. Giese. A Simple Ocean Data Assimilation analysis of the global upper ocean 1950–95. Part I: Methodology. *J. Phys. Oceanogr.*, 30:294–309, 2000. DOI: 10.1175/1520-0485(2000)030%3C0294:ASODAA%3E2.0.CO;2.

J. G. Charney. Dynamics of long waves in a baroclinic westerly current. *J. Meteor.*, 4:135–162, 1947.

J. G. Charney and P. G. Drazin. Geostrophic turbulence. *J. Geophys. Res.*, 66:83–109, 1961.

L. Cheng, J. Abraham, Z. Hausfather, and K. E. Trenberth. How fast are the oceans warming? *Sci. Adv.*, 363:128–129, 2019. DOI: 10.1126/science.aav7619.

C. de Lavergne, G. Madec, J. Le Sommer, A. J. G. Nurser, and A. C. Naveira Garabato. On the consumption of Antarctic Bottom Water in the abyssal ocean. *J. Phys. Oceanogr.*, 46:635–661, 2016a. DOI: 10.1175/JPO-D-14-0201.1.

C. de Lavergne, G. Madec, J. Le Sommer, A. J. G. Nurser, and A. C. Naveira Garabato. The impact of variable mixing efficiency on the abyssal overturning. *J. Phys. Oceanogr.*, 46:663–681, 2016b. DOI: 10.1175/JPO-D-14-0259.1.

C. de Lavergne, G. Madec, F. Roquet, R. M. Holmes, and T. J. McDougall. Abyssal ocean overturning shaped by seafloor distribution. *Nature*, 551:181–186, 2017. DOI: 10.1038/nature24472.

P. G. Drazin and W. H. Reid. *Hydrodynamic Stability*. Cambridge University Press, 2nd edition, 1981.

E. T. Eady. Long waves and cyclone waves. *Tellus*, 1:33–52, 1949. DOI: 10.1111/j.2153-3490.1949.tb01265.x.

R. Ferrari and C. Wunsch. The distribution of eddy kinetic and potential energies in the global ocean. *Tellus*, 62A:92–108, 2010.

R. Ferrari, M. F. Jansen, J. F. Adkins, A. Burke, A. L. Stewart, and A. F. Thompson. Antarctic sea ice control on ocean circulation in present and glacial climates. *Proc. Natl Acad. Sci. USA*, 111(24): 8753–8758, 2014. DOI: 10.1073/pnas.1323922111.

R. Ferrari, A. Mashayek, T. J. McDougall, M. Nikurashin, and J.-M. Campin. Turning ocean mixing upside down. *J. Phys. Oceanogr.*, 46: 2239–2261, 2016. DOI: 10.1175/JPO-D-15-0244.1.

P. Fofonoff and R. C. Millard Jr. Algorithms for computation of fundamental properties of seawater. Technical Report 53, UNESCO Tech. Pap. in Mar. Sci., 1983.

G. Forget, J.-M. Campin, P. Heimbach, C. N. Hill, R. M. Ponte, and C. Wunsch. ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation. *Geosci. Model Dev.*, 8:3071–3104, 2015. DOI: 10.5194/gmd-8-3071-2015.

Y. Fujii, E. Rémy, H. Zuo, P. Oker, G. Halliwell, F. Gasparin, M. Benkiran, N. Loose, J. Cummings, J. Xie, Y. Xue, S. Masuda, G. C. Smith, M. Balmaseda, C. Germineaud, D. J. Lea, G. Larnicol, L. Bertino, A. Bonaduce, P. Brasseur, C. Donlon, P. Heimbach,

Y. Kim, V. Kourafalou, P.-Y. Le Traon, M. Martin, S. Paturi, B. Tranchant, and N. Usui. Observing system evaluation based on ocean data assimilation and prediction systems: on-going challenges and a future vision for designing and supporting ocean observational networks. *Front. Mar. Sci.*, 6:417, 2019. DOI: 10.3389/fmars.2019.00417.

J. Gan, Z. Liu, and C. R. Hui. A three-layer alternating spinning circulation in the South China Sea. *J. Phys. Oceanogr.*, 46:2309–2315, 2016. DOI: 10.1175/JPO-D-16-0044.1.

P. R. Gent, J. Willebrand, T. J. McDougall, and J. C. McWilliams. Parameterizing eddy-induced tracer transports in ocean circulation models. *J. Phys. Oceanogr.*, 25:463–474, 1995. DOI: 10.1175/1520-0485(1995)025<0463:PEITTI>2.0.CO;2.

R. H. Grove. Global Impact of the 1789-93 El Niño. *Nature*, 393:318–9, 1998. DOI: 10.1038/30636.

N. Harnik, E. Heifetz, O. M. Umurhan, and F. Lott. A buoyancy-vorticity wave interaction approach to stratified shear flow. *J. Atmos. Sci.*, 65:2615–2630, 2008. DOI: 10.1175/2007JAS2610.1.

E. Heifetz, J. Mak, J. Nycander, and O. M. Umurhan. Interacting vorticity waves as an instability mechanism for magnetohydrodynamic shear instabilities. *J. Fluid Mech.*, 767: 199–225, 2015. DOI: 10.1017/jfm.2015.47.

B. J. Hoskins, M. E. McIntyre, and A. W. Robertson. On the use and significance of isentropic potential vorticity maps. *Q. J. Roy. Met. Soc.*, 111:877–946, 1985.

IOC, SCOR, and IAPSO. The international thermodynamic equation of seawater – 2010: Calculation and use of thermodynamic properties. Technical report, Intergovernmental Oceanographic Commission, Manuals and Guides No. 56, UNESCO, 2010.

M. F. Jansen. Glacial ocean circulation and stratification explained by reduced atmospheric temperature. *Proc. Natl Acad. Sci. USA*, 114(1): 45–50, 2017. DOI: 10.1073/pnas.1610438113.

C. P. Kelley, S. Mohtadi, M. A. Cane, R. Seager, and Y. Kushnir. Climate change in the Fertile Crescent and implications of the recent Syrian drought. *Proc. Natl. Acad. Sci. U.S.A.*, 112:3241–3246, 2015. DOI: 10.1073/pnas.1421533112.

S. Kobayashi, Y. Ota, Y. Harada, A. Ebita, M. Moriya, H. Onoda, K. Onogi, H. Kamahori, C. Kobayashi, H. Endo, K. Miyaoka, and

K. Takahashi. The JRA-55 reanalysis: General specifications and basic characteristics. *J. Met. Soc. Japan. Ser. II*, 93(1):5–48, 2015. DOI: 10.2151/jmsj.2015-001.

Y. Kostov, H. L. Johnson, and D. P. Marshall. AMOC sensitivity to surface buoyancy fluxes: the role of air-sea feedback mechanisms. *Climate Dyn.*, 53:4521–4537, 2019. DOI: 10.1007/s00382-019-04802-4.

I. Langmuir. Surface motion of water induced by wind. *Science*, 87: 119–123, 1938. DOI: 10.1126/science.87.2250.119.

M. Lévy, P. Klein, A.-M. Tréguier, D. Iovino, G. Madec, S. Masson, and K. Takahashi. Modifications of gyre circulation by sub-mesoscale physics. *Ocean Modell.*, 34:1–15, 2010. DOI: 10.1016/j.ocemod.2010.04.001.

G. Madec. NEMO ocean engine. *Note du Pôle de modélisation, Institut Pierre-Simon Laplace (IPSL), No. 27*, 2008.

J. Mak, S. D. Griffiths, and D. W. Hughes. Shear flow instabilities in shallow-water magnetohydrodynamics. *J. Fluid Mech.*, 788:767–796, 2016. DOI: 10.1017/jfm.2015.718.

J. Mak, S. D. Griffiths, and D. W. Hughes. Vortex disruption by magnetohydrodynamic feedback. *Phys. Ref. Fluids*, 2:113701, 2017. DOI: 10.1103/PhysRevFluids.2.113701.

J. Mak, J. R. Maddison, D. P. Marshall, and D. R. Munday. Implementation of a geometrically informed and energetically constrained mesoscale eddy parameterization in an ocean circulation model. *J. Phys. Oceanogr.*, 48:2363–2382, 2018. DOI: 10.1175/JPO-D-18-0017.1.

D. P. Marshall, D. R. Munday, L. C. Allsion, R. J. Hay, and H. L. Johnson. Gill's model of the Antarctic Circumpolar Current, revisited: The role of latitudinal variations in wind stress. *Ocean Modell.*, 97:37–51, 2016. DOI: 10.1016/j.ocemod.2015.11.010.

A. Mashayek and W. R. Peltier. The 'zoo' of secondary instabilities precursory to stratified shear flow transition. Part 1 Shear aligned convection, pairing, and braid instabilities. *J. Fluid Mech.*, 708:5–44, 2012a. DOI: 10.1017/jfm.2012.304.

A. Mashayek and W. R. Peltier. The 'zoo' of secondary instabilities precursory to stratified shear flow transition. Part 2 The influence of stratification. *J. Fluid Mech.*, 708:45–70, 2012b. DOI: 10.1017/jfm.2012.294.

T. J. McDougall. Potential enthalpy: A conservative oceanic variable for evaluating heat content and heat fluxes. *J. Phys. Oceanogr.*, 33:945–963, 2003. DOI: 10.1175/1520-0485(2003)033<0945:PEACOV>2.0.CO;2.

G. Meneghello, J. Marshall, J.-M. Campin, E. Doddridge, and M.-L. Timmermans. The ice-ocean governor: ice-ocean stress feedback limits Beaufort gyre spin-up. *Geophys. Res. Lett.*, 45:11293–11299, 2018. DOI: 10.1029/2018GL080171.

D. R. Munday, H. L. Johnson, and D. P. Marshall. Eddy saturation of equilibrated circumpolar currents. *J. Phys. Oceanogr.*, 43:507–532, 2013. DOI: 10.1175/JPO-D-12-095.1.

D. R. Munday, H. L. Johnson, and D. P. Marshall. The role of ocean gateways in the dynamics and sensitivity to wind stress of the early Antarctic Circumpolar Current. *Paleoceanography*, 30:284–302, 2015. DOI: 10.1002/2014PA002675.

W. H. Munk. Abyssal recipes. *Deep-Sea Res. Oceanogr. Abstr.*, 13(4): 707–730, 1966. DOI: 10.1016/0011-7471(66)90602-4.

W. H. Munk and C. Wunsch. Ocean acoustic tomography: A scheme for large scale monitoring. *Deep-Sea Res.*, 26:123–161, 1979. DOI: 10.1016/0198-0149(79)90073-6.

W. H. Munk and C. Wunsch. Abyssal recipes II: energetics of tidal and wind mixing. *Deep-Sea Res., Part 1, Oceanogr. Res. Pap.*, 45(12): 1977–2010, 1998. DOI: 10.1016/S0967-0637(98)00070-3.

J. Nycander. Horizontal convection with a non-linear equation of state: generalization of a theorem of Paparella and Young. *Tellus A*, 62:134–137, 2010. DOI: 10.1111/j.1600-0870.2009.00429.x.

W. M. F. Orr. The stability or instability of the steady motions of a liquid. part i: A perfect liquid. part ii: A viscous liquid. *Proc. R. Irish Acad. A*, 27:9–138, 1907.

F. Paparella and W. R. Young. Horizontal convection is non-turbulent. *J. Fluid Mech.*, 466:205–214, 2002. DOI: 10.1017/S0022112002001313.

N. A. Phillips. The general circulation of the atmosphere: a numerical experiment. *Q. J. Roy. Met. Soc.*, 82:123–164, 1956.

G. L. Pickard and W. J. Emery. *Descriptive Physical Oceanography: An Introduction*. Pergamon Press, 1990. DOI: 10.1016/C2009-0-11176-5.

H. R. Pillar, H. L. Johnson, D. P. Marshall, P. Heimbach, and S. Takao. Impacts of atmsopheric reanalysis uncertainty on Atlantic

overturning estimates at 25°N. *J. Climate*, 31:8719–8744, 2018. DOI: 10.1175/JCLI-D-18-0241.1.

A. Rabinovich, O. M. Umurhan, N. Harnik, F. Lott, and E. Heifetz. Vorticity inversion and action-at-a-distance instability in stably stratified shear flow. *J. Fluid Mech.*, 670:301–325, 2011. DOI: 10.1017/S002211201000529X.

L. F. Richardson and H. Stommel. Note on eddy diffusion in the sea. *J. Meteorology*, 5:238–240, 1948. DOI: 10.1175/1520-0469(1948)005<0238:NOEDIT>2.0.CO;2.

F. Roquet, G. Madec, L. Brodeau, and J. Nycander. Defining a simplified yet "realistic" equation of state for seawater. *J. Phys. Oceanogr.*, 45:2564–2579, 2015a. DOI: 10.1175/JPO-D-15-0080.1.

F. Roquet, G. Madec, T. J. McDougall, and P. M. Barker. Accurate polynomial expressions for the density and specific volume of seawater using the TEOS-10 standard. *Ocean Modell.*, 90:29–43, 2015b. DOI: 10.1016/j.ocemod.2015.04.002.

B. Schutz. *A First Course in General Relativity*. Cambridge University Press, 2nd edition, 2009.

T. A. Smith and P. Heimbach. Atmospheric origins of variability in the South Atlantic Meridional Overturning Criculation. *J. Climate*, 32:1483–1500, 2019. DOI: 10.1175/JCLI-D-18-0311.1.

G. J. Stanley. Neutral surface topology. *Ocean Modell.*, 138:88–106, 2019. DOI: 10.1016/j.ocemod.2019.01.008.

L. D. Talley, G. L. Pickard, W. J. Emery, and J. H. Swift. *Descriptive Physical Oceanography*. Academic Press, 6th edition, 2011.

G. I. Taylor. Eddy motion in the atmosphere. *Proc. Lond. Math. Soc.*, 20:196–212, 1921.

J. R. Toggweiler and B. Samuels. Effect of Drake passage on the global thermohaline circulation. *Deep-Sea Res., Part 1, Oceanogr. Res. Pap.*, 42(4):477–500, 1995. DOI: 10.1016/0967-0637(95)00012-U.

J. R. Toggweiler, J. L. Russel, and S. R. Carson. Midlatitude westerlies, atmospheric CO_2, and climate change during the ice ages. *Paleoceanography*, 21:PA2005, 2006. DOI: 10.1029/2005PA001154.

G. K. Vallis. *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press, 2006.

A. Verdy and M. Mazloff. A data assimilating model for estimating Southern Ocean biogeochemistry. *J. Geophys. Res. Oceans.*, 122: 6968–6988, 2017. DOI: 10.1002/2016JC012650.

R. G. Williams and M. J. Follows. *Ocean Dynamics and the Carbon Cycle: Principles and Mechanisms*. Cambridge University Press, 2011.

W. Wu, Z. Zhan, S. Pend, S. Ni, and J. Callies. Seismic ocean thermometry. *Science*, 369:1510–1515, 2020. DOI: 10.1126/science.abb9519.

C. Wunsch. *Modern Observational Physical Oceanography: Understanding the Global Ocean*. Princeton University Press, 2015.

L. Zanna, S. Khatiwala, J. M. Gregory, J. Ison, and P. Heimbach. Global reconstruction of historical ocean heat storage and transport. *Proc. Natl. Acad. Sci. USA*, 116:1126–1131, 2019. DOI: 10.1073/pnas.1808838115.