CLIMATE MATHEMATICS

CLIMATE MATHEMATICS

–Commonly Used Mathematics and Statistics Tools for Climate Science

Samuel S.P. Shen
San Diego State University and University of California San Diego

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

FOREWORD

This volume of Climate Mathematics is built on the class notes for a course of the same title (Course Catalog Number SIOC 290S) at Scripps Institution of Oceanography, University of California San Diego. SIOC 290 was first taught in a special summer session of five weeks in 2015, with nine instruction hours per week.

The course was designed for the students in the newly established Masters of Advanced Studies program in Climate Sciences and Policy. These students would need to understand, explain and present the results from climate models and observations. The students would use SIOC 290 to prepare themselves to take SIO 210 (Physical Oceanography) and SIO 217A (Atmospheric Thermodynamics). The students would learn probability, statistics, mathematics, and plotting skills to present and describe climate data from both observations and models, particularly about climate extremes and uncertainties.

The students may have different mathematics background, ranging from three semesters of calculus, linear algebra, differential equations, complex variables and basic statistics, to only Calculus I and basic statistics. Even the students who have taken many mathematics courses before may find that the mathematics and statistics in climate science are used in a different format. Thus, SIOC 290 will adopt an innovative instruction of mathematics and statistics to present the minimum amount of mathematics needed to make effective climate data interpretation and presentation. This climate math class will start from the basic concept of calculus (without limits), and go to Taylor series and line integral, from the basic concept of probability and statistics to advanced theory of sampling error estimation, and from R programming to more advanced plotting and visualization skills.

PhD students who need additional mathematics and statistics skills or who would like to learn modern tools and unconventional approaches to climate data analysis and models can also take this course.

GLOSSARY

DD Calculus	Descartes' direct calculus
D[f,a]	Derivative of a function f at its indecent variable equal to a, which is the same as $f'(a)$
I[f,a,b]	Integral of a function f from a to b, which is the same as $\in_a^b f(x)dx$

PART I

CALCULUS, LINEAR ALGEBRA, CLIMATE DATA REPRESENTATIONS, CLIMATE MODELS:
–DIFFERENTIAL EQUATIONS AS CALCULUS APPLICATIONS
–CLIMATE DATA MATRICES AS LINEAR ALGEBRA APPLICATIONS
–CLIMATE MODELS AS DIFFERENTIAL EQUATIONS APPLICATIONS

CHAPTER 1

DD CALCULUS –LEARNING BASIC CALCULUS CONCEPTS IN TWO HOURS

This chapter introduces DD calculus and describes the basic calculus concepts of derivative and integral in a direct and non-traditional way, without limit definition: Derivative is computed from the point-slope equation of a tangent line and integral is defined as the height increment of a curve. This direct approach to calculus has three distinct features: (i) it defines derivative and (definite) integral without using limits, (ii) it defines derivative and antiderivative simultaneously via a derivative-antiderivative (DA) pair, and (iii) it posits the fundamental theorem of calculus as a natural corollary of the definitions of derivative and integral. The first D in DD calculus attributes to Descartes for his method of tangents and the second D to DA-pair. The DD calculus, or simply direct calculus, makes many traditional notations and procedures unnecessary, a plus when introducing calculus to the non-mathematics majors. It has few intermediate procedures, which can help dispel the mystery of calculus as perceived by the general public. The materials in this paper are intended for use in a two-hour introductory lecture on calculus.

—Summary

1.1 Introduction: Necessity to dispel calculus mystery and simplify calculus notations

"Calculus" is not a commonly used word in daily life. The Oxford Dictionary indicates that the word comes from the mid-17th century Latin and literally means small pebble (such as those used on an abacus) for counting. The dictionary gives three meanings of "calculus": a branch of mathematics that deals with derivatives and integrals, a particular method of

Climate Mathematics,
First Edition.

calculation, and a hard mass formed by minerals. Obviously here we are interested in the first meaning. We wish to demonstrate that (i) the calculus method can be developed by analyzing steepness and height change of a curve, (ii) the method development can be achieved directly using Descartes' method of tangents and does not need an introduction of limit as prerequisite, and (iii) the basic method of calculus and a few simple examples can be introduced in a one-hour or two-hour lecture to a high-school level audience.

Calculus is one of the most important tools in a knowledge based society. Millions of people around the world learn calculus everyday. All engineering, science, and business major undergraduate students must take calculus. Many high schools offer calculus courses. The usefulness and power of calculus have been well recognized. Nonetheless, calculus is a mysterious subject to many people and is regarded by the general public as accessible only to a few privileged people with special talents. Tight schedules and high fail rates for the first semester calculus have given the course a reputation as a monster, a nightmare, or a phycological barrier for many students, some of whom are even STEM (science, technology, engineering and mathematics) majors. Calculus can be a topic that causes people at a social gathering to shake their heads in incomprehension, shy away from the daunting challenge of understanding it, or express effusive exclamation of awe and admiration. It is also sometimes associated with conspicuous nerdiness. In classrooms, the student-instructor relationship can be tense. Some students regard calculus instructors as inhuman and ruthless aliens, while instructors frequently joke about students' stupidity, clumsiness, or silly errors. Tedious and peculiar notations coupled with fiendish and complex approaches to calculus teaching and learning may have contributed to the above unfortunate situation.

A major cause of this mystery and scare of calculus is unnecessarily complex terminologies, including definite and indefinite integrals, derivatives defined as limits, definite integrals defined as limits, the difference between dx and Δx, and using ever-finer divisions of an area under a curve to approximate a definite integral (i.e., introducing the definite integral by area and limit), to name but a few. Additionally, the Riemann sum with its arbitrary point x_i^* in the interval $[x_i, x_{i+1}]$ complicates calculation procedures and adds more confusion. These conventional concepts and notations are essential for professional mathematicians who research mathematical analysis, but are absolutely unnecessary to the majority of calculus learners, who are majoring in engineering, science, business, or other non-mathematical or non-statistical fields.

In the first semester calculus, instructors repeatedly emphasize that an indefinite integral yields a function while a definite integral yields a real value. When an instructor discovers that some students still cannot tell the difference between a definite integral and an indefinite integral in the final exam, s/he becomes disappointed and complains that these terrible students did not pay attention to her/his repeated emphasis on the difference, but s/he rarely questions the necessity of introducing the concept of indefinite integrals and their notations.

The purpose of this paper is to dispel this mystery of calculus by introducing a more direct approach. We attempt to introduce the basic ideas of calculus in one hour without using the concept of limits. Our introduction to derivatives directly uses the idea of René Descartes' (1596-1650) method of tangents (Cajori,1985, pp.176-177; Coolidge, 1951; Susuki, 2005; Range, 2011), rather than indirectly uses the secant line method with a limit. We introduce derivatives and antiderivatives simultaneously, using derivative-antiderivative (DA) pairs. Our introduction to integrals is directly from the DA pairs and the height increment of an antiderivative curve. The height increment approach has been advocated by Q. Lin in China for over two decades (see Lin (2010) and Lin (2009) for two

recent examples). We define the area under the curve of an integrand by the integral, and then explain why the definition is reasonable. This is a reversal of the traditional definition, which defines an integral by calculating the area underneath the curve of an integrand.

We also demonstrate that an introduction to the basic ideas of calculus does not need to use many complicated notations. Thus, the notations of derivatives $f'(x)$ and integrals $I[f(x), a, b]$ in this paper are parsimonious, simple, computer friendly and non-traditional.

In the following sections, we will first introduce (a) the method of point-slope equation for calculating derivatives, and (b) DA pairs for calculating both antiderivatives and derivatives. Then we will introduce integrals as the height increment of an antiderivative curve. Next we will discuss the mathematical rigor of our direct calculus method, but this part of material does not need to be included in a one-hour lecture. Finally we provide a brief history on the development of calculus ideas and give our conclusions on the direct calculus from the perspective of Descartes' method of tangents and DA pairs.

1.2 Slope, and derivative and antiderivative pairs

When we drive on a steep highway, we often see a grade warning sign like the one in Figure 1. The 9% grade or slope on a highway means that the elevation will decrease 90 feet when the horizontal distance increases 1000 ft. The grade or slope is calculated by

$$m = \tan\theta = \frac{H}{L} = \frac{90}{1000} \tag{1.1}$$

i.e., the ratio of the opposite side to the adjacent side of the right triangle in Figure 1.

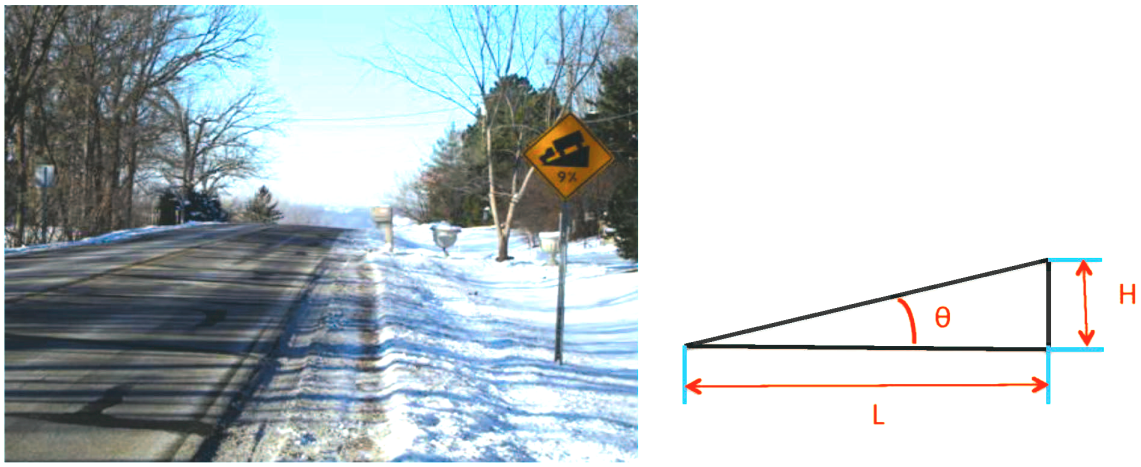


Figure 1.1 Highway grade sign and a right triangle to show slope.

The slope of a curve at a given point is defined as the slope of a tangent line at this point. Figure 2 shows three points: P, A and B. T_P represents the tangent line at P whose slope is defined as the slope of the curve at P. The tangent line's slope is used to measure the curve's steepness. Calculus studies (i) the slope $\tan\theta$ of a curve at various points, and (ii)

the height increment H from one point to another, say, A to B, as shown in Figure 2. Our geometric intuition indicates that height H and slope $\tan\theta$ are related because H increases rapidly if the slope is large for an upward trend. A core formula of calculus is to describe the relationship between these two quantities.

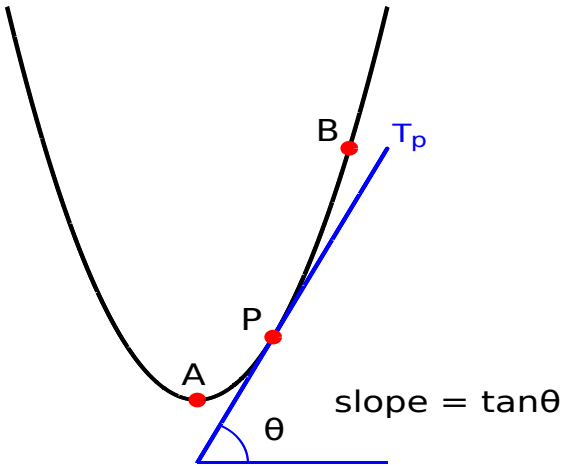


Figure 1.2 A curve, three points and a tangent line without coordinates.

Let us introduce the coordinates x and y and use a function $y = f(x)$ to describe the curve. We start with an example $f(x) = x^2$.

The tangent line of the curve at point $P(x_0, y_0)$ can be described by a point-slope equation

$$y - y_0 = m(x - x_0), \tag{1.2}$$

where $y_0 = x_0^2$, m is the slope to be determined by the condition that the tangent line touches the curve at one point. In fact, "tangent" is a word derived from Latin and means "touch."

The tangent line (1.2) and the curve

$$y = x^2 \tag{1.3}$$

have a common point $P(x_0, y_0)$ (See Figure 3), where x_0 must be a double root, since the straight line is a tangent line.

Substituting (1.3) to (1.2) to eliminate y, we have

$$x^2 - x_0^2 = m(x - x_0), \tag{1.4}$$

then

$$(x - x_0)(x + x_0) - m(x - x_0) = 0, \tag{1.5}$$

or

$$(x - x_0)(x + x_0 - m) = 0. \tag{1.6}$$

The two solutions of this quadratic equation are

$$x_1 - x_0 = 0 \tag{1.7}$$

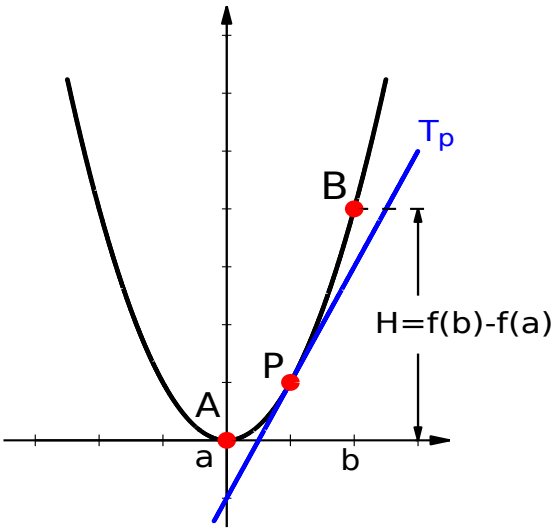


Figure 1.3 A curve, three points and a tangent line on xy-plane.

and

$$x_2 + x_0 - m = 0 \tag{1.8}$$

Since $P(x_0, y_0)$ is a tangent point, x_0 must be a repeated root, also called a double root. Hence, $x_1 = x_2 = x_0$. This yields

$$m = 2x_0. \tag{1.9}$$

Thus, we claim that the slope of the curve $y = x^2$ at x_0 is $2x_0$, at a is $2a$, at 3 is $2 \times 3 = 6$, and in general, at x is $2x$. The slope measures the steepness of the curve $y = f(x)$, i.e., the rate of the curve's height increase or decrease. The slope, or rate, varies from point to point. The slope is thus a derived quantity from the original function $y = f(x)$ and is called "derivative." We have the following definition.

Definition 1. (Definitions of derivative as slope and DA pair). *The slope $2x$ is called the derivative of x^2. Further, x^2 is called an antiderivative of $2x$. And $(2x, x^2)$ is called a derivative-antiderivative (DA) pair.*

For a general function $y = f(x)$, the slope of the curve $y = f(x)$ is called *derivative* and is denoted by $f'(x)$ or y', and $f(x)$ is called antiderivative of $f'(x)$. Thus, $(f'(x), f(x))$ is a DA pair.

If $f(x) = C$ is a constant, then $y = C$ represents a horizontal line whose slope is 0 for any x. Hence $(C)' = 0$, and $(0, C)$ is a DA pair.

If $f(x)$ is a linear function, then $y = \alpha + \beta x$ represents a straight line whose slope is β for any x, hence $(\alpha + \beta x)' = \beta$, i.e., $(\beta, \alpha + \beta x)$ is a DA pair.

Thus, the derivative's geometric meaning is the slope of the curve $y = f(x)$: $f'(x)$ is large at places where the curve $y = f(x)$ is steep. At a flat point, such as the maximum or minimum point of $f(x)$, the slope is zero since the tangent lines at these points are horizontal.

In addition to the geometric meaning, derivative has physical meaning, such as speed, and biological meaning, such as growth rate, as well as the meaning of the rate of change in almost any scientific field and anyone's daily life. As an example, for a car driven at $v = 60$ mph (miles per hour) for two hours, the total distance traveled is $s = v \times t = 60 \times 2 = 120$ mi. Here, (v, vt) or (v, s) is a DA pair for a general time t.

Free fall is another example. An object's free fall has its distance of falling equal to $s = (1/2)gt^2$ and its falling speed is $v = gt$, where $g = 32[ft/s^2]$ is the Earth's gravitational acceleration. Galileo Galilei (1564-1642) discovered this time-square relationship for the distance. Since derivative t^2 with respect to t is $2t$, we have $s' = \frac{1}{2}g(2t) = gt$. Thus, $(gt, (1/2)gt^2)$ or (v, s) is a DA pair.

In general, the meaning of derivative is the rate of change of the function $f(x)$ with respect to the independent variable x, which can be either time or spatial location.

Let us now return to the derivative calculation. The above tangent line approach for finding the slope for x^2 can be applied to the function $y = x^3$. It is to solve the following simultaneous equations

$$y - y_0 = m(x - x_0), \tag{1.10}$$

$$y = x^3. \tag{1.11}$$

Eliminating y, we have

$$x^3 - x_0^3 = m(x - x_0). \tag{1.12}$$

The factorization of this equation yields

$$(x - x_0)(x^2 + xx_0 + x_0^2 - m) = 0. \tag{1.13}$$

This factorization implies $x_1 - x_0 = 0$ and $x_2^2 + x_2x_0 + x_0^2 - m = 0$. The cubic equation has three solutions. Because $P(x_0, y_0)$ is the tangent point, x_0 is a repeated root of these two equations: $x_1 = x_2 = x_0$, which leads to

$$m = 3x_0^2. \tag{1.14}$$

Hence, we claim that the derivative of x^3 is $3x^2$, and x^3 is an antiderivative of $3x^2$. Namely, $(3x^2, x^3)$ forms a DA pair.

Following the tangent line approach, the above examples have demonstrated four DA pairs:

$$(0, C), (1, x), (2x, x^2), (3x^2, x^3). \tag{1.15}$$

The tangent line approach can be applied to any power function x^n, where n is a positive integer. The DA pair for x^n is

$$(nx^{n-1}, x^n). \tag{1.16}$$

This formula actually holds for any real number n with the exception of $n = 0$. For example, $(x^{1/2})' = (1/2)x^{-1/2}$. Proof of this claim is not easy, but fortunately the calculation of derivatives and antiderivatives can be easily done using computer programs. Many open source computer software packages are available to do this kind of calculation. For example, WolframAplha is one. At the website www.wolframalpha.com, you can enter a derivative command and a function, such as

```
derivative  x^(3/2)+4x^3
```

The computer will give you its derivative, plus other information, such as the graph of the derivative function (see Figure 4). You can also use WolframAlpha via a smart phone application.

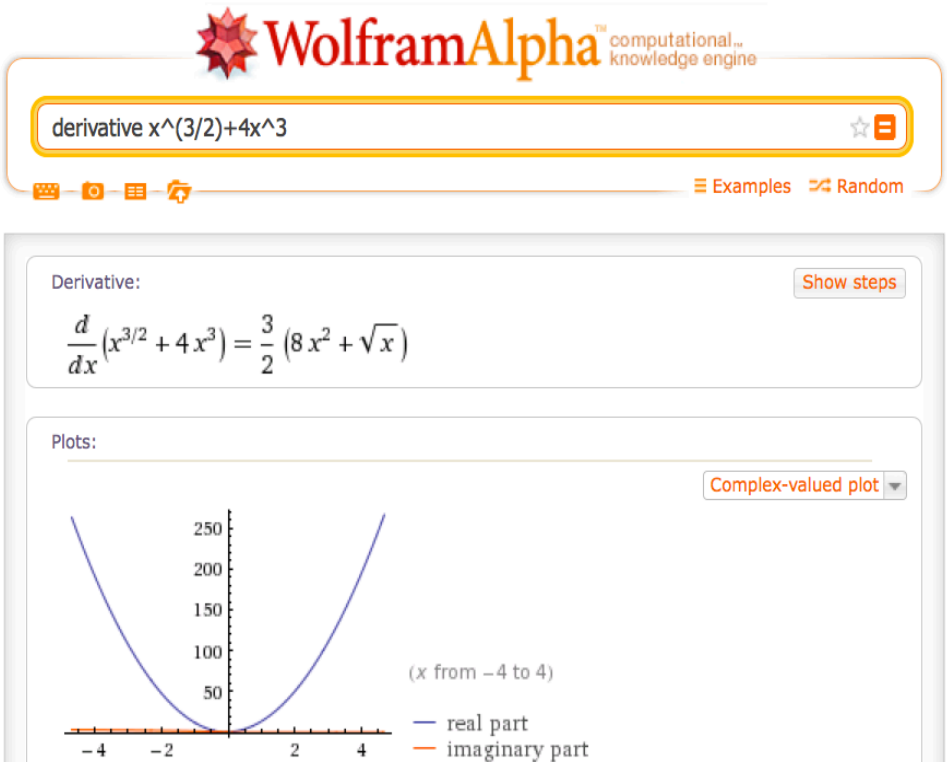


Figure 1.4 WalframAlpha derivative example.

To find an antiderivative, you can use a similar command

```
antiderivative  x^(3/2)+4x^3
```

Using this program, one can easily find DA pairs for commonly used functions. See the list below.

1. Exponential function: (e^x, e^x).
2. Natural logarithmic function: $(\frac{1}{x}, \ln x)$.
3. Sine function: $(\cos x, \sin x)$.
4. Cosine function: $(-\sin x, \cos x)$.
5. Tangent function: $(\sec^2 x, \tan x)$.

1.3 Height increment and integrals

When we trace a curve, we care about not only the slope, but also the ups and downs of the curve, i.e., the increment or decrement of the curve from one point to another. When

we drive over a mountain road, we also care about both steepness (i.e., slope) and elevation. Apparently, the slope and height increment are related. The slope has already been defined as derivative in the above section. In this section, the height increment is defined as *integral*, since the height increment or elevation increment is an integration process, or an accumulation process of point motion, measured by both speed and time.

For a function $y = f(x)$, its increment from $A(a, f(a))$ to $B(b, f(b))$ is $f(b) - f(a)$ as shown in Figure 3. Another notation for the increment is $f(b) - f(a) = f(x)|_a^b$. This height increment is used to describe the integral definition below.

Definition 2. (Definition of integral as height increment of a curve). *The function's increment $f(b) - f(a)$ from $A(a, f(a))$ to $B(b, f(b))$ is defined as the integral of the derivative function $f'(x)$ in the interval $[a, b]$ and is denoted by $I[f'(x), a, b] = f(b) - f(a)$. Here, $f'(x)$ is called the integrand, and $[a, b]$ is called the integration interval.*

Example 1. Given $f(x) = x$, $f'(x) = 1$, and $[a, b] = [0, 2]$, we have

$$I[f'(x), a, b] = I[1, 0, 2] = x|_0^2 = 2 - 0 = 2. \tag{1.17}$$

The area between $y = 1$ and $y = 0$ in the interval $[0, 2]$ is also 2 (see Figure 5 for $f'(x) = 1, a = 0$, and $b = 1$).

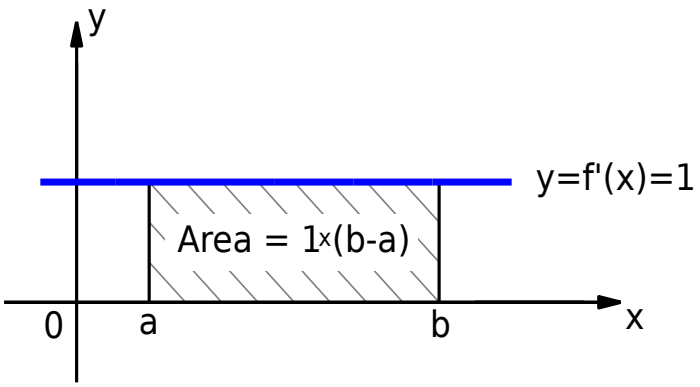


Figure 1.5 The area of a rectangle under a horizontal line.

Example 2. If we integrate speed $v(t)$, we will then get the distance $I[v(t), a, b]$ travelled from time $t = a$ to $t = b$. If $v(t)$ is a constant, say, 60 mi/hour, and if $a = 14 : 00$ and $b = 16 : 00$, then the integral $I[60, 14, 16] = 60t|_{14}^{16} = 60 \times (16 - 14) = 120$ mi is the total distance traveled from 2pm to 4pm. Here $60t$ is an antiderivative of 60. If we plot v as a function of t, then 120 is equal to the area of the rectangle bounded by $v = 60, v = 0, t = 14$ and $t = 16$ (See Figure 5 for $f'(x) = 60$, $a = 14$, and $b = 16$).

Example 3. Given $f(x) = (1/2)x^2$, $f'(x) = x$, and $[a, b] = [0, 1]$, we have

$$I[f'(x), a, b] = I[x, 0, 1] = (1/2)x^2|_0^1 = (1/2)(1^2 - 0^2) = 1/2. \tag{1.18}$$

The area under the integrand $y = x$ but above the x-axis in $[0, 1]$ is $1/2$ (see Figure 6 for $a = 0$ and $b = 1$).

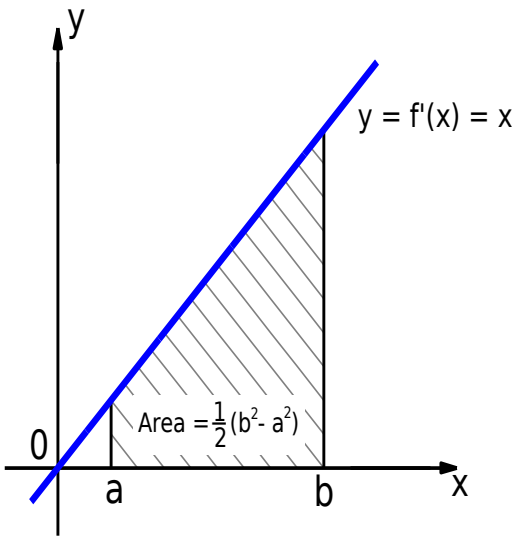


Figure 1.6 The area of a triangle under a straight line.

Example 4. In the free fall problem, the speed is a linear function of time, $v = gt$, and the integration $I[gt, 0, x] = (1/2)gx^2$ is the distance traveled from time zero to time x. The region bounded by $v = gt, v = 0, t = 0$ and $t = x$ is a right triangle with base equal to x, height gx and area $(1/2) \times x \times gx = (1/2)gx^2$. In this example, we have chosen to use x as an arbitrary right bound for the region. This x can can be any number, such as 1, 2, or 2.5.

In the above four examples, the area under a curve is equal to an integral. As a matter of fact, this inference of area equal to an integral is generally true. For an irregular region, we can simply use the integral $I[f'(x), a, b]$ as the definition of the area of the region bounded by $y = f'(x)$, x-axis, $x = a$ and $x = b$. The next section will justify this definition. As for the calculation of an integral, if one knows the relevant DA pair, the integral is a simple substitution $f(b) - f(a)$. Otherwise, one can use a computer to find the antiderivative, or to directly evaluate $I[f'(x), a, b]$. Similar to derivative software, there are many free online computer programs and smart phone apps for calculating integrals. Figure 7 shows the WolframAlpha calculation of the integral $I[x^4 + 2x, 0, 1]$ using the command

```
integrate x^4+2x from 0 to 1
```

The result is 6/5. i.e., $I[x^4 + 2x, 0, 1] = 6/5$.

The definition of an integral states that the integral of a function in an interval is the increment of its antiderivative in the same interval. One can write

$$I[g(t), a, b] = G(b) - G(a) \tag{1.19}$$

where $G(t)$ is an antiderivative of $g(t)$. Another way to express the above is

$$I[G'(u), a, b] = G(b) - G(a). \tag{1.20}$$

In the above two expressions, t and u are the integration variables, also called dummy variables. The integral values are independent of the choice of dummy variables. One

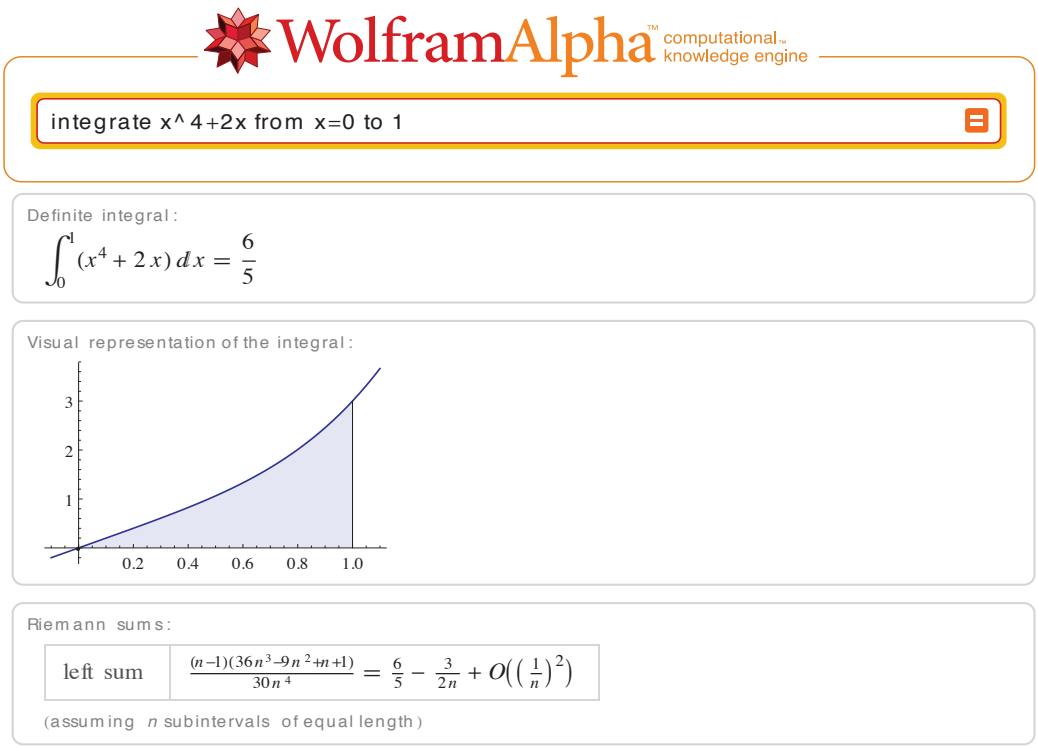


Figure 1.7 WalframAlpha integral example.

can use any symbol to represent this variable. In practical applications, if the independent variable is time, such as when speed is a function of time, t is often used as the independent variable.

Also according to the integral definition, the integral of $f'(t)$ in the interval $[a, x]$ is

$$I[f'(t), a, x] = f(x) - f(a). \tag{1.21}$$

Taking derivative of both sides of this equation with respect to x, we have

$$(I[f'(t), a, x])'_x = f'(x) \tag{1.22}$$

since $(f(a))'_x = 0$ due to $f(a)$ being a constant with respect to x and having a slope of zero. Here we have used a subscript x to indicate that the independent variable is x and the derivative is with respect to x.

Equation (1.22) is often called Part II of the Fundamental Theorem of Calculus (FTC), while the definition of an integral is actually often referred to as Part I of the FTC. Part II of the FTC is saying that an antiderivative can be explicitly expressed by an integral. Thus, the FTC makes a computable and close connection between slope and height increment, and enhances our intuitive sense that the height increment in an interval is closely related to the slope of our interested curve, i.e., our study function.

Example 5. $I[x, 0, 1] = (x^2/2)|_0^1 = 1^2/2 - 0^2/2 = 1/2$, because $(x, x^2/2)$ is a DA pair. The area between $y = x$ and $y = 0$ over $[0, 1]$ is 1/2 (=$I[x, 0, 1]$) (See Figure 6).

Example 6. $I[x^2, 0, 1] = (x^3/3)|_0^1 = 1^3/3 - 0^3/3 = 1/3$, because $(x^2, x^3/3)$ is a DA pair, or simply since x^2's antiderivative is $x^3/3$. The area of the right triangle with a curved hypotenuse bound by $y = x^2, y = 0$ and $x = 1$ is $1/3$ (=$I[x^2, 0, 1]$) (See Figure 8). The WolframAlpha command for this calculation is `integrate x^3 from 0 to 1`.

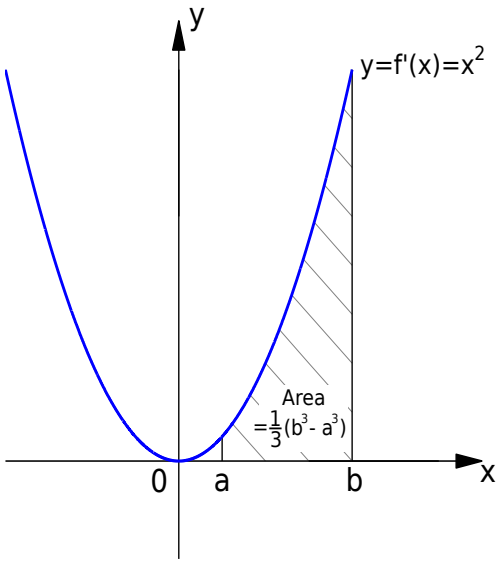


Figure 1.8 The area of a triangle of a curved hypotenuse when $a = 0$.

1.4 Discussion and mathematical rigor of direct calculus

Two points are discussed here. First, is our definition of area by an integral reasonable and mathematically rigorous? Second, besides using computer programs to calculate the DA pairs of complicated functions, can one provide a systematic procedure of hand calculation?

First, how do we know that our definition of area using an integral is reasonable? According to the Oxford dictionary,"area" is defined as "the extent or measurement of a surface or piece of land." The word "area" comes from the mid-16th century Latin, literally meaning a "vacant piece of level ground." The units of an area are "square feet","square meters", etc, meaning that the area of a region is equal to the number of equivalent squares, each side equal to a foot, a meter, or other units, that fit into the region. For the area between $y = f'(x)$ and $y = 0$ over $[a, b]$, the simplest measure is to use an equivalent rectangle of length $L = b - a$ and width W (See Figure 9). That is, the excess area above $y = W$ (the vertically stripped region) is moved to fill the deficit area (the horizontally striped region). The corresponding description by a mathematical formula is below:

$$I[f'(x), a, b] = L \times W = (b - a) \times W. \tag{1.23}$$

This can be true as long as we have

$$W = \frac{I[f'(x), a, b]}{b - a} = \frac{f(b) - f(a)}{b - a}. \tag{1.24}$$

We call this W the mean value of $f'(x)$ over the interval $[a, b]$. Geometrically, $W = (f(b) - f(a))/(b - a)$ is the slope of the secant line that connects points A and B of Figure 10. If $y = f(x)$ is not a straight line, then there must be a point m in $[a, b]$ whose slope is less than $(f(b) - f(a))/(b - a)$, and another point M whose slope is larger than $(f(b) - f(a))/(b - a)$, i.e.,

$$f'(m) \leq \frac{f(b - f(a)}{b - a} \leq f'(M). \tag{1.25}$$

Between $f'(m)$ and $f'(M)$, $(f(b) - f(a))/(b - a)$ must meet the mid-ground slope $f'(c)$ at a point c in $[a, b]$, i.e.,

$$W = f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{1.26}$$

This is the mean value theorem (MVT) of calculus. It states that

Theorem 1. (MVT). *There exists c in $[a, b]$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$ if $f'(x)$ has a value for every x in $[a, b]$.*

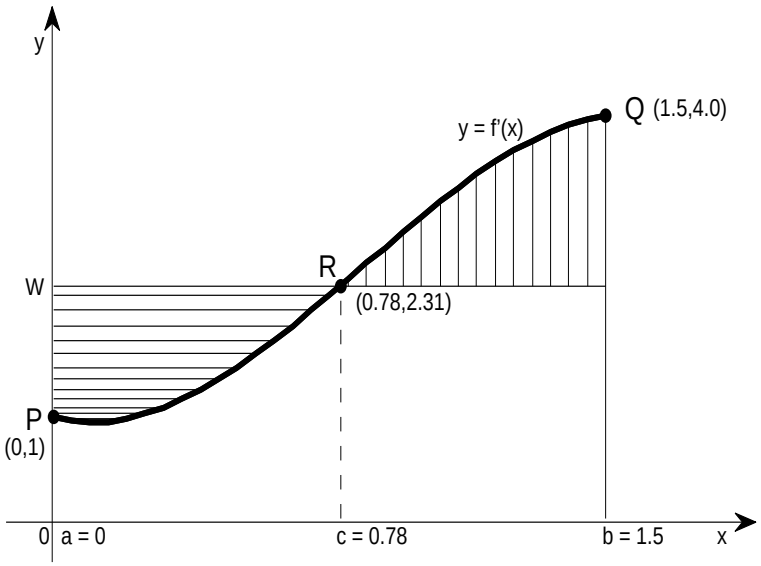


Figure 1.9 An area interpretation of an integral.

Geometrically, MVT means that there is at least one point c whose tangent line is parallel to the secant line AB. Of course, this holds if $y = f(x)$ is a straight line, in which case c can be any point in $[a, b]$.

Rigorous mathematics for MVT would require one to prove the above statement "$(f(b)-f(a))/(b - a)$ must meet the mid-ground slope $f'(c)$ at one point c in $[a, b]$," namely, it proves the existence of the point c. This is to prove the intermediate value theorem and is beyond the scope of this introductory lecture.

Therefore, the integral $I[f'(t), a, x] = f(x)-f(a)$ is the increment of the antiderivative from a to x, and is also the area for the region between the integrand derivative function

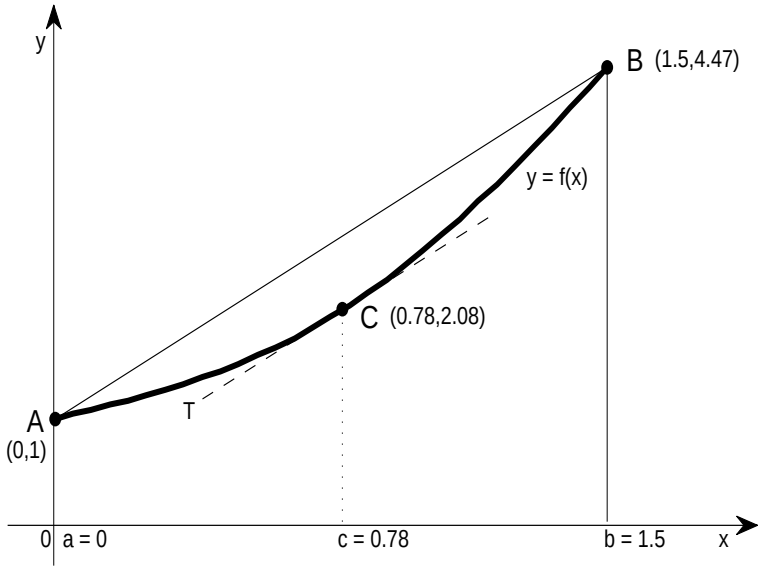


Figure 1.10 Illustration of the mean value theorem: There is a tangent line parallel to the secant line connecting points A and B.

and $y = 0$ in the interval $[a, x]$, i.e., the region bounded by $y = f'(t), y = 0, t = a$ and $t = x$. The traditional definition of an integral is from the aspect of an area that is defined as a sum of many rectangles of increasing narrow widths, under the condition of each width approaching zero. For non-mathematics majors and general public, the condition of each width approaching zero, which is a concept of limit, adds complexity and confusion to the traditional definition of integral. In contrast, the geometric meaning of our direct integral is the height increment of the antiderivative, not the area underneath of the derivative. The area is only regarded as an additional geometric interpretation according to the intermediate value theorem. Under this interpretation of area, we have the following example.

Example 7. $I[\sqrt{1-x^2}, 0, 1]$ is the area of a quarter unit round disc and is thus equal to $\pi/4$, since $y = \sqrt{1-x^2}$ represents a quarter circle in the first quadrant.

Calculating the slope using the factorization method works for polynomial functions, but the procedure is tedious. The procedure may not even work for transcendental functions like $y = \sin x$. The MVT provides another way to calculate the slope by using the slope of a secant line. In the above MVT, if B moves very close to A, then the mean value $(f(b) - f(a))/(b - a) = f'(c)$ in Theorem 1 is approaching the slope at A, since c is approaching a, forced by b approaching a. The formal writing is

$$\lim_{b \to a} \frac{f(b) - f(a)}{b - a} = f'(a). \tag{1.27}$$

This can also be considered a definition of derivative and is called *defining a derivative by a limit*. This procedure is efficient to calculate the derivative by hand and to derive many traditional derivative formulas in the earlier years of calculus.

1.5 Descartes' method of tangents and brief historical note

So, what is the origin of the direct calculus ideas described above? Numerous papers and books have discussed the historical development of calculus. Here we recount the development by a few major historical figures to sort out the origin of the main ideas of the calculus method outlined above. We cite only a few modern references and two works by Newton. Our focus is on (i) Descartes' method of tangents that is the earliest systematic way of finding the slope of a curve without using a limit, and (ii) Wallis' formulas of area, which were the earliest form of FTC. We do not intend to present a complete list of important works on the history of calculus.

In 1638, René Descartes (1596-1650) derived his method of tangents and included the method in his 1649 book, *Geometry* (see Cajori (1985), p.176). Descartes' method of tangents is purely geometric, constructing a tangent circle at a given point of a curve with the circle's center on the x-axis (Cajori (1985), pp. 176-177). Graphically, it is easier to draw a tangent circle than a tangent line using a compass and rulers. The tangent circle can be constructed using a radius line and moving the center on the x-axis so that the circle touches the curve at only one point. Then a tangent line can be drawn as the line perpendicular to the radial line of the circle at the tangent point.

Descartes' method of tangents also has an analytic description. The tangent circle is determined by the given point $P(x_0, y_0)$ on the curve and the moving center on the x-axis $(a, 0)$. The circle's equation is

$$(x-b)^2 + (y-0)^2 = (x_0-a)^2 + (y_0-0)^2. \tag{1.28}$$

The tangent condition requires that this equation and the curve's equation $y = f(x)$ have a double root at $P(x_0, y_0)$. This can determine a and hence the tangent circle. The radial line is determined by $P(x_0, y_0)$ and $(a, 0)$ and has its slope $m_R = y_0/(x_0 - a)$. The tangent line of the circle at point $P(x_0, y_0)$ is the tangent line of the curve at the same point. The slope of the tangent line is calculated as $m_T = -1/m_R = (a - x_0)/y_0$. In the above procedure, limit is not used.

Example 8. Use Descartes' method of tangents to find the slope of $y = \sqrt{x}$ at $(1, 1)$.

Substituting $y = \sqrt{x}$ into eq. (1.28), we obtain

$$(x-a)^2 + x = (1-a)^2 + 1. \tag{1.29}$$

This can be simplified to

$$x^2 + (1-2a)x + 2(a-1) = 0. \tag{1.30}$$

Knowing that $x = 1$ is a solution of this equation helps factorize the left-hand side

$$(x-1)(x+2-2a) = 0. \tag{1.31}$$

The double root condition for a tangent line requires $x_1 = x_2 = 1$. This leads to

$$x_2 + 2 - 2a = 1 + 2 - 2a = 0. \tag{1.32}$$

Hence, $a = 3/2$. The slope of the radial line is $m_R = 1/(1 - 3/2) = -2$. The slope of the tangent line is thus $m_T = 1/2$.

Although Descartes' method of tangents is complicated in calculation, its concept is simple, clear and unambiguous, and its geometric procedure is sound. It does not involve

small increments of an independent variable (developed by Fermat also in the 1630s), and hence it does not involve limits or infinitesimals. According to the point-slope equation of a tangent line presented earlier, the complexity of Descartes' method of tangents is unnecessary to calculate a slope. However, the point-slope form of a line was not known during Descartes' lifetime. According to Range [**6**], the point-slope form of a line was first introduced explicitly by Gaspard Monge (1746-1818) in a paper published in 1784. Thus, Monge's point-slope method of tangent appeared more than 100 years after Descartes' method of tangents.

Pierre de Fermat (1601-1665)'s method of tangents is similar to the modern method of differential quotient and uses a small increment (i.e., infinitesimal), which is ultimately set to be zero when the infinitesimal is forced to disappear from the denominator (Ginsburg et al., 1998). Thus, Fermat's method of tangents is more efficient for calculation from the point of view of limit, while Descartes' method of tangents is geometrically more direct and easier to plot by hand, and Monge's method of tangents is geometrically more direct. Fermat also used a sequence of rectangular strips to calculate the area under a parabola. His strips have variable width, which enabled him to use the sum of geometric series. This method of calculating an area can be traced back to Archimedes.

Archimedes (287-212 BC) 's method of exhaustion enabled him to find the area under a parabola. He used infinitely many triangles inscribed inside the parabola and also utilized the sum of geometric series.

Bonaventura Cavalieri (1589-1647) used rectangular strips of equal width to calculate the area under a straight line (i.e., a triangle) and under a parabola.

Around 1655-1656, John Wallis (1616-1703) derived algebraic formulas that represent the areas under the curve of simple functions, such as $y = kt$ and $y = kt^2$, from 0 to x (Ginsburg et al., 1998). Considering the existing work on tangents (i.e., slopes or derivatives) at that time, and considering the DA pair concept here, we thus may conclude that Wallis had already explicitly demonstrated, before Newton, the relationship between slope and area using examples, i.e., FTC.

Isaac Newton (1642-1727) attended Trinity College, Cambridge in 1660 and quickly made himself a master of Descartes' *Geometry*. He learned much mathematics from his teacher and friend Isaac Barrow (1630-1677), who knew the method of tangents by both Descartes and Fermat and also knew how to calculate areas under some simple functions. Barrow revealed that differentiation and integration were inverse operations, i.e., FTC (Cajori, 1985). Newton summarized the past work on tangents and area calculation, introduced many applications of the two operations, and made the tangent and area methods a systematic set of mathematics theory. Newton's method of tangents followed that of Fermat and had a small increment that eventually approaches zero. Namely, he used a sequence of secant lines to approach a tangent line as is done in modern calculus' definition of derivative. Although the "method of limits" is frequently attributed to Newton (see his book (Newton, 1729) entitled "*The Mathematical Principles of Natural Philosophy*" (p45)), he was not as adamant as Leibniz about letting an infinitesimal be zero at the end of a calculation. Newton was dissatisfied with the omitted small errors. He wrote that "in mathematics the minutest errors are not to be neglected" (see Cajori (1985), p198).

Newton's method of fluxions intended to solve two fundamental mechanics problems that are equivalent to the two geometric problems of slope and height increment of a curve pointed out earlier in in this paper:

"(i). The length of the space described being continually (i.e., at all times) given; to find the velocity of the motion at any time proposed.

(ii). The velocity of motion being continuously given; to find the length of the space de-

scribed at any time proposed" (see Cajori (1985), p.193, and Newton (1736), p.19).

The solution to these two problems also led to FTC, as geometrically interpreted in previous sections. The above statement of the two problems is directly cited from Newton's book "*The Method of Fluxions* "(Newton (1736), p.19), which was translated to English from Latin and published by John Colson.

Gottfried Wilhelm Leibniz (1646-1716) produced a profound work similar to Newton's that summarized the method of tangents and the method of area using a systematic approach. His approach has been passed on to today's classrooms, including his notations of derivative and integration.

Our description of calculus method has demonstrated that if we avoid calculating the area underneath a curve and define an integral by the height increment, we can readily extend Descartes' method of tangents to establish the theory of differentiation and integration by considering the slope (i.e., grade), DA pair, and height increment. The no-limit approach to calculus outlined in this paper is attributable to Descartes' original ideas, and is different from the those of Fermat, Newton and Leibniz. The FTC is attributable to Wallis' original ideas. Ginsburg et al. (1998) concluded that the query whether Leibniz plagiarized Newton's work on calculus is not really a valid question since the calculus ideas had already been developed by others before the calculus works of either Newton or Leibniz. Both just summarized the work of earlier mathematicians and developed differentiation and integration into a systematic branch of mathematics by using the methods of infinitesimals and limits. After their work, calculus became a very useful tool in engineering, natural sciences, and numerous other fields.

In addition to the aforementioned mathematicians, there are many others who contributed to the development of calculus, including Gregory of St. Vincent (1584-1667) from Spain, Gilles Persone de Roberval (1602-1675), Blaise Pascal (1623-1662), Christiaan Huygens (1629-1695), and Leonhard Euler (1707-1783). Augustin-Louis Cauchy (1789-1857) has been credited with the rigorous development of calculus from the definition of limits. Karl Weierstrass (1815-1897) corrected Cauchy's mistakes and introduced the delta-epsilon language we use today in mathematical analysis.

1.6 Conclusions on DD Calculus

We have introduced the concepts of derivatives and integrals without using limits. Geometrically, derivatives were introduced directly from the slope of a tangent line. Algebraically, derivatives and antiderivatives were introduced simultaneously as a DA pair. Then the integral was introduced as the height increment of the antiderivative function. This increment was geometrically interpreted as the area of the region bounded by the integrand function, the horizontal axis and the integration interval. A justification of this interpretation was given to demonstrate that this definition of area was reasonable and mathematically rigorous up to the proof of IVT which, like the Euclidean axioms, is intuitively true to most people in general public. At the end, we pointed out that limit was an efficient approach to calculate derivatives by hand and could help derive derivative formulas for complicated functions besides polynomials. We thus regard that the limit approach to calculus is an excellent computing method for finding derivatives by hand. In the pre-computer era, this limit approach was obviously critical in calculating derivatives of a variety of functions. In our current computer era, the limit approach to calculus is less essential and may be unnecessary for non-mathematical majors or general public at the introductory stage.

Although the ideas of direct calculus described in this paper come from practical applications, we have maintained the self-contained mathematical rigor and logic. Pure mathematical analysis regarding the structure of real line and the sequence approach to a compact set are not topics for this introduction. These analysis approaches mainly due to Cauchy and Weierstrass have certainly enriched calculus as begun by Archimedes, Descartes, Fermat, Wallis, Newton, Leibniz and others. However, our paper has shown that it is possible to introduce the basic concepts and calculation methods of calculus directly, without using limits.

One can also regard calculus as an extension of the trigonometry of a regular right triangle to the trigonometry of a curved-hypotenuse right triangle. For a regular right triangle, the slope (i.e., derivative = tangent of the angle) of the hypotenuse is derivative, and the vertical increment (i.e., an integral) is equal to the opposite side, which is the integral of the derivative (see Figure 11). It is obvious that the vertical increment and the slope are related and have the following relationships:

$$\tan\theta = \frac{BC}{AC} \qquad \text{(derivative)} \tag{1.33}$$

and

$$BC = \tan\theta \times AC \qquad \text{(integral).} \tag{1.34}$$

These two formulas are the FTC for the regular right triangle. The extension is from this straight line hypotenuse to the curved hypotenuse, such as a parabola or an exponential function. For the curved hypotenuse, the slope varies at different points and the total height increment is an integral. That is,

$$m = f'(x) \qquad \text{(derivative)} \tag{1.35}$$

and

$$BC = I[f'(x), a, b] \qquad \text{(integral).} \tag{1.36}$$

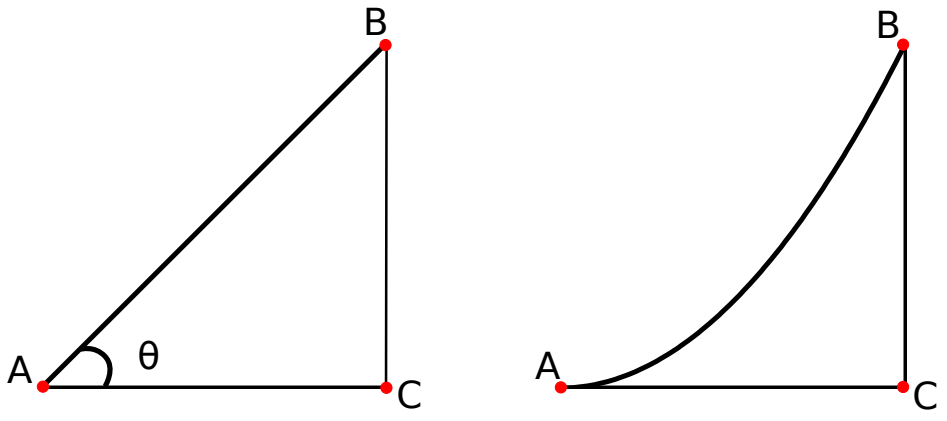


Figure 1.11 Calculus explained by triangles with straight and curved hypotenuses

1.7 Calculus from a Statistics Perspective

The following sections provide another approach to establishing the calculus method from the concept of mean, i.e., average of climate data. This approach is from a statistics perspective and can help calculus learners understand calculus ideas and analyze a function defined by data or sampling values from a given function, rather than an explicit mathematical formula. The basics of this approach are two averages: arithmetic mean and graphic mean. The arithmetic mean is used to define integral. Area is used to interpret the meaning of an integral. Antiderivative is introduced from integral, and derivative-antiderivative pair is introduced as a mathematical operation entity. The graphic mean is an average speed in an interval and is used to interpret the meaning of a derivative.

How can one use elementary statistics method to help explain ideas of calculus? The purpose of this paper is to provide an approach to developing calculus methods from a statistics perspective. Our approach is motivated in part by mathematical modeling (Kaplan et al., 2012), and can be helpful for enhancing the link between elementary statistics courses and calculus courses. The link is developed from statistics to calculus and is different from the traditional connection from calculus to statistics.

A mathematical function $y = f(x)$ can be represented in four ways (see p11 of Stewart (2008)): (i) a description in words or text, (ii) a table of at least two columns of data, (iii) a graph in the x-y coordinates plane, and (iv) an explicit mathematical formula. Conventional calculus textbooks (e.g., Anton et al. (2008) and Stewart (2008)) almost exclusively deal with Case (iv): functions defined by explicit formulas. The graphic representations (i.e., Case (iii)) are dealt with as a supporting tool for the function formula. Case (ii), data-represented functions, is usually considered the category of statistics, which analyzes data and makes inferences for the data's implications. Hence, most calculus textbooks do not deal with this case even in the modern era of computers. However, the rapid development of data-based information in our practical life requires us to broaden the traditional statistics methods, including interconnections between statistics and calculus. Today's speed and availability of personal computers and smart phones make it possible to take new and innovative approaches to calculus for functions represented in any of the four ways. This paper will develop the calculus ideas from statistics perspective. We will build the calculus concept for the function of Case (ii) and make the concept applicable to the functions of all the four cases. We use two types of averages: arithmetic mean and graphic mean. The arithmetic mean and law of large numbers (LLN) are used to define an integral, and graphic mean is to define a derivative.

We do not intend to use this approach to replace the conventional way of teaching calculus. Instead, our proposed approach may be used as supplemental material in today's classrooms of conventional calculus and hence provides a new perspective for explaining calculus ideas to students. Further, we do not claim that this approach is superior to other approaches to calculus. Every approach to calculus has its own disadvantages. Ours is not an exception.

Mathematicians have experimented the new approach in two different courses. Samuel Shen of San Diego State University (SDSU) used it in Calculus I for engineering and physical sciences in the Fall 2011, where he spent two hours altogether at the beginning of the course on the new approach in a class of 120 students. After the two hours, he asked his students to explain the concept of integral to their grandmother. Dov Zazkis taught this material to four students in the summer of 2011 in the third mathematics course at SDSU in a sequence of four courses for prospective elementary school teachers. This third mathematics course's main contents are on probability and statistics. He made an

educational pedagogy study from this course and spent over four hours on this approach with emphasis on understanding the concept of integrals. Therefore, our modest goals of the calculus from statistics perspective are (a) to provide a set of supplemental materials to explain the the concept of calculus from a non-traditional perspective, and (b) to provide research opportunities for mathematics education.

To simplify the description of our approach to the integral and derivative concept, we limit our functions to those that are continuous and smooth with positive domain and positive range, when discussing the Case (iv) function $y = f(x)$. These limits do not affect the rigor of the mathematics developed here and can be easily removed when a more sophisticated mathematics of calculus is introduced.

1.8 Arithmetic mean, sampling, and average of a function

For a given location on Earth, its temperature's variation with respect to time forms a functional relationship. Consider the arithmetic mean for the annual surface air temperature (SAT) data (in units [°C]) data from 1951-2010 at Fredericksburg (38.32°N, 77.45°W), which is 80 kilometers southwest of Washington DC, United States (See Table 1). The data pairs $(x_i, f_i)(i = 1, 2, \cdots, n)$ represent the tabular form of a function (the Case (ii) function) with x_i for time and f_i for temperature, and n is 60 for this case. *Arithmetic mean* is defined as

$$\bar{f} = \frac{\sum_{i=1}^{n} f_i}{n}. \tag{1.37}$$

For the SAT data in Table 1, the mean is 13.2[°C].

Table 1. Annual mean SAT at Fredericksburg from 1951- 2010

Annual mean surface air temperature at Fredericksburg (38.32°N, 77.45°W) Virginia, United States. The temperature units are [°C]. The data are from the U.S. Historical Climatology Network. http://www.ncdc.noaa.gov/oa/climate/research/ushcn/											
1951	13.1	1961	12.9	1971	13.2	1981	12.5	1991	14.3	2001	13.7
1952	13.4	1962	12.3	1972	12.8	1982	12.7	1992	12.5	2002	14.3
1953	14.4	1963	12.4	1973	13.7	1983	12.8	1993	13.2	2003	13.0
1954	13.7	1964	13.2	1974	13.3	1984	13.0	1994	13.2	2004	13.6
1955	12.6	1965	12.7	1975	13.0	1985	13.6	1995	13.2	2005	13.7
1956	13.1	1966	12.4	1976	12.6	1986	13.4	1996	12.6	2006	14.2
1957	13.3	1967	12.3	1977	13.1	1987	13.4	1997	13.0	2007	14.1
1958	12.1	1968	12.8	1978	12.3	1988	12.6	1998	14.7	2008	13.7
1959	13.6	1969	12.5	1979	12.5	1989	12.9	1999	13.9	2009	13.2
1960	12.6	1970	13.1	1980	12.7	1990	14.3	2000	13.0	2010	14.2

Next we explore the data that are samples from the Case (iv) functional values. The samples of independent and dependent variables are $(x_i, y_i), i = 1.2. \cdots, n$, where the Case (iv) function is denoted by $y = f(x)$. The sampling of x is usually done in three ways.

1. Uniform sampling: The distance between each pair of neighborhood samples is the same, i.e., $h = x_i - x_{i-1}$ is the same for each i. The sample points can be determined

by $x_i = x_{i-1} + h, i = 1, 2, \cdots, n$. An example is the SAT sampling in Table 1 whose h is one year. Another example is a uniform sampling of a function shown in Figure 1.

2. Random sampling: Random sampling is, by name, a probability process and can be determined by a predefined way using a statistics software. Many software tools for generating random numbers are available in pubic domain for free, such as WolframAlpha: `www.wolframalpha.com`. The command `RandomReal[{0,2},10]` yields 10 random real numbers in the interval $[0, 2]$. A trial of this command is $\{1.25685, 1.02584, 1.26752, 0.813785, 0.224482, 0.905491, 1.73265, 1.44104, 1.95108, 0.926807\}$
Of course, each trial yields a different random result.

3. Convenience sampling: This sampling may be neither as evenly spaced as the uniform sampling nor completely random, but the samples are taken from convenient and practical locations, such as the location of weather stations, which could be neither at the top of Himalayas nor over the midst of Pacific. Thus, convenient sampling is often encountered in practice, such as the data from geology, meteorology, hydrology, agriculture and forestry, but suffers possible drawbacks of bias and incompletion. When given a Case (iv) function, this sampling is unnecessary, but for a function of the other three cases in practical applications, this sampling is essential to describe a function for applications.

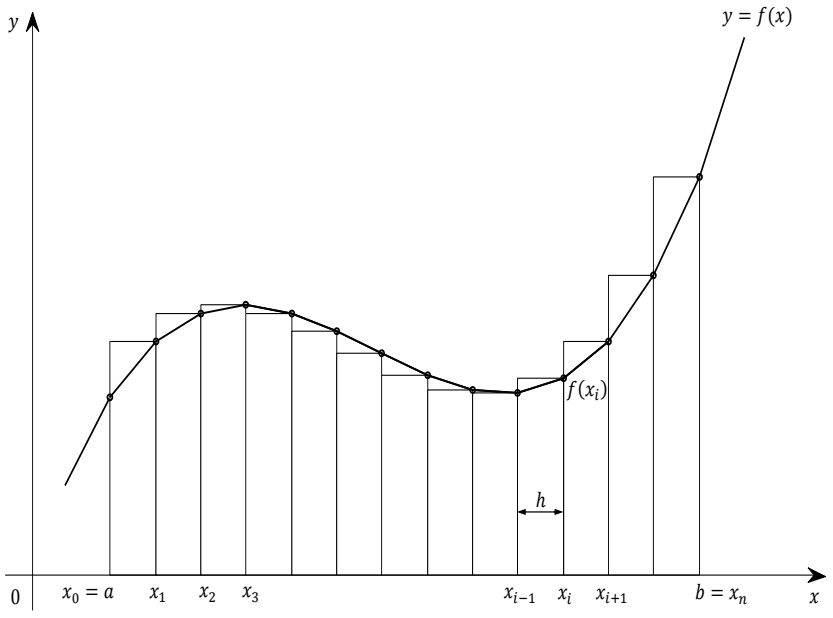


Figure 1.12 Uniform sampling for a function $y = f(x)$ in $[a, b]$. The area under the curve is approximated by the sum of the narrow rectangular strips

Let us explore two examples of calculating arithmetic means of sample data from a function: $y = x$ and $y = x^2$.

From our intuition or assisted by a figure (see Figure 2 with $a = 0$), we can infer that the mean of the function $y = x$ in $[0, 1]$ is $1/2$. This result can be simulated by different sampling methods.

1. Uniform sampling: We use a sample of size 100 with the first sample at $x = 0.01$ and last sample at $x = 1$. These 100 points divide the interval $[0, 1]$ into 100 equal sub-intervals of length $1/100$:

$$x_1 = 0.01, x_2 = 0.02, \cdots, x_i = \frac{i}{100}, \cdots, x_{100} = 1.$$

$$y_i = x_i, i = 1, 2, \cdots, 100$$

The arithmetic mean of data $y_i, i = 0, 1, \cdots, 100$ is equal to

$$\hat{\bar{y}} = \frac{\sum_{i=1}^{100} y_i}{100} = \frac{\sum_{i=1}^{100}(i/100)}{100} = \frac{\sum_{i=1}^{100} i}{100 \times 100} = 0.505, \tag{1.38}$$

which is very close to the exact average value 0.5.

2. Random sampling: Random sampling usually does not sample the end points. Our 100 samples are at the internal points of $[0, 1]$, i.e., over $(0, 1)$. WolframAlpha command `RandomReal[{0,1},100]` generates 100 random numbers whose mean is 0.5312, which is close to the exact result 0.5. Again, the result 0.5312 is different each time the command is implemented due to the random nature of the `RandomReal` generator. It is intuitive that larger samples should be likely to lead to more accurate approximation. When using 1000 samples, a result is 0.4981, which is very close to 0.5.

Next we consider $y = x^2$ in $[0, 1]$ (see Figure 3). The uniform sampling of using 100 points as above yields the following:

$$\hat{\bar{y}} = \frac{\sum_{i=1}^{100} y_i^2}{100} = \frac{\sum_{i=1}^{100}(i/100)^2}{100} = \frac{\sum_{i=0}^{100} i^2}{100^2 \times 100} = \frac{100 \times (100+1)(2 \times 100+1)/6}{100^2 \times 100}$$
$$= \frac{101 \times 201}{6 \times 100^2} = .33835$$

The random sampling with WolframAlpha uses a command `RandomReal[{0,1},10]^2` to generate the data, and the text result can be copied and used to calculate the mean by using WolframAlpha. A result is 0.36726. The accurate mean value should be 0.33333, which can be computed from n samples when n is very large.

WolframAlpha has limited statistical computing power. The open source statistics software R and the commonly used MS Excel have more statistical computing power and can be used for practical applications and research. An Excel calculation of the sample mean for $y = x^2$ in $[0, 1]$ for different number of samples is given in Table 2, which indicates that, in general, the accuracy of the mean improves as the sample size n increases. As a matter of fact, the improvement of the accuracy can be quantified by the error $\sigma/\sqrt{n}$ where σ is the standard deviation of the population being sampled. This is a result of the Central Limit Theorem (CLT) in statistics (see a basic statistics text, e.g., Johnson and Bhattacharyya (1996) and Wackerly et al. (2002)). The CLT states that the mean of samples with sufficiently large sample size is normally distributed. CLT further asserts that

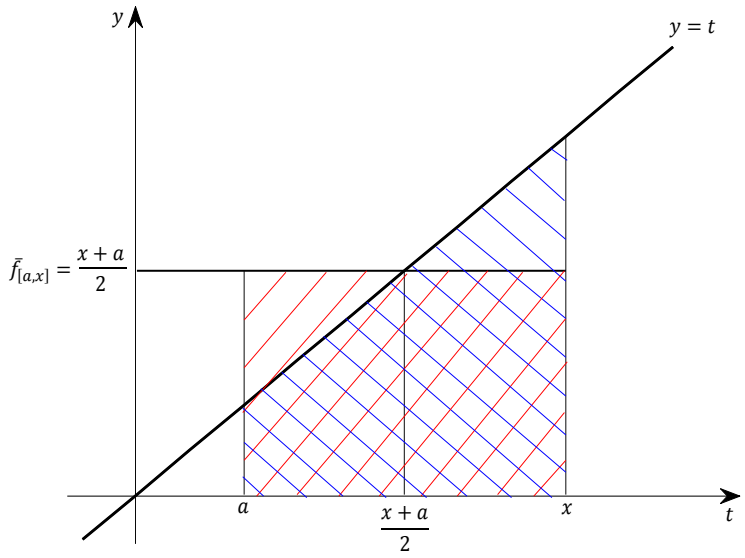


Figure 1.13 Arithmetic mean of a linear function.

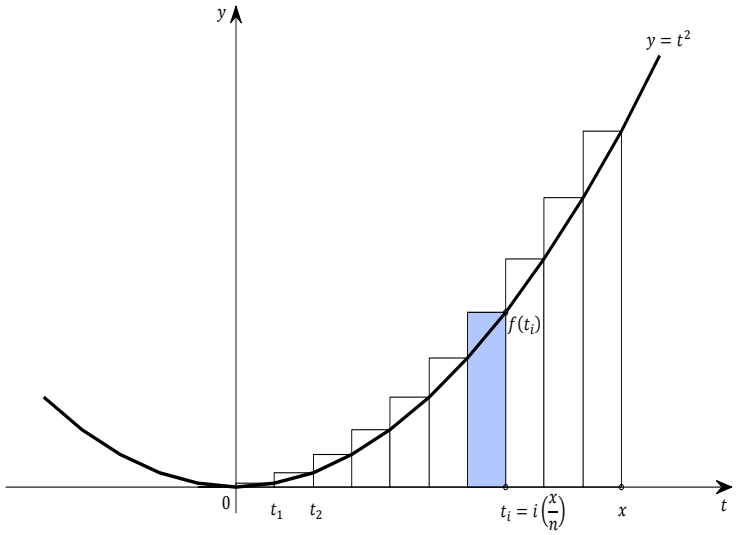


Figure 1.14 Uniform sampling of a parabolic function.

(i) the expected value of the sample mean is the same as the population mean, and (ii) the variance of the sample mean is $(1/n)$th of the population variance. These two assertions are widely assumed in practical statistical applications and computer calculations, such as Monte Carlo simulations. This is normal distribution result can be intuitive to students and may be accepted as an axiom. However, instructor does not need to teach the CLT to students in this first introduction of calculus from statistics perspective. This is like the case

that the first introduction of conventional calculus does not need to prove the existence of a limit from the rigorous $\delta - \epsilon$ argument.

Table 2. Convergence of the sample mean as the sample size increases.

Sample Size n	10	100	1,000	10,000	100,000	1,000,000
Uniform Sampling	0.3850	0.3384	0.3338	0.3334	0.3333	0.3333
Random Sampling						
Trial 1	0.4350	0.3505	0.3463	0.3344	0.3315	0.3332
Trial 2	0.3560	0.3058	0.3284	0.3325	0.3331	0.3332
Trial 3	0.4640	0.3518	0.3355	0.3317	0.3357	0.3332
Average	0.4183	0.3360	0.3367	0.3329	0.3335	0.3332

Let $\hat{\bar{f}}[n]$ denote the mean from n samples. Then, it is almost always true that $\hat{\bar{f}}[n]$ approaches the true mean of the function as n increases. This convergence has a probability equal to one. The probability for this to be false is zero. This intuitive conclusion is the strong LLN (see Wacherly et al., 2002), which asserts that the event of the following limit being true has a probability equal to one:

$$\lim_{n\to\infty} \hat{\bar{f}}[n] = \bar{f}. \tag{1.39}$$

Here $\bar{f}$ is the limit of the sequence $\{\hat{\bar{f}}[1], \hat{\bar{f}}[2], \hat{\bar{f}}[3], \cdots\}$, or simply $\{\hat{\bar{f}}[n]\}$.

For example, when x^2 is uniformly sampled by n points over $[0, 1]$, the mean is

$$\hat{\bar{f}}[n] = \frac{\sum_{i=1}^{n}(i/n)^2}{n} = \frac{\sum_{i=0}^{n} i^2}{n \times n^2} = \frac{n(n+1)(2n+1)/6}{n^3} = \frac{n+1}{n} \times \frac{2n+1}{6n}.$$

As $n \to \infty$, the first factor goes to 1 and the second to 2/6=1/3. Thus, the above mean approaches $1/3 \approx 0.33333333$ as $n \to \infty$.

The above leads to the definition of *average of a function* $y = f(x)$ in an interval $[a, b]$

$$\bar{f} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(x_i)}{n}, \tag{1.40}$$

where $\{x_i\}_{i=1}^{n}$ are sampled from $[a, b]$ by uniform sampling, random sampling, and possibly convenience sampling. We also call $\bar{f}$ *arithmetic mean*, or just *mean*, of the function $y = f(x)$.

1.9 Definition of integral, antiderivative, and DA pair

Definition 1. (Definition of integral). it If $\bar{f}$ is the average of the function $f(x)$ over the interval $[a, b]$, then $(b - a) \times \bar{f}$ is defined as the integral of $f(x)$ over $[a, b]$, denoted by

$$I[f, a, b] = (b - a)\bar{f} \tag{1.41}$$

This serves as the definition of the definite integral in conventional calculus. Since we will not introduce the concept of indefinite integral, we thus treat "integral" here as the "definite integral."

From the examples of mean in the above section, we can calculate the following integrals:

$$I[x, 0, 1] = (1 - 0) \times (1/2) = 1/2,$$

and

$$I[x^2, 0, 1] = (1 - 0) \times (1/3) = 1/3.$$

Then, what is the graphic meaning of integral from the above definition of integral? We use uniform sampling to make an interpretation. Since $\hat{\bar{f}}[n] \approx \bar{f}$, we have

$$I[f, a, b] \approx (b - a)\hat{\bar{f}}[n] = (b - a)\frac{\sum_{i=1}^{n} f_i}{n} = \sum_{i=1}^{n} \frac{b - a}{n} f_i.$$

Here $f_i = f(x_i)$ is the sample value of the function at the sampling location x_i (see Figure 1). Since this is uniform sampling, $x_{i+1} - x_i = h = (b - a)/n$. Thus, each term in the above sum

$$\frac{b - a}{n} f_i$$

is the area of a rectangular strip with base $h = (b-a)/n$ and height f_i, as shown in Figure 1. Thus, $(b-a)\hat{\bar{f}}[n]$ is the area of under the echelon, formed by n rectangular strips whose heights are determined by the function values from a uniform sampling $f(x_i) = f(ih), i = 1, 2, \cdots, n$. This sum approaches the true area, denoted by S, of the region under the curve $y = f(x)$ in $[a, b]$. The ever improving approximation as $n \to \infty$ is a process of limit and is denoted by

$$\lim_{n\to\infty} (b - a)\hat{\bar{f}}[n] = S = I[f, a, b]. \tag{1.42}$$

Dividing both sides by $b - a$ leads to the limit of $\hat{\bar{f}}[n]$ as $n \to \infty$

$$\lim_{n\to\infty} \hat{\bar{f}}[n] = \bar{f}. \tag{1.43}$$

As mentioned earlier, the existence of this limit is guaranteed by LLN, and the probability for this limit to fail is zero.

Therefore, the geometric meaning of the integral $I[f, a, b]$ is the area of the region bounded by $y = f(x)$, $y = 0$, $x = a$ and $x = b$.

Following the Shen and Lin (2014)'s idea of derivative-antiderivative (DA) pair, we can introduce the concept of antiderivative and derivative. We define the integral $I[f, a, x]$ as *antiderivative* of $f(x)$ and is denoted by $F(x)$, where a is an arbitrary constant and x is regarded as a variable. We also call $f(x)$ the derivative of $F(x)$ and is denoted by $F'(x) = f(x)$. Because x is regarded as a variable, antiderivative $F(x)$ is a function. The function pair $(f(x), F(x))$ is called a *DA pair*.

From the geometric meaning of integral, $F(x) = I[f, 0, x]$ is the area of the region bounded by $y = f(t)$, $y = 0$, $t = 0$ and $t = x$ over the $t - y$ plane (see Figure 3).

Example 1. Evaluate the antiderivative of $f(x) = 1$.

The area of the rectangle bounded $y = 1$, $y = 0$, $t = 0$ and $t = x$ over the $t - y$ plane is x. Thus, $F(x) = x$, and $(1, x)$ is a DA pair.

Example 2. Evaluate the antiderivative of $f(x) = x$.

The area of the triangle bounded $y = t$, $y = 0$, $t = 0$ and $t = x$ over the $t - y$ plane is $x^2/2$ (See Figure 2). Thus, $F(x) = x^2/2$, and $(x, x^2/2)$ is a DA pair, i.e., $(x^2/2)' = x$ and $I[t, 0, x] = x^2/2$.

Example 3. Evaluate the antiderivative of $f(x) = x^2$.

This is a problem of calculating the area of a curved triangle (See Figure 3). We can use the uniform sampling as we have done above for $y = t^2$ in the interval $[0, 1]$, but now for the interval $[0, x]$.

$$\hat{\hat{f}}[n] = \frac{\sum_{i=1}^{n}(xi/n)^2}{n} = x^2\frac{\sum_{i=0}^{n} i^2}{n \times n^2} = x^2\frac{n(n+1)(2n+1)/6}{n^3} = x^2\frac{n+1}{n} \times \frac{2n+1}{6n}.$$

This mean approaches $x^2/3$ as $n \to \infty$. Hence,

$$F(x) = (x - 0) \times x^2/3 = x^3/3.$$

We thus have a DA pair $(x^2, x^3/3)$, i.e., $(x^3/3)' = x^2$ and $I[t^2, 0, x] = x^3/3$.

It is tedious to find an antiderivative by definition as shown above. Fortunately, many free software packages, such as WolframAlpha, are now available over the internet and apps for smart phones (Shen and Lin (2014)). For example, in the pop up box from WolframAlpha, type `integral x^4`. Press the enter key. The WolframAlpha returns the antiderivative $x^5/5$. So $(x^4, x^5/5)$ is a DA pair. In the same way, WolframAlpha can generate the following DA pairs.

1. Power function: $(x^n, x^{n+1}/(n + 1))$.
2. Exponential function: (e^x, e^x).
3. Natural logarithmic function: $(\frac{1}{x}, \ln x)$.
4. Sine function: $(\cos x, \sin x)$.
5. Cosine function: $(-\sin x, \cos x)$.
6. Tangent function: $(\sec^2 x, \tan x)$.

One can also use `www.wolframalpha.com` to find derivatives. For example, in the pop up box, type `derivative x^5/5` and press enter. It returns x^4. This verifies that $(x^5/5)' = x^4$.

1.10 Calculation of an integral $I[f, a, b]$

We can use antiderivative to calculate an integral. From the area meaning of an integral (see Figure 4), we have the following formula:

$$I[f, c, d] = F(d) - F(c). \tag{1.44}$$

This is also denoted by $I[f, c, d] = F(x)|_c^d$. Namely, the area of the region bounded by $y = f(x)$, $y = 0$, $x = c$, and $x = d$ is equal to the difference of the area bounded by $y = f(x)$, $y = 0$, $x = a$, and $x = d$ minus the area bounded by $y = f(x)$, $y = 0$, $x = a$, and $x = c$. This is shown in Figure 4: $I[f, c, d]$ is the area CDD′C′, equal to area ODD′O′minus area OCC′O′.

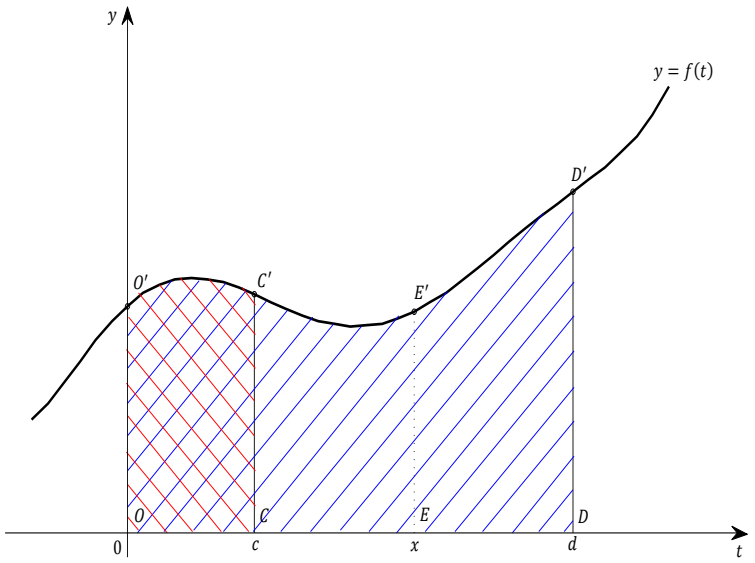


Figure 1.15 Areas and their differences under a curve: an illustration for FTC.

In traditional calculus, this way of calculating an integral is called Part II of the Fundamental Theorem of Calculus (FTC), and the DA pair is called Part I of FTC (e.g., Stewart (2008)).

Example 4. Evaluate $I[x^2, 0, 1]$.
Since $F(x) = I[x^2] = x^3/3$, $I[x^2, 0, 1] = x^3/3|_0^1 = 1^3/3 - 0^3/3 = 1/3$.

Example 5. Evaluate $I[sin^2(x), 0, \pi]$.
In `www.wolframalpha.com`, the command `integrate sin^2 x` yields $F(x) = (1/2)(x - \sin x \cos x)$. Thus,

$$\begin{aligned} I[\sin^2 x, 0, \pi] &= (1/2)(x - \sin x \cos x)|_0^\pi \\ &= (1/2)(\pi - \sin \pi \cos \pi) - (1/2)(0 - \sin 0 \cos 0) = \pi/2. \end{aligned}$$

One can also use `www.wolframalpha.com` to calculate the integral directly by using the command `integrate [sin^2 x, 0, pi]` or `integral [sin^2 x, 0, pi]`. This command directly returns value $\pi/2$.

1.11 Use average speed and graphic mean to interpret the meaning of a derivative

In the above we have successfully introduced the concepts and calculation techniques of integral and derivative from statistics perspective. Further, the meaning of integral of a function in $[a, b]$ is well interpreted as the value increment of its anti-derivative function from $x = a$ to $x = b$. This section attempts a remaining task of interpreting the meanings of derivative of a function and makes connections with the conventional way of defining a derivative.

If one drives along a freeway from Exit 6 at 2pm and arrives at Exit 98 at 4pm, one's average speed is $(98-6)/(4-2) = 46$ mph. In general, we use $s(t)$ to represent the location of the car at time t, then the distance traveled by the car from time t_1 to time t_2 is $s(t_2) - s(t_1)$. The average speed is

$$\bar{v} = \frac{s(t_2) - s(t_1)}{t_2 - t_1}. \tag{1.45}$$

This kind of average is called a *graphic mean*, in contrast to the arithmetic mean discussed earlier. Figure 5 gives a schematic illustration of function $y = s(t)$. The graphic mean is thus the slope of the secant line, i.e., the purple line that connects P_1 and P_2 in Figure 5. Namely, $\bar{v} = \tan\theta$.

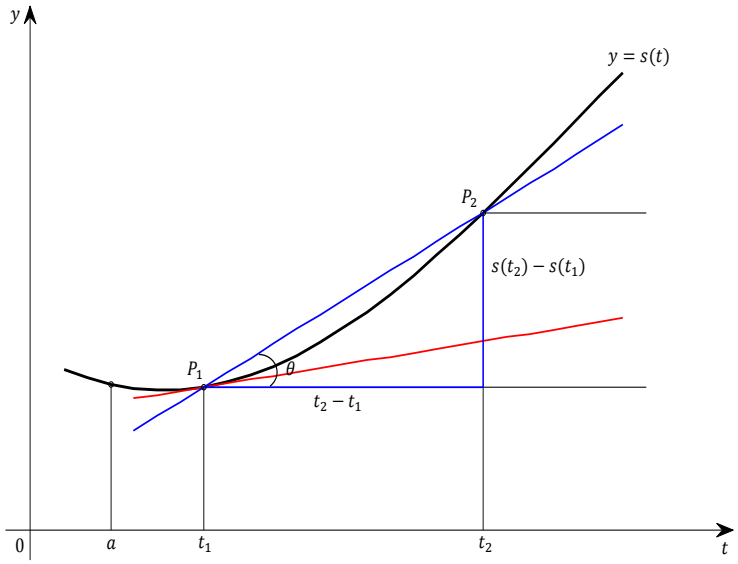


Figure 1.16 Illustration of a graphic mean.

The instant speed at a time, say t_1, is approximately the average speed in an ever smaller interval $[t_1, t_2]$, which means t_2 goes to t_1. The limit expression is

$$\lim_{t_2 \to t_1} \frac{s(t_2) - s(t_1)}{t_2 - t_1} = v(t_1). \tag{1.46}$$

This is regarded as the *slope* of $y = s(t)$ at t_1, i.e., the slope of the tangent line (i.e., the red line in Figure 5) of $y = s(t)$ at t_1. We would like to show that $s(t)$ is an antiderivative of $v(t)$, i.e., $s(x) = I[v(t), 0, x]$. We divide $[0, x]$ uniformly into n sub-intervals of width $h = x/n$. Within interval $[x_i, x_{i+1})$, the distance traveled is approximately $v(x_i)h$. The total distance traveled in $[0, x]$ is $\sum_{i=1}^{n} v(x_i)h$, which goes to $s(x) - s(0) = I[v, 0, x]$ as $n \to \infty$. Thus, $(v(t), s(t))$ is a DA pair, i.e., the derivative of distance with respect to time is speed, and the integral of speed is the total distance traveled. Equation (1.46) provides another way of defining derivative in addition to the derivative introduced via DA pair. This definition of derivative by limit in (1.46) is the most popular way of introducing derivatives in today's classrooms. It is also the main idea of Isaac Newton (1642-1727)'s

approach to fluxion (see Newton (1736) for his book of *Method of Fluxion*, 1736). Before Newton, Pierre de Fermat (1601-1665) developed a method of tangents in 1629 using a small increment (i.e., infinitesimal) which is finally set to be zero when the infinitesimal disappears from the denominator (see Cajori (1985) and Ginsburg et al. (1998)). Newton's method of fluxion or tangent is similar to Fermat's but introduced the concept of limit, although he did not explicitly use a mathematical notation or formula to express the limit [see Newton (1729) for his book "*The Mathematical Principles of Natural Philosophy*" (1729, p45)].

In summary, a derivative of $F(x)$ is a limit of its graphic mean. One can use the formulas of DA pairs given in the last section to find the derivative of a given function. For example, to find the derivative of x^2, we use the DA pair $(x^n, x^{n+1}/n+1)$. With $n = 1$, the pair becomes $(x, x^2/2)$. So $(x^2/2)' = x$, hence $(x^2)' = 2x$.

1.12 Conclusions and discussion

We have used arithmetic mean to define integral. This definition leads to the DA pair concept, integral's area meaning, and FTC. We have used graphic mean to interpret the meanings of a derivative as the slope of a curve or speed of moving object. The graphic mean also provides another way of calculating derivative by limit, an application of Fermat's method of tangents. Computer calculation for DA pairs is used here in place of traditional derivation of derivative formulas using limit and graphic mean. Before today's popularity of computer and internet, the method of introducing derivative by limit was very useful in the calculation of calculus problems, since the method can easily be used to derive various kinds of derivative formulas for non-polynomial functions, compared to the DA pair approach $I[f, 0, x]$. However, with today's easy access to notebook computers, smart phones, and publicly available software, finding derivatives and integrals for a given function can be readily done on the internet. Thus, today's calculus teaching may shift its emphasis from the limit-based hand calculations to the web-based calculations, concepts, interpretations, and applications. Limit is a difficult, subtle, puzzling and intermediate procedure for calculus. Although in our statistics calculus here the limit notation is introduced to state the existence of the arithmetic mean following LLN, we hardly used the concept of limit in formulation derivations and actual calculations. Statistical calculus' rigorous background needs LLN, which requires a rigorous proof but its conclusion has a certain degree of intuition and can be treated as an axiom like those in Euclidean geometry. Thus, the calculus based on LLN implies that, with today's handy computer technology, the calculus basics can be axiomized and the calculus method can be developed directly with minimum and intuitive theory without the need of some unnecessary, intermediate and complex procedures. In this way, calculus can be properly taught in high schools or in a workshop of a few hours as proposed in Lin (2010).

REFERENCES

1. H. Anton, I.C. Bivens, and S. Davis, *Calculus: Early Transcendentals, Single Variable*, 9th ed. John Wiley and Sons, New York, 880pp, 2008.
2. F. Cajori, *A History of Mathematics* (pp. 162-198), 4th ed., Chelsea Publishing Co., New York, 534pp, 1985.
3. J.L. Coolidge, The story of tangents. *American Math. Monthly* **58** (1951) 449-462.

4. D. Ginsburg, B. Groose, J. Taylor, and B. Vernescu, The History of the Calculus and the Development of Computer Algebra Systems, *Worcester Polytechnic Institute Junior-Year Project*, www.math.wpi.edu/IQP/BVCalcHist/calctoc.html, 1998

5. R. A. Johnson and G.K. Bhattacharyya, *Statistics: Principles and Methods.* 3rd ed., John Wiley and Sons, New York, 720pp, 1996.

6. D. Kaplan, D. Flath, R. Pruim, and E. Marland, *MAA Ancillary Workshop: Teach Modeling-based Calculus*, at Sheraton Boston on January 3rd, 2012. `www.causeweb.orgworkshopjmm12_modeling`.

7. Leung, K., C. Rasmussen, S.S.P. Shen, and D. Zazkis, 2014: Calculus from a statistics perspective, *The College Mathematics Journal*, 45, 377-386. Also see the original manuscript at the Cornell University Library public access arXiv (http://arxiv.org/abs/1406.2731) with the same title authored by Shen, Zazkis, Leung and Rasmussen.

8. Q. Lin, *Calculus for High School Students: from a Perspective of Height Increment of a Curve*, People's Education Press, Beijing, 2010.

9. Q. Lin, *Fastfood Calculus*, Science Press, Beijing, 2009.

10. I. Newton, *The Method of Fluxions and Infinite Series with Application to the Geometry of Curve-Lines*, Translated from Latin to English by J. Colson, Printed for Henry Woodfall, London, 339pp, 1736. [http://books.google.com]

    `http://books.google.com/books?id=WyQOAAAAQAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false`

11. I. Newton, *The Mathematical Principles of Natural Philosophy*, Translated from Latin to English by A. Motte, Printed for Benjamin Motte, London, 320pp, 1729. [http://books.google.com]

    `http://books.google.com/books?id=Tm0FAAAAQAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=true`

12. R.M. Range, Where are limits needed in calculus? *Amer Math Monthly* **118** (2011) 404-417.

13. Shen, S.S.P., 1995: Climate Sampling Errors, Lecture Notes at University of Tokyo, 74pp.

14. S.S.P., Shen, and Q. Lin, 2014: Two hours of simplified teaching materials for direct calculus, *Mathematics Teaching and Learning*, No. 2, 2-1 - 2-6. English translation available at Cornell University Library public access arXiv entitled DD Calculus with file number: arXiv:1404.0070. http://arxiv.org/abs/1404.0070

15. J. Stewart, *Single Variable Calculus - Early Transcentals*. 6th ed., Thomson Brooks/Cole, Belmont, 763pp, 2008.

16. J. Susuki, The lost calculus (1637-1670): Tangency and optimization without limits. *Mathematics Mag.* **78** (2005) 339-353.

17. D.D. Wackerly, W. Mendenhall III, and R.L. Scheaffer, *Mathematical Statistics with Applications. 6th ed., Duxbury, 853pp, 2002.*

EXERCISES

1.1 Climatology is defined as the mean state, or normal state, of a climate parameter, and is calculated from a period of time called the climatology period (e.g., 1961-1990). Thirty years are commonly in the climate community as the standard length of the climatology period. Due to the highest density of weather stations in 1961-1990, people

often use 1961-1990 as their climatology period, although some now chose 1971-2000 or 1981-2010. Surface air temperature (SAT) is defined as the temperature inside a Stevenson's screen box about 2 meters above the ground. Daily maximum temperature (Tmax) is the maximum temperature measured inside the screen box by a maximum temperature thermometer within 24 hours.

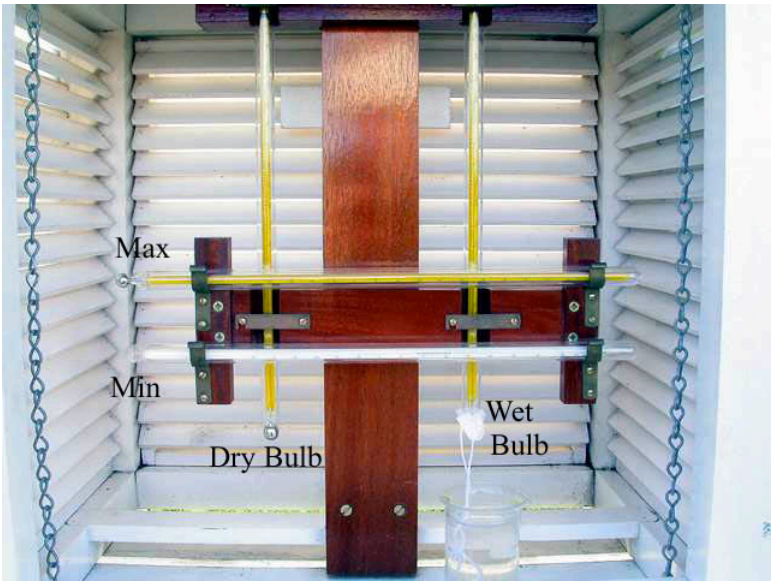


Figure 1.17 Inside a Stevenson's screen box, invented by Thomas Stevenson in 1864, and recommended by the World Meteorological Organization (WMO) to measure Tmax and Tmin using two thermometers. The data were recorded every 24 hours. Tmax and Tmin are the previous 24 hours temperature extremes and depend on the time of observation. Thus, the observations have the time of observation bias (TOB) due to the inconsistent time of data recording. USHCN dataset has the TOB corrected data, as well as the raw data.

Go to the United States Historical Climatology Network (USHCN) website

`http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn_map_interface.html`

and find the climatology of the August Tmax at Cuyamaca station (USHCN Site No. 042239) near San Diego using the 1961-1990 climatology period.

1.2 Express the Tmax climatology as an integral when regarding Tmax as a function of time t, using the definition of an integral from the statistics perspective.

1.3 Use Excel or R or another computer program to find the trend of the January mean Tmin at Cuyamaca station in the period of

a) 1951-2010,
b) 1961-2010,
c) 1971-2010, and
d) 1981-2010.

1.4 Trend and derivative:

a) Use derivative to explain the trends of the above exercise, and

b) Plot the January Tmin as a function of time from 1951-2010. Use the curve and its derivative to explain the rate of change and the change of the Cuyamaca station Tmin for a given period of time. Use derivative and integral as much as you can.

CHAPTER 2

BASICS OF R PROGRAMMING

An important feature that makes this book different from other mathematics and statistics tool book for climate science is that we refer to modern computers to do the complex and tedious algebra. Students can thus focus on correct usage of the tools with accurate statement of assumptions supported by climate data and models, because they are relieved from the mountains of formulas, rules and theorems. Among many software packages are used in climate community, R's popularity has dramatically increased in the last a few years due to its enormous power of handling big data. We thus choose to include the basics of R for this book. A student who has mastered the R examples used in this book should be able to handle most cases of climate data analysis and presentations for both research and applications.

2.1 Download and install R software package

For Windows users, visit the website

`https://cran.r-project.org/bin/windows/base/`

to find the instructions of R program download and installations.

For Mac users, visit

`https://cran.r-project.org/bin/macosx/`

For details about the publicly open access R-Project, visit

Climate Mathematics,
First Edition.

https://www.r-project.org/

The beginners of R would find it very difficult to navigate through this official, formal, detailed, and massive R-Project documentation to learn the program. Fortunately, many excellent tutorials for a quick learn of R programming are available online and in Youtube. One can google around and find a couple of his preferred tutorials.

2.2 Youtube tutorial: for true beginners

A very good and slow paced youtube tutorial: Ch. 1. An Introduction to R

https://www.youtube.com/watch?v=suVFuGET-0U

2.3 Youtube tutorial: Reading in csv files into R

https://www.youtube.com/watch?v=QkE8cp0B9gg

2.4 R tutorial for climate science

https://www.isse.ucar.edu/ams/present/elsner.pdf

http://www.ats.ucla.edu/STAT/r/

http://stats.stackexchange.com/questions/72421
/showing-spatial-and-temporal-correlation-on-maps

CHAPTER 3

BASIC STATISTICAL METHODS FOR CLIMATE DATA ANALYSIS

Statistics originated from Latin "status" meaning "state" and is a suite of scientific methods that analyze data and make credible conclusions. Statistical methods are routinely used for climate data, such as calculating the climate normal of precipitation at a weather station, claiming the global warming based on a significant positive linear trend of the surface air temperature (SAT) anomalies, and inferring a significant shift from the lower North Pacific sea level pressure (SLP) state to a higher state. The list of questions such as the above can be infinitely long. The purpose of this chapter is to provide basic concepts and a user manual on the commonly used statistical methods in climate data analysis, so that the users can make credible conclusions with a given error probability.

R-programs will be supplied for examples in this chapter. Users can easily apply these programs and the given formulas in this book for their data analysis needs without prerequisite background of calculus and much statistics. To interpret the statistics results in a meaningful way, domain knowledge of climate science should be very useful when using the statistical concepts and calculations results to state the conclusions from specific climate datasets.

Our statistical methods have two objectives: Make credible inference about the climate state with a given error probability based on the analysis of climate data, and design the optimal climate data sampling strategies so that the observed data can form the basis of making objective and reliable conclusions. We will thus present a list of statistical indices, such as mean and variance, for climate data, and then look into the probability distributions and statistical inferences using these distributions.

3.1 A list of statistical indices for a set of data

Mean, variance, standard deviation, skewness, kurtosis, median, 5th percentile, 95th percentile, quantiles.

3.2 A set of statistical figures for climate data

Box plot, histogram, scatter plot, qq-plot, linear regression and trend.

3.3 Probability distributions

Normal distribution, t distribution; χ^2-distribution, standard normal distribution, probability density function (pdf), large samples, central limit theorem

3.4 Confidence interval

Compute the confidence interval at 95% confidence level for normal distribution and for t-distribution.

3.5 Statistical inferences

Inference for mean for normal distributions and large sample sizes (i.e., sd is given): mean; confidence interval for mean; inference for mean for normal distribution but small sample sizes (i.e., sd is not given and unknown); inference about standard deviation and χ^2-distribution.

3.6 Sample size requirement for a given error probability

Type I and Type II errors, and sample size calculation for a given confidence level or a significance level.

3.7 Uncertainties and sample designs

The problem of sample size design with given error probability.

3.8 Linear trend

Linear trend and its inference.

3.9 Reference websites

`http://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm`

http://empslocal.ex.ac.uk/people/staff/dbs202/cag/courses/MT37C/course-d.pdf

http://www.climate-service-center.de/imperia/md/content/csc/projekte/csc-report13_englisch_final-mit_umschlag.pdf

EXERCISES

3.1 The two most commonly used datasets of global average annual mean surface air temperature (SAT) anomalies are those credited to the research groups of Jim Hansen (relative to 1951-1980 climatology period) and Phil Jones (relative to 1961-1990 climatology period):

`http://cdiac.ornl.gov/trends/temp/hansen/hansen.html`

`http://cdiac.ornl.gov/trends/temp/jonescru/jones.html`

a) Find the average anomalies of each 15 years, starting at 1880. Use t-distribution to find the confidence interval of each 15-year SAT average at 95% confidence level using t-distribution. You can use either Hansen's data or Jones' data. Figure SPM.1(a) of IPCC 2013 (AR4) is a helpful reference.

b) Find the hottest and the coldest 15-year periods from 1880-2014, which has nine 15-year periods. Use t-distribution to check whether the hottest and the coldest 15-year periods are different from zero.

c) Discuss the differences between the Hansen and Jones datasets.

3.2 To test if the average of temperature in Period 1 is significantly different from that in Period 2, one can use t-statistic

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{3.1}$$

where $\bar{x}_i$ and s_i^2 are the sample mean and variance of the Period i ($i = 1, 2$). The degree of freedom (i.e., d.f.) of the relevant t-distribution is equal to the smaller $n_1 - 1$ and $n_2 - 1$. The null hypothesis is that the two averages do not have significant differences, i.e., their difference is zero (in statistical sense with a confidence interval). The alternative hypothesis is that the difference is significantly different from zero. Now you can choose to one-sided test when the difference is positive. Use a significance level of 5% or 1%, or another level at your own choice.

a) Choose two 15-year periods which have very different average anomalies. Use the higher one minus the lower one. Use this t-test method for one-sided test to check if the difference is significantly great than zero.

b) Choose two 15-year periods which have very close average anomalies. Use the higher one minus the lower one. Use this t-test method for two-sided test to check if the difference is not significantly different from zero.

CHAPTER 4

MATRICES, MATRIX ALGEBRA, AND MULTIVARIATE REGRESSION

A matrix here is an N-row and Y-column of numbers, which are called elements. Precipitation data [Units: mm/day] at multiple stations and multiple days form a matrix, normally with stations [marked by station ID] counted in rows and time [Units: day] counted in columns. The daily minimum surface air temperature (Tmin) data for the same stations and the same period of time form another matrix.

As a daily life example, the ages of the audience sitting in a movie theater of rows and columns of chairs form a matrix. Their weights form another matrix. Their bank account balance still another, and so on.

A slightly higher level of matrix concept is the coefficient matrix of a group of linear equations. Solve linear equations are very common in science and engineering. Any numerical solutions of a partial differential equation- climate model will have to solve a set of linear equations. The famous Leontief's economic supply-demand balance model is a set of linear equations that can be written in the matrix form.

A simple elementary school kid's problem reads like this: The sum of two brothers' age is 20 years and their difference is 4. What are the ages of the brothers? One can easily guess that the older brother is 12 and the younger one is 8. An eight-year old kid can most likely figure this out. The idea can be extended to a more general form of linear equations.

If we form a set of equations, which would be

$$\begin{aligned} x_1 + x_2 &= 20 \\ x_1 - x_2 &= 4 \end{aligned} \tag{4.1}$$

when x_1 and x_2 stand for the brothers ages.

The matrix form of these equations would be

$$\mathbf{Ax} = \mathbf{b} \tag{4.2}$$

which involves three matrices:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 20 \\ 4 \end{bmatrix}. \tag{4.3}$$

The single column n-row matrix is often called an n-dimensional vector.

Although one can easily guess the solution to the matrix equation is $x_1 = 12$ and $x_2 = 8$, a more consistent computing may be done by R using the following commands

```
A<-matrix(seq(1:4),2)
b<-seq(1:2)
A[1,1]=1
A[1,2]=1
A[2,1]=1
A[2,2]=-1
b[1]=20
b[2]=4
solve(A,b)
#[1] 12  8 This is the result x1=12, and x2=8.
```

This R method can solve much more complicated linear equations, such as an equation of 100 unknowns rather than 2 in this example.

This solution may be represented as

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}, \tag{4.4}$$

where A^{-1} is the inverse matrix of $\mathbf{A}$, i.e.,

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{4.5}$$

where

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{4.6}$$

is called identity matrix, which is like value 1.0 in our commonly used real number system.

This chapter will discuss three topics: space-time decomposition of a climate data matrix, matrix algebra (including matrix division (i.e., inverse matrix) and linear transform), reduce a matrix to its simplest possible form, and the matrix application in multivariate linear regression.

4.1 Matrix algebra and echelon form of a matrix

4.1.1 Matrix algebra

Addition, subtraction, multiply a matrix by a scalar constant, multiply a matrix by a matrix, a matrix divided by a matrix, and matrix inverse.

4.1.2 Independent row vectors and row echelon form

Example 1:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \tag{4.7}$$

Example 2: 5-deg global precipitation data

`http://www.ats.ucla.edu/stat/r/library/matrix_alg.htm`

4.2 Space-time representation of climate data

SVD (singular value decomposition)

`https://stat.ethz.ch/R-manual/R-devel/library/base/html/svd.html`

4.3 Covariance, EOFs, and PCs

Climate data interpretations of SVD results

4.4 Multivariate linear regression using matrix notations

Formulas and R programs

EXERCISES

4.1 SVD space-time decompositions and data reconstruction.

SVD decomposition: Use R and SVD method to decompose the 5-degree reconstructed annual precipitation data (PrcpRecon.csv) into space, variance, and time matrices: u, d, and v.

EOF-PC reconstruction: Use the SVD's u, d, and v matrices to reconstruct the original data matrix, $m = u \cdot d \cdot v^T$, where v^T is the transpose matrix of v. Here, u's column vectors are EOFs which represent spatial variations; the diagonal matrix's elements are eigenvalues representing variances; and v's row vectors are principal components (PCs) which represent the temporal variation.

a) Choose a case of 4 grid boxes and 2 years.
b) Choose a case of 4 grid boxes and 3 years.
c) Discuss the above results, using plain text or figures or both.

4.2 A covariance matrix C can be computed from a space-time observed anomaly data matrix X which has N rows for spatial grid boxes and Y columns for time in years:

$$C = X \cdot X^T / Y \tag{4.8}$$

This is an $N \times N$ matrix.

a) Choose a Y matrix from the reconstructed 5-degree annual precipitation data and calculate a covariance matrix for $N = 5$ and $Y = 6$.
b) Use R to find the inverse matrix of the covariance matrix C.

CHAPTER 5

ENERGY BALANCE MODELS FOR CLIMATE AND DIFFERENTIAL EQUATIONS

An equation is any formulation of a balance using one equal to other. A differential equation is an equation involving derivatives.

The solution of $2x = 3$ is $x = 3/2 = 1.5$ is a number. The geometric meaning of this equation is that it is the x-coordinate of the intersection point of the slant straight line $y = 2x$ and the horizontal line $y = 3$. One can plot the lines and find the intersection, and hence can find the solution geometrically.

The linear equations of more than one variable is often represented in a matrix formula discussed in the last chapter:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{5.1}$$

The solution of this equation in matrix form is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{5.2}$$

where $\mathbf{A}^{-1}$ is the inverse matrix of $\mathbf{A}$. Of course, there is a condition for this solution to exist, which is that $\mathbf{A}^{-1}$ exists, i.e., the coefficient matrix $\mathbf{A}$ is invertible. For

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 20 \\ 4 \end{bmatrix}. \tag{5.3}$$

, the solution is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{A}^{-1}\mathbf{b} = (1/2)\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 20 \\ 4 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix}, \tag{5.4}$$

i.e., $x_1 = 12$ and $x_2 = 8$.

The geometric meaning of this solution is that the solution is the coordinates of the intersection point of two straight lines on the (x_1, x_2) plane. One is defined by

$$x_1 + x_2 = 20, \tag{5.5}$$

another by

$$x_1 - x_2 = 4. \tag{5.6}$$

One can thus also find the solution geometrically by plotting the lines and finding the intersection point.

A quadratic equation, such as $x^2 - 1 = 0$, normally has two solutions, which are x-coordinates of the intersection points of $y = x^2 - 1$ with x-axis. This particular function $y = x^2 - 1$ represents an up-open parabola and has two intersection points with x-axis: $x = -1$ and $x = 1$.

The algebraic solution of a general quadratic equation

$$ax^2 + bx + c = 0 \tag{5.7}$$

is

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}. \tag{5.8}$$

Of course, the general parabola $y = ax^2 + bx + c$ may intersect with x-axis at two points (i.e., two distinct solutions), one point (two overlapped solution, or called double root, or repeated root), and no point when the entire parabola is above or below the x-axis.

A differential equation is an equation that involves at least one derivative. The simplest one is

$$\frac{dy}{dx} = 0. \tag{5.9}$$

What is the solution of this differential equation? It must be a relation between x and y since $\frac{dy}{dx}$ involves x and y. Thus, a solution must be a function (i.e., an expression for the relationship between x and y). What kind of $y - x$ function can make this equation satisfied? One can guess that this particular simple differential equation, denoted by DE hereafter, is that y is a constant. It can be any constant since a constant's derivative (i.e. slope) is always zero. So the general solution is

$$y = C, \tag{5.10}$$

where C denotes a constant. To fix the solution, we need another condition, called initial condition. Say $y = 2$ when $x = 0$ (the initial state), then the fixed unique solution is $y = 2$, the state does not change (because the change rate is always zero).

Another simple example of DE is $\frac{dy}{dx} = x$, whose general solution is $y = x^2/2 + C$.

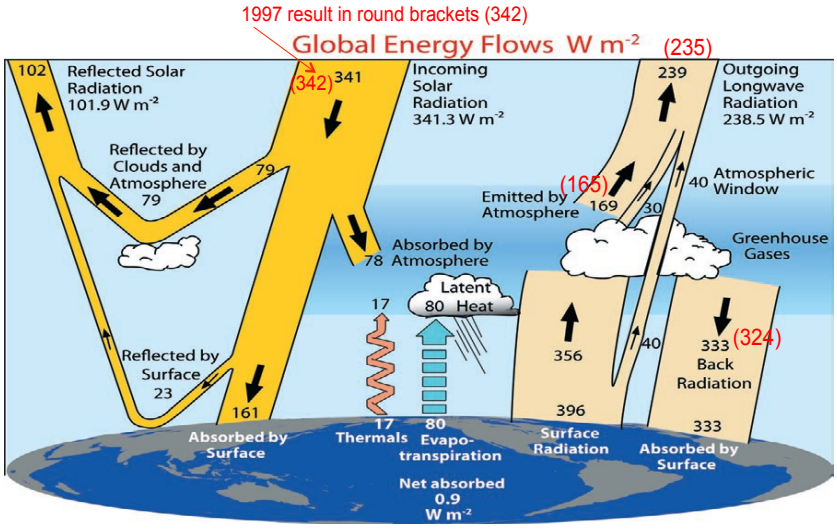
This chapter will use the ideas of equations, particularly the DE concepts, to model the Earth's air temperature. The unknown variable y in the model equation will be temperature.

5.1 Zero-dimensional Energy Balance Model for Earth's Constant Temperature Climate

5.1.1 Earth's energy budget

The energy balance of the incoming solar energy and the outgoing energy to the outer space through radiation and reflection forms an equation, which is called an energy balance

model (EBM) for climate. This is the simplest, yet very important, climate model for the study of long-term climate change. After all, the warming of the Earth surface is powered by the solar energy and regulated by ocean water and other Earth surface materials, such as ice, air, and plants. The updated observation by satellite shows a net incoming energy of 0.9 watts per square meter (See Figure 5.1), which is equal to the incoming solar energy of minus the outgoing total energy and can be a cause of surface air warming still regulated by the water body in the vast ocean.



Carbon dioxide forcing is about 2 w/sq m. The net heating is 0.9 w/sq m.

Balance of the global energy budget: Kevin Trenberth et al. (*Bull. Amer. Meteo. Soc.*, 2009), an update from his 1997 results marked in red.

Figure 5.1 Energy balance of the Earth's surface: incoming solar radiation and outgoing energy via radiation and reflection (Trenberth et al. (2009)).

This small amount of net incoming energy (0.9 watts/square meter) kept on the surface , including the entire atmosphere and ocean, has a large uncertainty, since the energies emitted by atmosphere to the outer space (mainly via clouds) and radiated by to Earth's surface (also via clouds) are hard to measure due to the complexity of clouds. The 2009 value of cloud radiation back to the Earth surface was 333 w/sq. m, while the 1997's value was 324 w/sq. m; the difference is 9 w/sq. m. The energy emitted by atmosphere was 169 w/sq. m. in 2009 publication, while the 1997 value was 165 w/sq. m; the difference is 4 w/sq. m. These large uncertainties in clouds' influence on energy leads to a question of whether the 0.9 w/sq. m is significantly different from zero. One can make an inference on a null hypothesis and an alternative hypothesis when the uncertainties are quantified. Another way to ask the question is: what is the confidence interval (CI) of the mean 0.9 w/sq. m at 95% confidence level? If the CI includes zero, then the net incoming energy 0.9 w/sq. m is not significantly difference from zero. Thus, the global warming in the last 150 years cannot be solely based on the assessment of a small positive net incoming energy. Climate dynamics must be involved, taking into account of ocean water, land processes,

major atmospheric and oceanic circulation patterns including MJO, Walker Circulation, Hadley Circulation, AO, QBO, ENSO, PDO, AMO, global monsoon systems, and polar ice dynamics and Tibetan Plateau's orographic effects. Including all these effects would form a general circulation model (GCM) for climate, which is the kind of climate models developed and run by major climate research centers. The high resolution GCMs must be run on supercomputers, such as those of 4-by-4 km resolution models. Most GCMs still have their resolution larger than 100 km.

GCM in climate community also refers to a Global Climate Model, which often has general circulations.

The EBM discussed here ignores these circulations and considers only the energy balance: the energy comes to the Earth system and the energy goes out from the system.

5.1.2 A uniform snowball Earth

Since sun is far away from Earth, the Earth's one side receives sun's radiation as straight line rays shown in Figure 5.2. The power of solar rays is called solar constant, denoted by S, which is about 1,365 w/sq.m (at the lower activity phase of sun spots)

`http://science.nasa.gov/science-news/science-at-nasa/2003/17jan_solcon/`

and varies with time around this value both randomly and periodically because of solar activities, such as the 11-year cycle of sun's dark spots.

The entire Earth receives solar radiance on one side at a given moment. The total energy flux is equivalent to that going through a round disk of the Earth's radius R (about 6,400 km). The round disk's area is πR^2. The energy is distributed to the entire Earth surface whose area is $4\pi R^2$. Thus, the per unit square's solar irradiance received by the Earth's surface is

$$S_{solar} = S\frac{\pi R^2}{4\pi R^2} = S/4 = 1,365/4 = 341.25[wm^{-2}]. \tag{5.11}$$

Some solar irradiance is reflected back to the outer space and is determined by Earth's reflectivity, α, which is approximately 0.32 for our current Earth surface conditions. Thus, the solar energy received by Earth is

$$E_{in} = (1-\alpha)(S/4). \tag{5.12}$$

The solar radiance received by Earth can have large variations due to the conditions of Earth's surface and clouds (through the variation of α value) and due to the solar activities like the sun spots (through the variations of S values), and can be as large as a few percent, possibly up to 7% claimed by some.

The other part is radiation given out by the Earth. It is universal that every body radiates energy, so does the Earth. It radiates like our humanbody via infrared waves, which are long compared to the incoming solar irradiance waves that can perpetrates transparent glasses and plastic membrane, while the infrared cannot. This mechanism makes the greenhouse work, and the phenomenon is called the greenhouse effect. The small amount of transparent carbon dioxide (CO2) can produce the greenhouse effect: short wave comes in and long waves are trapped. The CO2 amount in air is measured by the famous Keeling curve, dedicated to David Keeling, a late SIO professor who started to observe the CO2 data from 1958. Today (August 20, 2015)'s CO2 reading is 397.93 ppm (parts per million). This number was around 320 ppm in 1958.

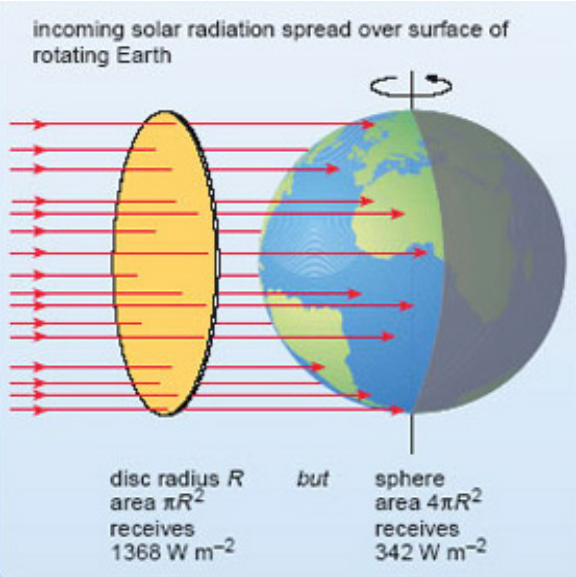


Figure 5.2 Earth's one side receives sun's radiation as straight line rays .

The long wave radiation to the outer space by Earth follows the Stefan-Boltzmann blackbody law

$$E_{bb} = \epsilon\sigma T^4, \tag{5.13}$$

where

$$\sigma = 5.670373 \times 10^{-8}[Wm^{-2}K^{-4}] \tag{5.14}$$

is called Stefan-Boltzmann constant, and $0 < \epsilon \leq 1$ is the dimensionless emissivity of Earth surface. In the Stefan-Boltzmann law, temperature is in the unit of Kelvin degrees, which is 273+ Celsius degrees. We first take ϵ to be one.

If the Earth temperature does not change and the incoming energy is equal to the outgoing energy, then we have an equation of energy balance $E_{in} = E_{bb}$, i.e.,

$$\epsilon\sigma T^4 = (1 - \alpha)(S/4). \tag{5.15}$$

We can easily solve this algebraic equation

$$T^4 = \frac{(1-\alpha)(S/4)}{\epsilon\sigma} = \frac{(1-0.32)(1365/4)}{1.0 \times 5.670373 \times 10^{-8}} = 40.92 \times 10^8[^\circ K]^4 \tag{5.16}$$

Then the temperature is

$$T^4 = 40.92[^\circ K]^4, \qquad T = 253[^\circ K], \qquad T = -20[^\circ C] \tag{5.17}$$

As the Earth surface air temperature, this T value is too low. Even it is too low when T is regarded as the average of a column of air weighted by the air concentration. This $T = -20[^\circ C]$ is a uniform temperature snowball Earth, everywhere having a below frozen temperature.

To make it a uniform Earth surface of other types different from snow or ice, we can tune the model by tuning the parameters in the above EBM solution

$$T = \left[\frac{(1-\alpha)(S/4)}{\epsilon\sigma}\right]^{1/4} - 273 \quad [^\circ C] \tag{5.18}$$

One way is to tune down the Earth emissivity ϵ from 1.0 to 0.6, justified by the greenhouse effect of heat trapping. Then,

$$T = \left[\frac{(1-0.32)(1365/4)}{0.6 \times 5.670373 \times 10^{-8}}\right]^{1/4} - 273 = 14 \quad [^\circ C] \tag{5.19}$$

Although this number seems reasonable and is the global average SAT of the current Earth, it is impossible to demonstrate by observation to support the emissivity parameter's value $\epsilon = 0.6$. In addition, this is the uniform temperate Earth, perhaps the the wet-towel-wrapped Earth, which has no difference between equator and poles. Not a reality. Nonetheless, this gives us a knob to turn for tuning a climate model, i.e., how the heat is trapped due to the non-perfect emissivity.

Yet, another problem emerges: when the Earth surface is wet-towel wrapped, its reflectivity α is different from that of snowball Earth. Thus, the Earth is a nonlinear system. All the parameters are related in a nonlinear way.

5.1.3 EBM for a uniform Earth with nonlinear albedo feedback

To make the model a one step close to reality, we make reflectivity depend on the surface temperature T. One model is below

$$\alpha(T) = 0.5 - 0.2 \times \tanh((T-265)/10), \tag{5.20}$$

which means $\alpha = 0.3$ when it is ice-covered for low temperature and has a high albedo (α is called co-albedo), and 0.7 when it is ice-free for a high temperature and a low albedo (see Figure 5.3). Then, the new EBM is

$$\epsilon\sigma T^4 = (1 - (0.5 - 0.2 \times \tanh((T-265)/10))(S/4). \tag{5.21}$$

Here 265°K is regarded as the ice formation temperature, and $\tanh$ is the hyperbolic tangent function, a smooth step function which 1.0 at infinity and -1.0 at negative infinity. Solving this highly nonlinear equation for T by hand is impossible. Computer can solve it or one can solve it graphically as shown in Figure 5.3.

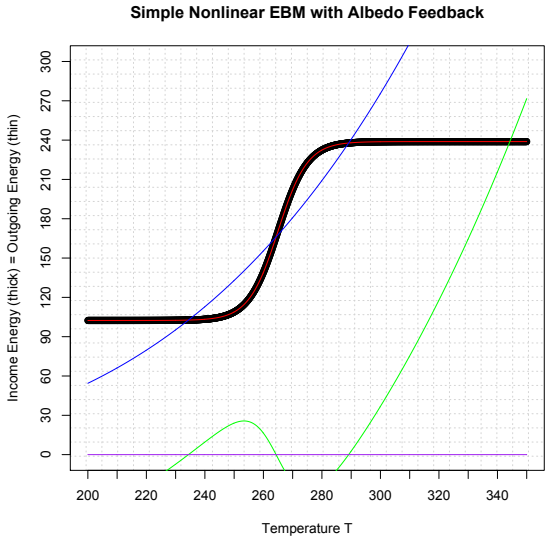


Figure 5.3 Reflectivity α as a nonlinear smooth step function of temperature T, and the graphic solution of an albedo-feedback EBM.

Figure 5.3 can be produced using the following parameter values and R commands.

```
S<-1365
ep<-0.6
sg<-5.670373*10^(-8)
T<-seq(200,350, by=0.1)
y1<-(1-(0.5 - 0.2 * tanh ((T-265)/10)))*(S/4)
y2<- ep*sg*T^4
plot(T, y1, xlim=c(200, 350), ylim=c(0,300),
xaxp=c(200, 350, 15), yaxp=c(0, 300, 10),
panel.first = grid(30, lty = "dotted", lwd = 1),
main="Simple Nonlinear EBM with Albedo Feedback",
ylab="Income Energy (thick) = Outgoing Energy (thin)",
xlab="Temperature T")
lines(T, y1,col="red")
lines(T, y2,col="blue")
lines(T, y2-y1,col="green")
y3<-0.0*T
lines(T, y3,col="purple")
# The 3 intersections of the green and purple lines
# are three solutions.
```

The three solutions of the albedo feedback nonlinear EBM's solution are $234, 264, 289°$K. The third solution is $16°$C, close the current Earth. The first solution is $-39°$C and is a deeply frozen all ice Earth, corresponding to a global scale ice age. The middle solution will be proved to be unstable in the next section.

These three numerical solutions were found by the following R commands.

```
S<-1365
ep<-0.6
sg<-5.670373*10^(-8)
f <- function(T){return(ep*sg*T^4 -
(1-(0.5 - 0.2 * tanh ((T-265)/10)))*(S/4))}
uniroot(f,c(220,240))
uniroot(f,c(260,275))
uniroot(f,c(275,295))
```

5.2 Zero-dimensional EBMfor Earth's Time Varying Temperature

The Earth's temperature varies with respect to time. The temperature change times the heat capacity is the power needed to drive the change. This relation yields the time dependent EBM

$$C\frac{dT}{dt} = (1 - (0.5 - 0.2 \times \tanh((T - 265)/10))(S/4) - \epsilon\sigma T^4, \qquad (5.22)$$

where C is the heat capacity of the Earth surface, including all air, ice, soil and water, particularly the ocean water. An estimated value of C for the current Earth is $C = 2.08 \times 10^8 [JK^{-1}m^{-2}]$.

This EBM allows temperature T to change with respect to time and involves a derivative. Any equation involves derivatives is called a differential equation. Thus, this EBM is a differential equation model. The derivative $\frac{dT}{dt}$ is slope, which can also be interpreted as the rate of temperature change. The rate of change is proportional to the driving power, the incoming short wave solar power minus the outgoing long wave power through blackbody radiation. The rate is inversely proportional to the heat capacity, since it is easier to heat air than water because water has much larger heat capacity than air.

The solution of a differential equation is a function. For example, the solution of $\frac{dy}{dt} = t-1$ is $y = t^2/2-t+D$ where D is an arbitrary constant. An initial condition $t = 0, y = 3$ can determine the constant D which is 3. Thus, the initial value problem of the differential equation is $y = t^2/2 - t + 3$.

Solving a differential equation by hand can be very difficult. Solving the above EBM by hand is impossible. Many computer program packages, such as R's deSolve package, can solve DEs for us. We will discuss the DE solutions later in this book. The remaining of this sub-section discusses the stability of the three solution found the previous sub-section.

When the right hand side of the above EBM is equal to zero, the derivative $\frac{dT}{dt}$ is zero. The model is at an equilibrium state, the incoming power equal to the outgoing power. If an equilibrium state can remain at this state or its vicinity after a small power unbalance, called small perturbation in mathematical terminology, then the equilibrium is stable. The climate can stay at this equilibrium state and its vicinity for a period of time. Otherwise, the equilibrium is unstable and the equilibrium is only a mathematical solution, and can only be a transient state of climate. The climate can never stay at an unstable equilibrium, which is thus not observable.

We can argue that the largest solution 289°K among the three in the previous subsection is stable. For a small positive perturbation, $T > 289$, $C\frac{dT}{dt} < 0$ since the red line (incoming power) is below the blue (outgoing power). The negative derivative means T goes down, so goes back to 289. For a small negative perturbation, $T < 289$, $C\frac{dT}{dt} > 0$ means T goes up, so also goes back to 289. Therefore, a small perturbation around 289 makes the climate system go back to 289, hence the system is stable.

One can use the similar method to argue that the smallest solution 234°K is also stable, but the middle solution 264°K is unstable.

5.3 Raise the complexity of climate models

The EBMs we have discussed are very simple, but the time-varying albedo feedback EBM is already complex enough to prevent a hand calculation. Most climate models today are very complex and take into account of various kinds of factors and parameters, such as Earth tilt effect, polar ice cap, ocean-land differences, land use land cover, and clouds. The inclusion of these factors often requires high spatiotemporal resolution, which can currently reach 1 km space resolution and 1 sec of time resolution for global models and reach 100 m resolution by regional models.

Despite the popular use of complex GCMs of high resolution, the EBM of moderate complexity can provide some useful and back-in-envelope estimate that can guide scientists to look into the main issues and practitioners for some strategic decisions. In addition to the few EBMs discussed earlier, the next level of complexity is to have a latitude-band dependence of the temperature, higher temperature at equator and lower at poles. This introduces the latitude dependence of the temperature, which is now a function of both time and latitude $T(t, \phi)$ where ϕ is latitude. The EBM for this $T(t, \phi)$ will involve derivatives with respect to both t and ϕ, which are partial derivatives. The resulted EBM will be partial differential equation (PDE). In contrast, when a DE involves derivative with respect to only one variable, say t, the derivative is called ordinary derivative and the equation is called the ordinary differential equation (ODE), sometimes simply written as DE.

A further complexity is the inclusion of the Earth's tilt, which produces seasonality.

Still further complexity is the consideration of land and ocean differences, where longitude must be considered. The temperature now depends on three variable: time, latitude and longitude $T(t, \phi, \theta)$ where θ denotes longitude.

5.4 ODE basics

ODE examples by R.

EXERCISES

5.1 Tune the snowball uniform Earth EBM parameters to find three types of Earth climate conditions. Discuss the numerical results generated by R.

5.2 Repeat the above problem for the albedo-feedback nonlinear EBM.

5.3 A simplified albedo-feedback nonlinear EBM can be solved by hand. This model assumes the following:

(a) $\alpha = 0.7$ when $T < -8°$C, and $\alpha = 0.3$ when $T > -8°$C. Regard $-273.15°$C as the absolute zero temperature. Thus T[°K]=273.15+ T[°C]. The incoming solar radiation power is

$$E_{in} = (1 - \alpha)\frac{S}{4}. \tag{5.23}$$

(b) The outgoing Earth radiation is approximated by a Budyko formua:

$$E_{out} = 189.4 + 2.77T[Wm^{-2}] \tag{5.24}$$

where T's units is [°C].

Find three equilibrium solutions for T of the EBM $E_{out} = E_{in}$ based on these two assumptions. Plot relevant figures about this EBM. Find your solutions by hand and calculators only. No R please.

5.4 Use R to develop a table to document the relative differences between the Earth radiation computed by the Stefan-Boltzmann law and that by Budyko's linear approximation formula, when T is in (-60,60)°C. Use the following parameters: $\epsilon = 0.6, S = 1,365$. Discuss and comment on your numerical results. Think about the real Earth whose radiation is large over the equator area and small over the polar regions. What is the meaning of our assumption of T being in (-60,60)°C? Plot some figures to help you describe your numerical results and your ideas.

CHAPTER 6

CLIMATE SCIENCE TOPICS OF CALCULUS I: DERIVATIVES

Calculus method of cut and add is used every modeling of science and engineering. Climate science is not an exception. This chapter some commonly used calculus methods in climate science, such as linear approximation, Newton's method, linearization of the Stefan-Boltzmann blackbody radiation law, Taylor expansion, partial derivatives, multiple integrals, and line integrals. Each calculus method is introduced by a climate science example.

6.1 Stefan-Boltzmann law and Budyko's approximation

For a perfect blackbody, the Stefan-Boltzmann law of radiation is

$$E_{pbb} = \sigma T^4 \tag{6.1}$$

where temperature T is in degree Kelvin.

Earth is not a perfect blackbody. Its radiation to the outer space is reduced by a factor, called emissivity ϵ. Thus, the Earth's long wave radiation is

$$E_{bb} = \epsilon \sigma T^4. \tag{6.2}$$

In the previous discussion on EBM, two values of emissivity ϵ were used: $\epsilon = 1.0$ and $\epsilon = 0.6$. Our simple EBM, often called toy model in the climate research community due its extreme simplicity, yields a reasonable surface temperature 14 $\circ$C for $\epsilon = 0.6$, and

snowball or ice ball Earth for $\epsilon = 1.0$. Perhaps Earth has experienced both cases: $\epsilon = 0.6$ and 1.0 (or close to 1.0). So what is the right value for ϵ? Can we use observed data to find out the ϵ value for the current Earth?

Mikhail I. Budyko (1961) used a linear regression to model the Earth's outgoing radiation. From the statistics point of view, he detoured from the Stefan-Boltzmann law and directly relied on linear regression

$$E_{Bud} = A + BT \tag{6.3}$$

where temperature T's units is °C, not °K as in the Stafan-Boltzmann law, and the units for the regression coefficients are $A[Wm^{-2}]$ and $A[Wm^{-2}(^{\circ}C)^{-1}]$.

According to North and Coakley (1979), the best fit based on the northern hemisphere yields

$$A = 315.6[Wm^{-2}], \qquad B = 4.62[Wm^{-2}(^{\circ}C)^{-1}]. \tag{6.4}$$

Figure 6.1 show the curves of outgoing long wave radiation energy when using $\epsilon = 0.6$ for the Stefan-Boltzmann's radiation formula and using $A = 315.6[Wm^{-2}], B = 4.62[Wm^{-2}(^{\circ}C)^{-1}]$ for Budyko's radiation formula, both in the units of [° C]. The two radiation formulas are below

$$E_{SB} = 0.6 \times (5.670373 * 10^{-8}) \times (273.15 + T)^4, \tag{6.5}$$

$$E_{Bud} = 315.6 + 4.62T \tag{6.6}$$

in the nearly possible Earth's surface temperature range $[-50, 50]$°C. Here 273.15 comes from the conversion of the units from [° K] to [° C], and -273.15[° C] is the absolute zero of [° K].

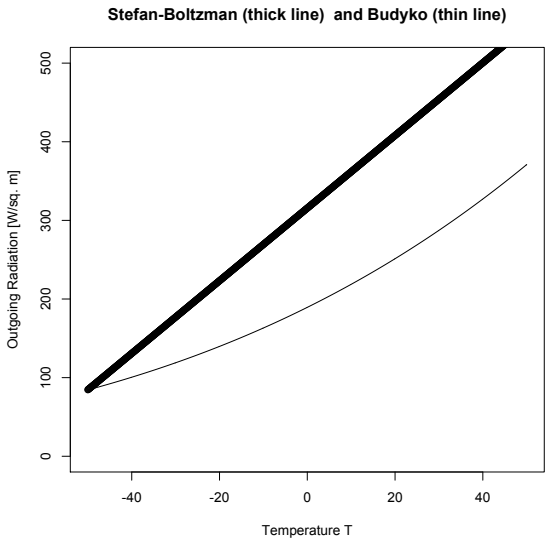


Figure 6.1 Comparison of theEarth's outgoing radiation according to Stefan-Boltzmann law and that according to Budyko's formula.

Figure 6.1 is produced by the following R program

```
A<-315.6
B<-4.62
ep<-0.6
sg<-5.670373*10^(-8)
T<-seq(-50,50, by=0.1)
plot(T,A+B*T, xlim=c(-50,50), ylim=c(0,500),
main="Stefan-Boltzman (thin line) and Budyko (thick line)",
xlab="Temperature T", ylab="Outgoing Radiation [W/sq. m]")
lines(T, ep*sg*(273.15+T)^4)
```

The figure indicates near-straight line approximation for Stefan-Boltzman law, although the figure shows that there is a difference between Budyko and Stefan-Boltzmann formulas. The near-straight line property of the Stefan-Boltzmann law in this temperature range $[-50, 50]$°C allows to fit the regression better using improved data. For example, $A = 189.4[Wm^{-2}], B = 2.77[Wm^{-2}(^{\circ}C)^{-1}]$ yields a very agreement between Budyko's radiation and the Stefan-Boltzmann radiation (see Figure 6.2).

This figure is produced by R

```
A<-189.4
B<-2.77
ep<-0.6
sg<-5.670373*10^(-8)
T<-seq(-50,50, by=0.1)
plot(T, 189.4 + 2.77*T, xlim=c(-50,50), ylim=c(0,500),
type="o",col="blue",
main="Stefan-Boltzman (thin line)  and Budyko (thick line)",
xlab="Temperature T", ylab="Outgoing Radiation [W/sq. m]")
 lines(T, ep*sg*(273.15+T)^4)
```

With this R program, one can tune the two "knobs": A and B. We can find that $A = 189.4[Wm^{-2}], B = 2.77[Wm^{-2}(^{\circ}C)^{-1}]$ looks like the best fit. Why? It is the best linear approximation, or just linear approximation. In general, to approximate a curve by a tangent straight line is called linear approximation.

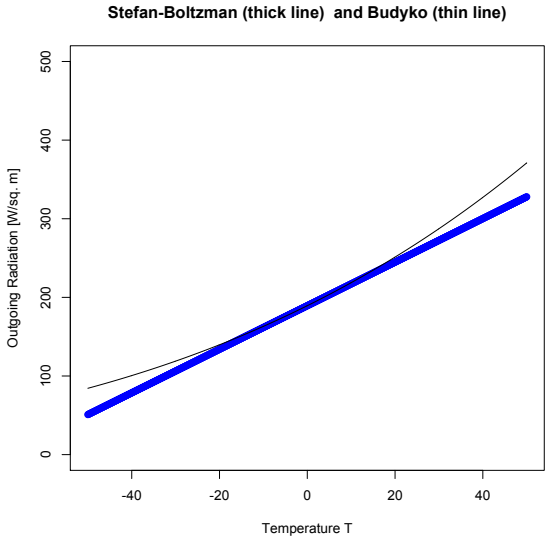


Figure 6.2 The best agreement between Stefan-Boltzmann law and Budyko's formula.

6.2 Linear approximation

It is clear that the best linear approximation of the curve $y = f(x)$ at a point $x = a$ is the tangent line at $(a, f(a))$. The point-slope equation of the tangent line is

$$y - f(a) = f'(a)(x - a), \tag{6.7}$$

where $f'(a)$ is the derivative of the function $f(x)$ at $x = a$. The linear approximation is

$$y = L(x) = f(a) + f'(a)(x - a) \tag{6.8}$$

EXAMPLE 6.1

Find the linear approximation of

$$f(x) = (1 + x)^4 \tag{6.9}$$

near $x = 0$. Solution: One can use WolframAlpha to find the derivative $f'(x) = 4(1 + x)^3$. When $x = 0$, $f'(0) = (1 + 0)^3 = 1$. $f(0) = (1 + 0)^4 = 1$. Thus the linear approximation is

$$y - f(0) = f'(0)(x - 0) \to y - 1 = 4 \times (x - 0) \to y = 4x + 1. \tag{6.10}$$

This example can be applied to the Stefan-Boltzmann law in the following way.

$$\begin{aligned} E_{SB} &= 0.6 \times (5.670373 \times 10^{-8}) \times (273.15 + T)^4 \\ &= 0.6 \times (5.670373 \times 10^{-8}) \times 273.15^4 \left(1 + \frac{T}{273.15}\right)^4. \end{aligned} \tag{6.11}$$

Table 6.1 Linear approximation of $f(x) = (1+x)^4$ by $L(x) = 1 + 4x$

x	f(x)	L(x)	Error[%]
-0.3	0.2401	-0.2	183
-0.2	0.4096	0.2	51
-0.1	0.6561	0.6	9
0.0	1.0000	1.0	0
0.1	1.4641	1.4	4
0.2	2.0736	1.8	13
0.3	2.8561	2.2	23

The example $(1+x)^4$ applies to this formula with $x = T/273.15$ which is between -0.2 and 0.2 and can be considered small.

Table 6.1 shows that the error is relatively small, less than 25%, when x is in $[-0.15, 0.30]$. When we apply this to the factor $x = T/273.15$, this x interval is transferred into the T interval $[-41, 82]^\circ$ C. This range includes almost all the temperature regimes over the Earth, except some parts of Antarctic. Therefore, Budyko's simple linear long wave radiation model is applicable to almost anyplace on Earth. The Budyko's radiation energy results are not sensitive to the two parameters, around $A = 189.4[Wm^{-2}], B = 2.77[Wm^{-2}(^\circ C)^{-1}]$. Hence, the model and its results are robust.

Table 6.1 can be produced by the following R program

```
x <- seq(-0.3,0.3, by=0.1)
fx <- c(1:7)
lx <- c(1:7)
ex <- c(1:7)
for(i in 1:7){fx[i]=(1+x[i])^4
                    lx[i]=1+4*x[i]
      ex[i]=((1+x[i])^4-(1+4*x[i]))/((1+x[i])^4)*100
             }
```

6.2.1 Bisection method for solving nonlinear equations

Some simple equations can be easily solved. For example, $2x = 1$'s solution is $x = 1/2$, and $x^2 + 2x + 1 = 0$ has repeated root $x_1 = x_2 = -1$ which can be found by quadratic root formula or factorization. However, most equations in nature cannot be solved by hand, such as the time-independent EBM in Sub-section 5.1.3.

$$\epsilon\sigma T^4 = (1 - (0.5 - 0.2 \times \tanh((T - 265)/10))(S/4). \tag{6.12}$$

R program solved this equation and yielded three solutions: $234, 264, 289^\circ$K. These are numerical solutions, which were found by R in Sub-section 5.1.3 using `uniroot(f,c(220,240))` for 234. So most equations in climate science computers have to be used to solve them, because either they are nonlinear, or they are too many, or both. Global climate models need to solve equations of thousand variables, and require numerical solutions of a large dimensional matrix equation. Most of the numerical methods for solving a nonlinear equation is based on the linear approximation, called Newton's method.

Before describing Newton's method, let us examine an a very simple method, the method of bisection. This brutal-force method is to find a root between a negative point and a positive point. Suppose $f(x) = 0$ is to equation to be solved for x. We assume that $y = f(x)$ is a continuous function, whose curve is continuous on $xy-$plan. If $x = c$ is a root, then $f(c) = 0$. A frequent case is that $f(x)$ is negative on one side of c and positive on the other side. That is, if $f(c) < 0$ and $f(b)?0$, then there is a root a between $x = a$ and $x = b$ under an assumption that the function $f(x)$ is continuos, i.e., the curve $y = f(x)$ does not have jumps, holes, and extremely fast oscillation. The bi-section method is a procedure to search for a. The algorithm is below.

Make an initial guess that a in $[a, b]$ and compute $f(a)$ and $f(b)$. If $f(a)$ and $f(b)$ have different signs, then take this pair a and b, else make another initial guess. The first successful guess is called a_1 and b_1. The bisection of this pair is $(a_1 + b_1)/2$. Compute $f((a_1 + b_1)/2)$. From the signs of $f(a_1), f(b_1)$ and $f((a_1 + b_1)/2)$, one can determine which half of the bisected interval $[a_1, (a_1 + b_1)/2]$ or $[(a_1 + b_1)/2, b_1]$. For the chosen half, we have the root a in $[a_2, b_2]$. Similarly, one can find $[a_3, b_3]$. Since the interval $[a, b]$ gets bisected each time, the total length of the remaining half is approaching zero, and the two end points $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ approach the root c.

This method is very simple and easy to understand, but it takes many steps to find a solution. We say that the convergence rate is very slow.

Another drawback is that it cannot find a double root, both sides of which $f(x)$ have the same sign.

EXAMPLE 6.2

Find a root of $f(x) = (1 + x)^4 - (2 + x) = 0$ between $x = 0$ and $x = 1$ using the bisection method.

6.3 Newton's method

Newton's method is a much fast approach and uses the tangent line linear approximation. To find a root of $f(x) = 0$, we make an initial guess x_0, then compute $f(x_0)$. From the point $(x_0, f(x_0)$, we draw a tangent line

$$y - f(x_0) = f'(x_0)(x - x_0) \tag{6.13}$$

and compute the intersection point of this line and the x-axis, i.e., finding the root for this linear equation, which is very easy. The solution of this linear equation as $y = 0$ is x_1

$$x_1 = x_0 - f(x_0)/f'(x_0). \tag{6.14}$$

This x_1 is the second, but calculated guess. From here, one can calculate the next guess, x_2, and so on. Therefore, we have a nice iteration algorithm for root-finding:

$$x_{n+1} = x_n - f(x_n)/f'(x_n), \qquad n = 0, 1, 2, 3, \cdots \tag{6.15}$$

An R- program for Newton's method is below

```
newton <- function(f, tol=1E-12,x0=1,N=20) {
        h <- 0.001
```

```
        i <- 1; x1 <- x0
        p <- numeric(N)
        while (i<=N) {
                df.dx <- (f(x0+h)-f(x0))/h
                x1 <- (x0 - (f(x0)/df.dx))
                p[i] <- x1
                i <- i + 1
                if (abs(x1-x0) < tol) break
                x0 <- x1
        }
        return(p[1:(i-1)])
}
```

To use this function, called `newton`, we need to specify the function, the error tolerance, the initial guess, and the maximum number of steps allowed. For example, to find the roots for $x^3 + 4x^2 - 10 = 0$ near 1.0, we can write the following code

```
f <- function(x) { x^3 + 4*x^2 -10 }
root <- newton(f, tol=1E-12, x0=1, N=10)
root
[1] 1.454256 1.368917 1.365238 1.365230 1.365230 1.365230 1.365230
```

One can easily verify this solution:`f(1.365230`, which is $-2.215123e-07$, very close to zero. Only 7 iterations are used. See the figure below for the above method.

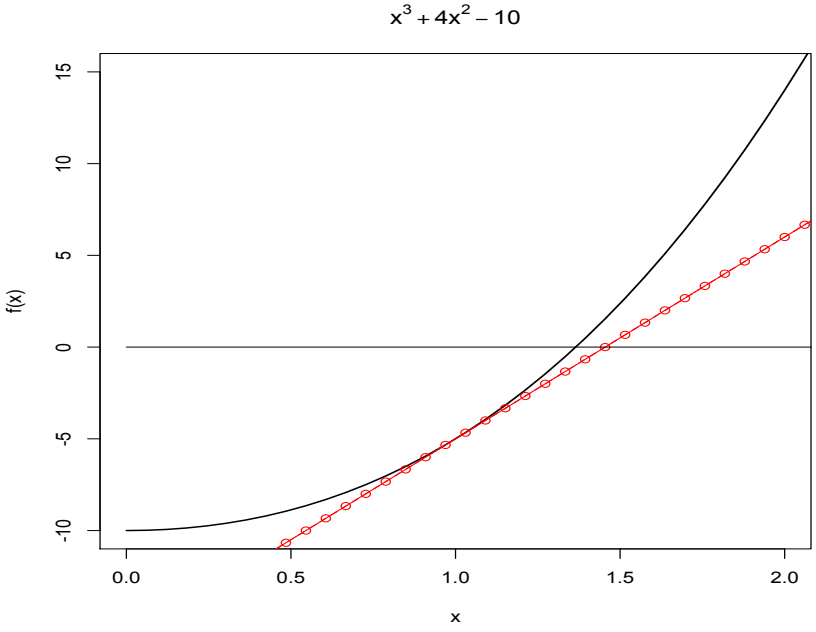


Figure 6.3 Illustration of Newton's method for finding a root from the initial guess $x_0 = 1.0$.

This figure can be produced by the following R commands when in the same R session as above Newton's procedure.

```
plot(x,f(x), type='l', lwd=1.5,
main= expression(x^3 + 4*x^2 -10),
xlim=c(0,2),ylim=c(-10,15))
lines(x,-5+11*(x-1),type="l", lty=2, col="red")
lines(x,0*x)
```

One can now try to use the Newton's method to find a root of $f(x) = (1+x)^4 - (2+x) = 0$ between $x = 0$ and $x = 1$ by starting at $x_0 = 1$. The result is $x = 0.2207441$.

```
function(x) { (x+1)^4 -(2+x) }
root <- newton(f, tol=1E-12, x0=1, N=10)
root
[1] 0.5809697 0.3335992 0.2359959 0.2210877 0.2207447
0.2207441 0.2207441 0.2207441
[9] 0.2207441
```

We can use Newton's method to find equilibria of the albedo-feedback EBM solutions for temperature T.

```
S<-1365
ep<-0.6
sg<-5.670373*10^(-8)
f <- function(T){return(ep*sg*T^4 -
(1-(0.5 - 0.2 * tanh ((T-265)/10)))*(S/4))}
newton <- function(f, tol=1E-12,x0=1,N=20) {
        h <- 0.001
        i <- 1; x1 <- x0
        p <- numeric(N)
        while (i<=N) {
                df.dx <- (f(x0+h)-f(x0))/h
                x1 <- (x0 - (f(x0)/df.dx))
                p[i] <- x1
                i <- i + 1
                if (abs(x1-x0) < tol) break
                x0 <- x1
        }
        return(p[1:(i-1)])
}
```

To find the three equilibria solutions, we place different initial guesses, each of which will converge to a solution. Our numerical experiments are below:

```
root1 <- newton(f, tol=1E-12, x0=220,N=20)
root1
#  [1] 235.6965 234.3860 234.3817 234.3817 234.3817
root2 <- newton(f, tol=1E-12, x0=270,N=20)
root2
[1] 262.0567 264.5071 264.3378 264.3377 264.3377 264.3377
root3 <- newton(f, tol=1E-12, x0=300,N=20)
 root3
[1] 289.9086 289.1469 289.1401 289.1401 289.1401 289.1401
```

The three equilibria solutions are: 234, 264, and 289 °K. The solutions converge quickly, only in 5 or 6 steps with a high procession of 1×10^-12. The first solution is the same as the `uniroot(f,c(220,240))` solution in sub-section 5.1.3. In fact, the R routine `uniroot` also uses Newton's method. Our R code here explicitly defines the algorithm for the Newton's method.

Newton's method is very efficient when one knows the reasonable range of possible solutions in advance from physical reasonings. For EBM, we certainly should limit our first guess between (200,340) since the Earth's surface air temperature is most likely in the range of (-70,70)°C. If our initial guess is far away from this range, Newton's method may still yield a solution, which, however, may not be what one expected. For example, when we make an initial guess $T_0 = 100$,

```
root5 <- newton(f, tol=1E-12, x0=100,N=20)
root5
# [1] 827.2544 623.5417 474.8968 372.5619 313.3648
# 292.0552 289.2231 289.1402
# [9] 289.1401 289.1401 289.1401
```

Newton's method still gives a solution, but takes more steps. Instead of yielding the smallest solution 234 °K, it results in the largest solution 289 °K.

6.4 Higher order derivatives

Slope, i.e., derivative, is an indicator of a mountain path's changing rate of height with respect to horizontal distance. The slope can vary and has its derivative, the second derivative, that indicates how curved the path is. The parabola path's second derivative is a constant.

Speed is the rate of change of the position with respect to time. When one accelerates his car from a stop sign with speed equal to zero to the full cruise speed of 65 miles per hour, the speed changes and the car experiences an acceleration period to reach a constant cruise speed. The acceleration measures the changing rate of speed, and is a derivative of a derivative, i.e., the second derivative.

For example, $h = -9.8t^2/2 + 20t + 1[m]$ is the height of a ball being tossed up at the initial height $1[m]$ and initial speed $20[m/s]$ (A baseball pitcher can make this speed!). The speed is $s = dh/dt = -9.8t + 20[m/s]$. The acceleration is $a = ds/dt = -9.8[m/s^2]$, which is the gravitational constant, or called the gravitational acceleration of Earth. Due to the negative acceleration because the gravity pulls down the ball to slow down the ball which has an initial speed of $20[m/s]$. When the ball reaches the highest point, where the ball's speed becomes zero: $-9.8t + 20 = 0$ yields $t = 20/9.8$, approximately 2 seconds. The highest point is approximately as $H = -9.8 \times 2^2/2 + 20 \times 2 + 1 = 21[m]$, about 5 stories high. After the ball reaches the highest point, it starts to drop, i.e., into the stage of negative speed. So the Earth's gravity puts on a force on the ball and causes the negative acceleration, according to the Newton's second law $F = ma$, to the up-tossed ball whose initial positive speed changes to zero and then negative. Here, the air resistance and other forces are neglected.

The units of force is $[newton]$ or $[N]$, which is the force needed to make $1[kg]$ ball have 1 $[m/s^2]$ acceleration. The gravitational force acted on a $0.1[kg]$ ball is thus $0.1[kg] \times 9.8[m/s^2] = 0.98[kg.m/s^2]$ or $0.98[N]$.

Mathematically, one can extend the second derivative, or called second order derivative, to the third derivative, fourth derivative, and so on. The notations are below.

(i) The first derivative, or simply derivative: $f'(x), f', D[f,x], D[f], df/dx, \frac{df}{dx}, y', \dot{y}, \frac{dy}{dx}$

(ii) The second derivative, or second order derivative: $f''(x), f'', D_2[f,x], D_2[f], d^2f/dx^2, \frac{d^2f}{dx^2}, y'', \ddot{y}, \frac{d^2y}{dx^2}$

(iii) The third derivative, or third order derivative: $f'''(x), f''', D_3[f,x], D_3[f], d^3f/dx^3, \frac{d^3f}{dx^3}, y''', \frac{d^3y}{dx^3}$

(iv) The fourth derivative, or fourth order derivative: $f^{(4)}(x), f^{(4)}, D_4[f,x], D_4[f], d^4f/dx^4, \frac{d^4f}{dx^4}, y^{(4)}, \frac{d^4y}{dx^4}$

(v) The fifth derivative, or fifth order derivative: $f^{(5)}(x), f^{(5)}, D_5[f,x], D_5[f], d^5f/dx^5, \frac{d^5f}{dx^5}, y^{(5)}, \frac{d^5y}{dx^5}$

(v) The nth derivative, or nth order derivative: $f^{(n)}(x), f^{(n)}, D_n[f,x], D_n[f], d^nf/dx^n, \frac{d^nf}{dx^n}, y^{(n)}, \frac{d^ny}{dx^n}$

R program can find the derivatives for us. For example, to find $D[x^2]$, the R command is

```
D(expression(x^2),"x")
# 2 * x
```

More examples are below

```
D(expression(exp(-x^2)),"x")
# -(exp(-x^2) * (2 * x))
D(expression(sin(-3*t)-2*cos(4*t-0.3*pi)),"t")
# -(cos(-3 * t) * 3 - 2 * (sin(4 * t - 0.3 * pi) * 4))
```

To find the nth order derivatives, just apply the D command n times. For example, the following R commands can find the second order derivative for the ball-tossing problem:

```
D(expression(-g*t^2/2 + v0*t + h0),"t")
# v0 - g * (2 * t)/2
# Find derivative of this result function
# in order to find the second derivative
D(expression(v0 - g * (2 * t)/2),"t")
# -(g * 2/2)
```

Or one can write a small loop in R to define the nth order derivative.

R derivative function notations are more rigid than WolframAlpha, which has an artificial intelligence to detect what one intends to compute.

6.5 Partial derivatives

The above derivative is for a curve in a 2D space, or on the $xy - plane$. We live in a 3D space or 4D space if we count time as another dimension. When walking on a mountain, we face the choice of different paths, one is a shortcut but steep (i.e., large slope or large derivative), others are longer roads but not as steep (i.e., smaller slopes or derivatives). Thus, the slope or derivative on a surface depends on direction, called directional derivative. Look at Figure 6.4, which is the northern hemisphere.

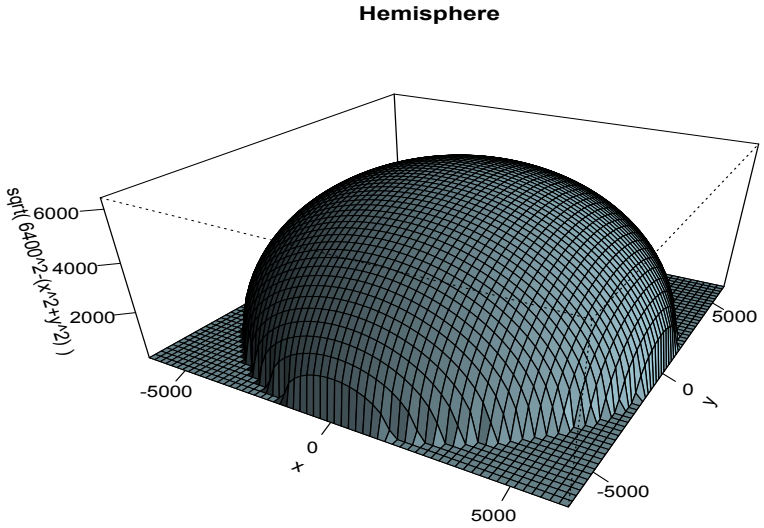


Figure 6.4 Northern hemisphere.

```
x <- seq(-6400, 6400, length= 60)
y <- x
f <- function(x, y) { sqrt(6400^2-(x^2+y^2)) }z <- outer(x, y, f)
z[is.na(z)] <- 1
op <- par(bg = "white")
persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue",
ltheta = 120, shade = 0.75, ticktype = "detailed", main="Hemisphere",
xlab = "x", ylab = "y", zlab = "sqrt( 6400^2-(x^2+y^2) )"
) -> res
round(res, 3)
```

The derivative along the x-axis direction is denoted by

$$\frac{\partial u}{\partial x} \tag{6.16}$$

for the function $u(x, y)$. Similarly,

$$\frac{\partial u}{\partial y} \tag{6.17}$$

denotes the derivative of the function $u(x, y)$ along the y-axis.

```
x <- seq(-6400, 6400, length= 120)
y <- x
f <- function(x, y) { r <- sqrt(x^2+y^2); sqrt(6400^2-r^2) }
z <- outer(x, y, f)
z[is.na(z)] <- 1
op <- par(bg = "white")
persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue",
ltheta = 120, shade = 0.75, ticktype = "detailed", main="Hemisphere",
xlab = "x", ylab = "y", zlab = "sqrt( 6400^2-(x^2+y^2) )"
) -> res
round(res, 3)
```

In general, the derivative along the direction, $\vec{n}$ of θ angle counterclockwise from the x-axis, the directional derivative is

$$\frac{\partial u}{\partial \vec{n}} = \frac{\partial u}{\partial x}\cos\theta + \frac{\partial u}{\partial y}\sin\theta. \tag{6.18}$$

When $\theta = 0$, this reduces to $\frac{\partial u}{\partial x}$, and when $\theta = \pi/2$, this reduces to $\frac{\partial u}{\partial y}$.

One can use the same R derivative command to find partial derivatives. For example,

```
 D(expression(sqrt(6400^2- (x^2+y^2))),"x")
# -(0.5 * (2 * x * (6400^2 - (x^2 + y^2))^-0.5))
D(expression(sqrt(6400^2- (x^2+y^2))),"y")
# -(0.5 * (2 * y * (6400^2 - (x^2 + y^2))^-0.5))
```

Let us look at a point's directional derivative. Check a point on the Greenwich line (i.e., meridional line) on the sphere that approximates Earth: $r = 6,400[km], lat = 45°, lon = 0°$ (Puynormand, France). This point's x, y coordinates are $x = r\cos\phi\cos\theta = 6400\cos 45°\cos 0° = 4,525[km], y = r\cos\phi\sin\theta = 0[km], z = 6400\cos 45° = 4,525[km]$.

The x-directional derivative is

$$\frac{\partial u}{\partial x} = -(x(6400^2 - (x^2 + y^2))^{-0.5})) = -\frac{x}{z} = -1.0. \tag{6.19}$$

This is the rate of z coordinate's change along the great circle, and is dimensionless since it is a change of height [in the units of length] with respect to horizontal distance [another length]. It is negative since the x-axis direction is chosen in the decreasing direction of the great circle at Puynormand, France.

The y-directional derivative is

$$\frac{\partial u}{\partial y} = -(y(6400^2 - (x^2 + y^2))^{-0.5})) = -\frac{y}{z} = 0. \tag{6.20}$$

Since the z-coordinate does not change along a latitude band.

So, the meridional direction gives the steepest path, while the zonal direction gives the most flat path. One may choose to travel from Puynormand to the North Pole via a slope between these two values, i.e., a spiral way to the North Pole. Say, we choose a path that have an angle $\gamma = 4^\circ$ from the zonal direction to the right, equivalent to about 7% slope. Then, the direction of this path is given by $\vec{n} = (-\sin\gamma, \cos\gamma) = (-0.069756, 0.997564)$. The directional derivative at Puynormand along the $\gamma = 4^\circ$ path is

$$\frac{\partial u}{\partial \vec{n}} = \frac{\partial u}{\partial x} \times (-0.069756) + \frac{\partial u}{\partial y} \times 0.0 = 0.069756. \tag{6.21}$$

The vector $(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y})$ is called gradient. It is denoted by

$$\nabla u = (\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}). \tag{6.22}$$

This vector points to the direction at which u has the fastest change. For example, at Puynormand, the fastest change of the z coordinate is along the meridional direction, thus

$$\nabla z = (-1, 0). \tag{6.23}$$

However, when it is off the meridional line, the gradient is not along the x-axis direction anymore since the cross section of the sphere along the x-axis is no longer the great circle, while the gradient as the maximum slope should be always along the great circle. For example, at San Diego (32°N, 118°W), its x, y coordinates are $x = -2,548[km], y = 4,792[km], z = 3,391[km]$. The gradient of z is

$$\begin{aligned} \nabla z &= \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) \\ &= (-x/z, -y/z) \\ &= (0.75, -1.41). \end{aligned} \tag{6.24}$$

This vector is in the direction $(0.47, 0.88)$, which is along San Diego's great circle direction.

An air mass in the atmosphere is subject to many types of forces: pressure gradient force (PGF), centrifugal force (called Coriolis force in meteorology), gravitational force, frictional force, and more. PGF is a force that caused by the difference between a high pressure on one side and a low pressure on the other side. The pressure difference thus results in a force toward the lower pressure side and forms wind. Let $p(x, y)$ be the pressure field at a give time. The PGF is

$$\mathbf{F}_p = -\nabla P/\rho \tag{6.25}$$

where ρ is the density of the air mass. This formula leads to the PGF's units [newton/kg], that is the force per unit mass.

Under the action of PGF and the Coriolis force due to Earth's rotation, the geotropic wind is formed. See Figs. 6.5 and 6.6 for two cases: wind directions on a globe and wind directions over a local hurricane region.

The pressure gradient goes from higher pressure outside to the lower pressure center, which is the hurricane center. Thus, PGF points to the center. Coriolis' force is in the opposite direction to PGF and is proportional to the wind speed. When PGF is

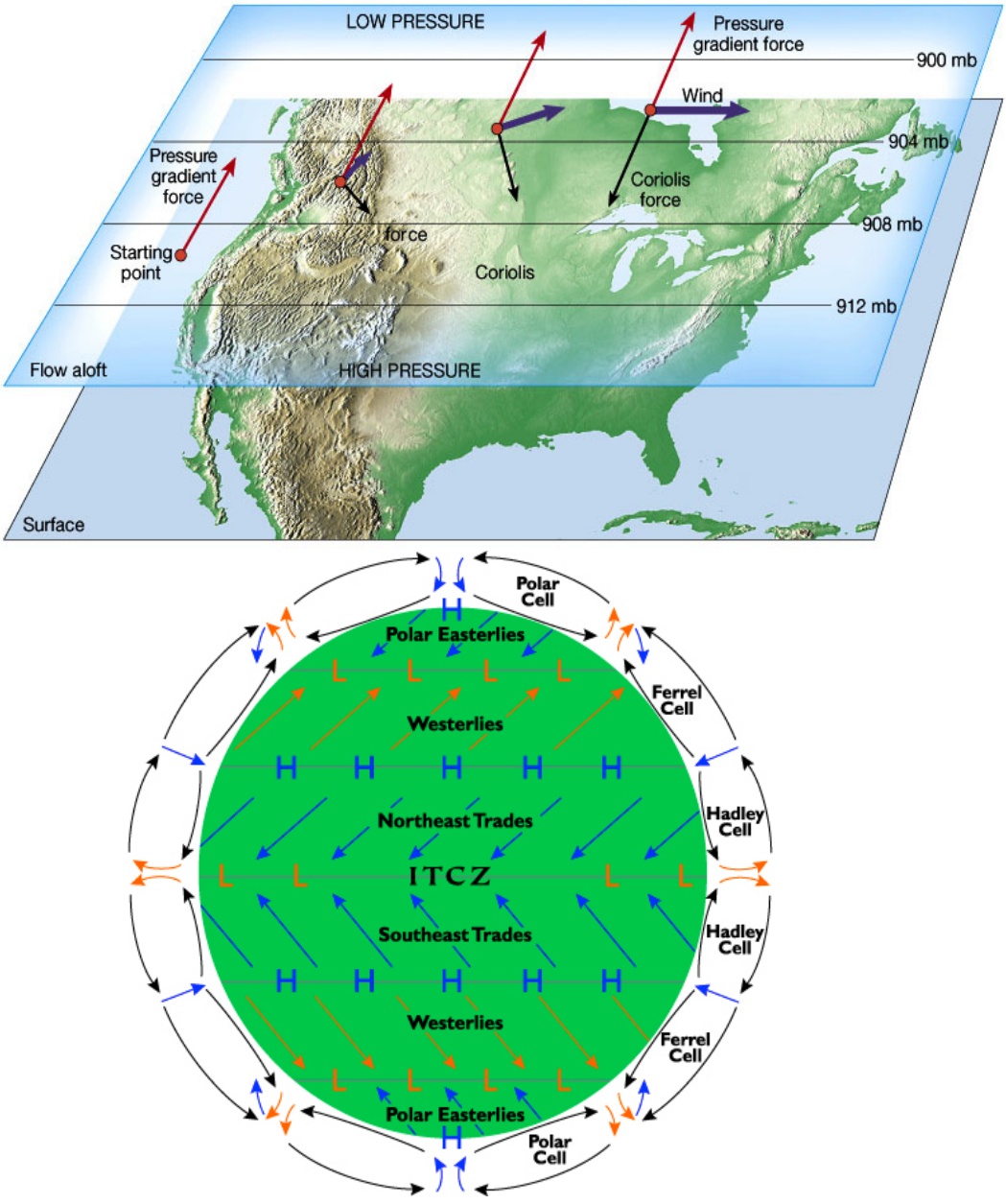


Figure 6.5 Wind directions of global atmospheric circulation.

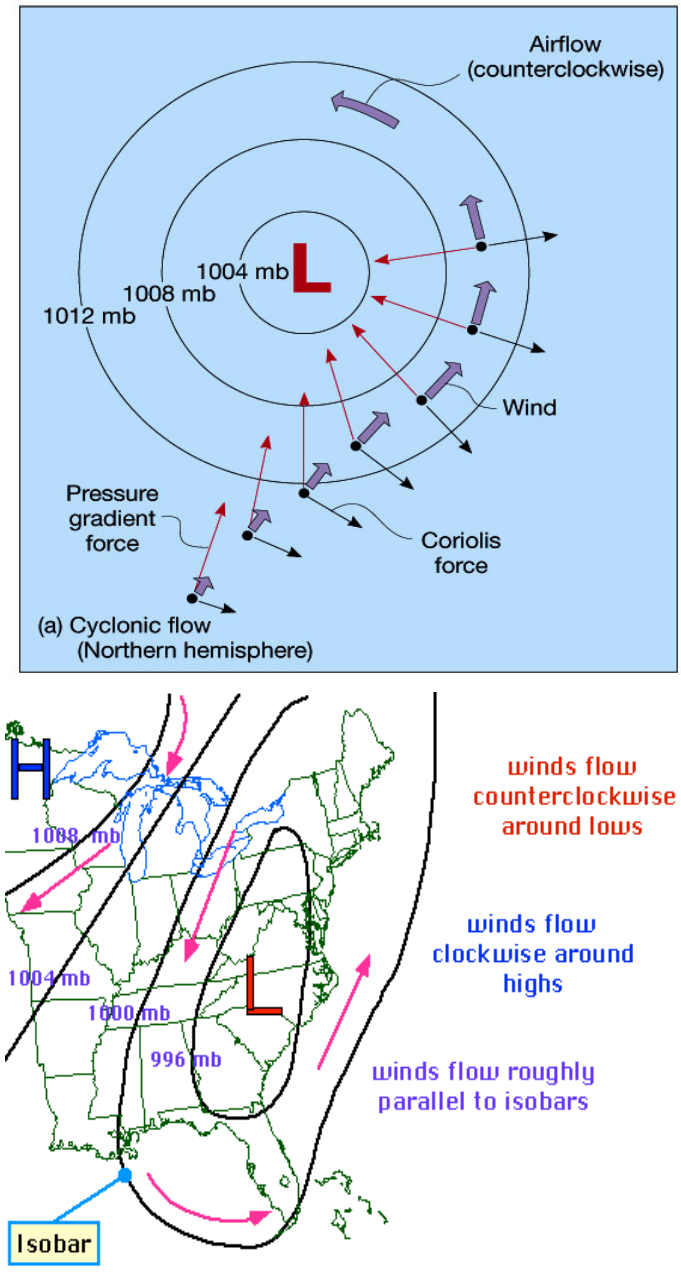


Figure 6.6 Wind directions over a local hurricane region.

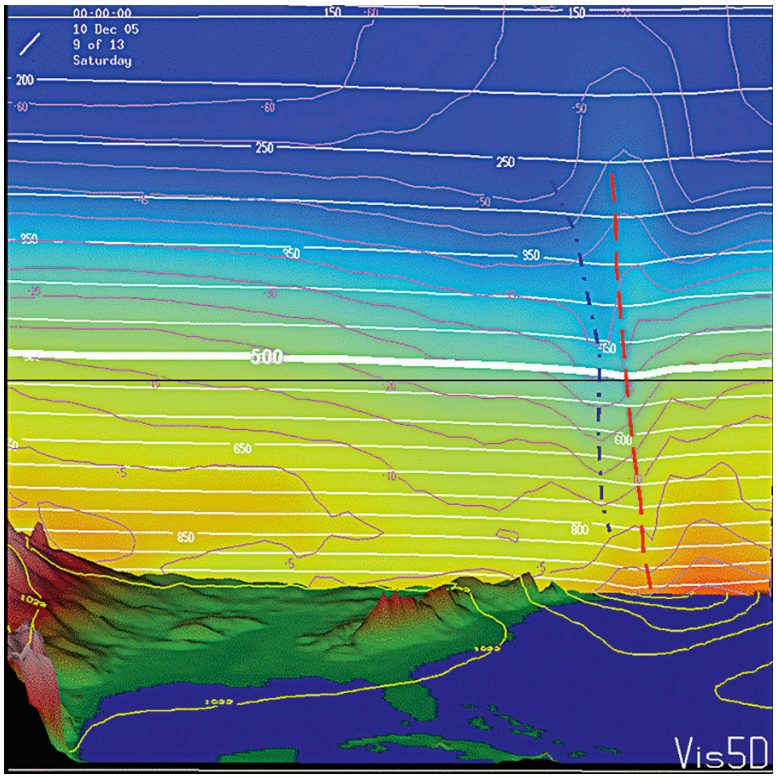


Figure 6.7 The temperature and pressure variations west to east cross-section from 0000 UTC 10 Dec 2005 across the United States. The white lines are isobars, from ground around 1000 mb to the stratosphere around 150 mb (about 14-17 km above the ground). The purple lines are isothermal contours on which the temperature is the same. The yellow lines are sea level pressure contours. The thick dot-dash blue line is a temperature trough which means a cold air goes from west to the east. The background color shows warm (red and yellow) and cold (blue). The thick dashed red line is a pressure trough that leads the temperature front to the east. This figure is from www.vos.noaa.gov/MWL/dec_08/milibar_chart.shtml

large, then wind is strong, then the Corilis force is stronger and balances the PGF to force. These two balanced forces make the wind go along the isobars, on which the air pressure is the same. Strong wind around a low pressure center can form a vortex, a hurricane, when the pressure difference is sufficiently large.

6.6 Spatiotemporal variations of temperature field

The temperature of our atmosphere decreases as the height increases. The colder the higher. When a Boeing 747 cruises at abut 10 km altitude, its outside temperature is often $-52°$C, while the ground temperature may be $20°$C. The air temperature

Figure 6.7 shows the temperature and pressure variations west to east cross-section from 0000 UTC 10 Dec 2005 across the United States.

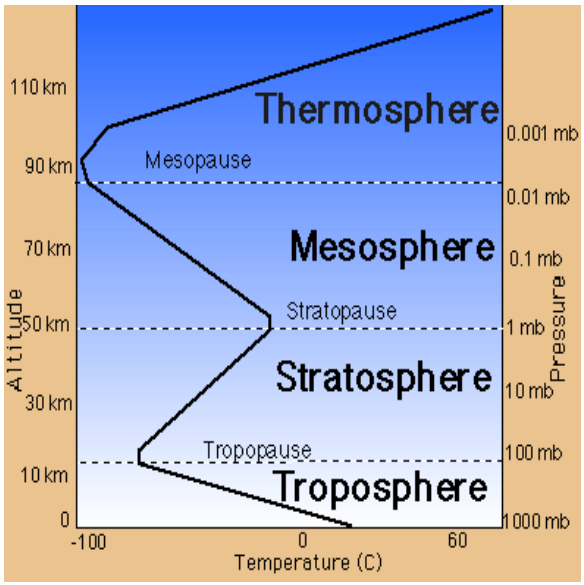


Figure 6.8 Vertical profile of atmospheric temperature.

This figure clearly shows that the atmospheric temperature changes with respect to both time and space. Space wise, the vertical change is much faster than the horizontal change except over a temperature trough.

The atmospheric temperature $T(x, y, z, t)$ is a function in a 4D space: three in space and one in time. The horizontal change of $T(x, y, z, t)$ is small, often with zero slope, while the vertical change is large, namely, $|\frac{\partial T}{\partial x}| << |\frac{\partial T}{\partial z}|$ in most cases, except the temperature troughs.

The vertical temperature change may be illustrated by Fig. 6.8.

This figure shows that the temperature is a function of height, but put function value on the horizontal axis and the independent variable height on the vertical axis. The derivative $\partial T/\partial z < 0$ below tropopause, about 11-12 km heigh, the temperature getting lower when height increasing, as we experience during our airplane taking off period. However, in the stratosphere, where air is very sparse and the atmospheric pressure drops below 100 mb, the temperature increases with height, i.e., $\partial T/\partial z > 0$. Climate scientists often care about only the atmosphere below stratopause where the atmospheric pressure drops to 1 mb.

The atmosphere temperature change rate with respect to height is called lapse rate:

$$\Gamma = -\frac{\partial T}{\partial z}. \tag{6.26}$$

If $\Gamma < 0$ in an atmosphere layer, then this layer is said to have a temperature inversion. The temperature inversion in troposphere may happen in a polar winter, where the surface temperature is very cold, while the air temperature is slightly warmer.

Temperature inversion can also happen in a very small region inside a convective storm where the atmosphere becomes very unstable and cold and hot air masses mix violently.

The ocean water temperature also varies with depth. Usually the top layer of water has little change with depth. This is called the isothermal layer and its depth varies

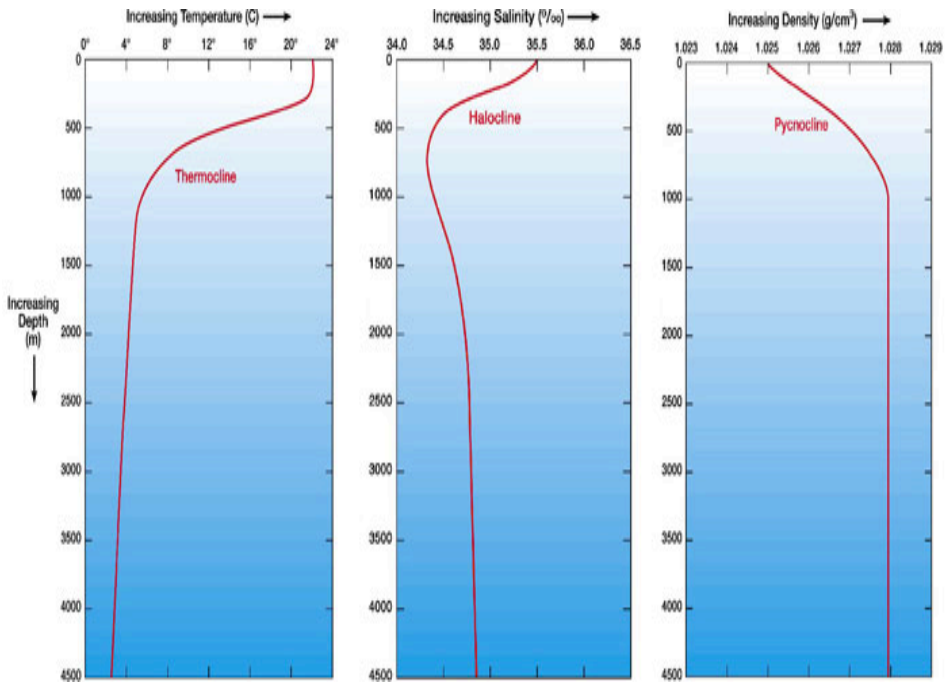


Figure 6.9 Vertical profile of ocean water temperature, salinity, and density.

from 200-300 meters in mid-latitude, to less than 100 meters in low-latitude, and to only 20-40 meters in polar areas. The depth of the isothermal layer has seasonal cycles. See Figure 8.9 for a typical profile of the ocean temperature change with depth.

Water has its highest density when at $4°$C. The bottom of ocean has no access of sunlight and the ocean bottom flow is relatively smooth, and has has little temperature change. Thus, the ocean water temperature below 2,000 meters are all around $4°$C and varies little. Because of this, the global buoy observations for the ocean temperature is mainly focusing on the top 2,000 meters.

The fast transition zone from the top isothermal layer to the deep and almost isothermal layer is called thermocline. The thermocline layer may have a depth of 200-1,000 meters.

If we use z-axis pointing up to measure the depth from the sea level, then $\partial T/\partial z \approx 0$ in the top isothermal layer, a large positive value in the thermocline, and still positive but with a very small value in the deep ocean.

The ocean water salinity and density also vary with depth shown in Fig. 6.9.

The temperature gradient $(\nabla T)(t)$ for a given time is along the fastest increase direction of the temperature. For a nice weather, this gradient is pointing down to Earth in the troposphere. However, it can point up or another direction in a convective storm zone where cold and hot air are mixed.

6.7 Taylor series as a high order approximation

6.7.1 Taylor theorem

The complexity of functions may listed in the following increasing order

1. Constant function: $y = c$ whose derivative is zero and antiderivative is $cx + C$;
2. Linear function: $y = a + bx$ whose derivative is $y = b$ and antiderivative is $ax + bx^2/2 + C$;
3. Quadratic function: $y = a + bx + cx^2$ whose derivative is $b + 2cx$ and antiderivative is $ax + bx^2/2 + cx^3/3 + C$;
4. The general polynomial functions of order n;
5. Exponential function $y = e^x$;
6. Logrithmic functions: $y = \ln(x)$;
7. Trigonometric functions: $y = \sin(x)$; and more

Because of simplicity of the linear function, i.e., a straight line, we have explored linear approximation in the last two sections. Here we extend the same idea to the general polynomial functions of order n

$$g(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n. \tag{6.27}$$

This is an nth-order polynomial with coefficients $a_i (i = 0, 1, 2, \cdots, n)$. We have already demonstrated good approximation by straight lines. It is expected that a quadratic approximation is even better. For most applications, one does not need to go beyond $n = 2$.

Taylor theorem states that

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + O(x^{n+1}) \tag{6.28}$$

where x is nearby zero and $O(x^{n+1})$ stands for the error term which is small since x is nearby zero, hence $|x| < 1$.

This theorem can be proved in the following direct-integration approach:

$$\begin{aligned} f(x) &= f(0) + I[f'(t_1), 0, x] \\ &= f(0) + I[f'(0) + I[f''(t_2), 0, t_1], 0, x] \\ &= f(0) + \{I[f'(0), 0, x] + I[I[f''(t_2), 0, t_1], 0, x]\} \\ &= f(0) + I[f'(0), 0, x] + \{I[I[f''(t_2), 0, t_1], 0, x]\} \\ &= f(0) + f'(0)x + \{I[I[f''(t_2), 0, t_1], 0, x]\} \end{aligned} \tag{6.29}$$

$$\tag{6.30}$$

The first two terms are the linear approximation of $f(x)$ at $x = 0$. One can continue the integral to formula the second or higher orders of approximation. The formula is below

$$a_k = \frac{f^{(k)}(0)}{k!}, \qquad k = 1, 2, \cdots, n. \tag{6.31}$$

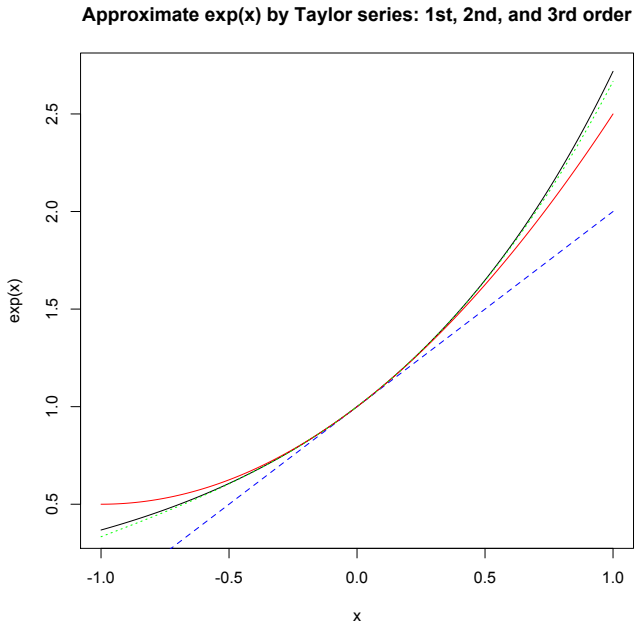


Figure 6.10 Taylor series approximation to a function around $x = 0$: the 1st oder (blue), 2nd order (red), and 3rd order (green).

6.7.2 Taylor expansion for exponential functions

All the derivatives of e^x are e^x which is equal to 1.0 at $x = 0$. Namely, $D[e^x] = e^x$, $D^{(n)}[e^x] = e^x$ and $e^0 = 1$ at $x = 0$. Taylor theorem implies that

$$P_n(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} \tag{6.32}$$

is the nth order polynomial approximation to the exponential function e^x around $x = 0$. The above is true because all the derivatives of e^x is e^x which is equal to 1.0 at $x = 0$.

Figure 6.10 shows the Taylor series approximation to e^x around $x = 0$ in an interval $[-1, 1]$. When $n = 1$, it is a linear approximation. When $n = 2$, it is a quadratic approximation, more accurate than the linear approximation. When $n = 3$, it is the cubic approximation (the green line in the figure), which has almost no distinction from the regional function e^x when $x \in (-0.5, 0.5)$. The approximation is extremely good. This example clearly shows the efficiency of Taylor polynomial approximation.

6.7.3 An example of the second order approximation

Simpson's rule of integration. To be written in the future.

EXAMPLE 6.3

Using a 5th order Taylor series, develop a mathematical model for the temperature profile in ocean water with depth.

EXERCISES

6.1 Practice Midterm $\#1, 20points$: Use integral to describe rainfall history of San Diego, or another location you are familiar with. You may use integral to describe precipitation surplus or deficit. Hint: You can use at least one figure. The English text must be longer than 100 words.

6.2 Practice Midterm $\#2, 20points$: Average US residents' bank balance is $\$5,000$. The bank balance is normally distributed. A group of 25 samples was taken. The sample data have a mean equal to $\$5,000$ and standard deviation of $1,000$. Find the confidence interval of this group of samples at 95% confidence level.

6.3 Practice Midterm $\#3, 20points$: A ball is been shot up straight at 10 $[m/s]$ at an initial height $2[m]$. How long it will take the ball to reach the maximum point? Hint: Choose the gravitational acceleration to be $9.8[m/s^2]$, and find an approximate answer. The general formula is $h = -gt^2/2 + v_0t + h_0$.

6.4 Practice Midterm $\#4, 20points$:Find the linear approximation of $f(x) = x^2 - 1$ at $x = 1.5$. If this 1.5 is used as the first guess, find the next approximate root of $f(x) = 0$ by Newton's method.

6.5 Practice Midterm $\#5, 20points$: The following is the SVD results

```
mat
     [,1] [,2]
[1,]    1    1
[2,]    1   -1

svd(mat)
$d
[1] 1.414214 1.414214

$u
           [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,] -0.7071068  0.7071068

$v
     [,1] [,2]
[1,]   -1    0
[2,]    0   -1
```

Use A=UDV' to recover the first column of

```
mat
     [,1] [,2]
[1,]    1    1
[2,]    1   -1
```

Show detailed calculations of all the relevant matrices and vectors. Use space-time decomposition to describe your results. Extra 5 bonus points: Describe the space and temporal modes, and their corresponding variances or energies.

CHAPTER 7

CLIMATE SCIENCE TOPICS OF CALCULUS II: INTEGRALS

This chapter includes some integrals commonly used in climate sciences, such as geopotential, work done by pressure, heat capacity, the first and second laws of thermodynamics, convergence and divergence, and conversation of mass, momentum, and energy.

7.1 Geopotential

The geopotential Φ is the potential energy of unit mass at a certain altitude z with respect to the sea level. The mathematics expression is an integral

$$\Phi = I[g, 0, z], \tag{7.1}$$

where g is the gravitational acceleration which is a function of latitude and geometric elevation from sea level. The geopotential Φ's units is $[m^2 s^{-2}]$ in SI system, which is the work done to raise 1 [Kg] mass from sea level to z level.

The geopotential height is

$$Z = \frac{1}{g_0} I[g, 0, z], \tag{7.2}$$

where g_0 is the standard gravity at the mean sea level, and is 9.80665 $[m/s^2]$, or 32.174 ft/s^2. Z's units is [m] or [ft]. Thus, $Z = f(\phi, \theta, z, t)$ is a function of lat, lon, actual altitude, and time.

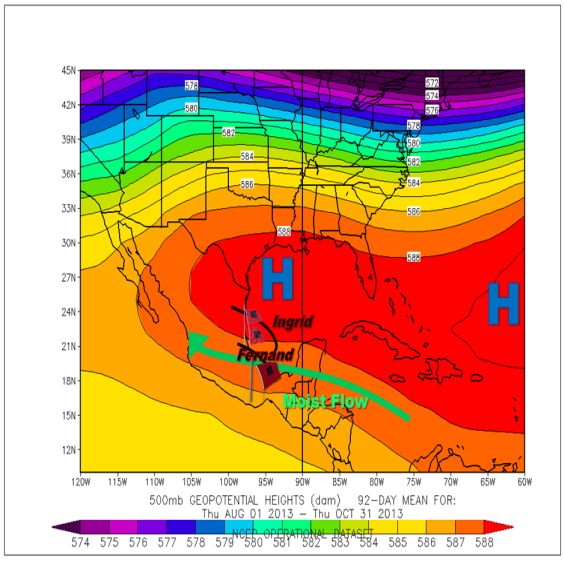


Figure 7.1 Geopotential height of 500 mb for the US and the adjacent Atlantic ocean during the 2013 hurricane season. Hurricane Ingrid: September 12-17, 2013, peak wind 85 mph, minimum surface pressure 983. Tropical Storm Fernand: August 25-26, 2013, peak wind 60 mph, minimum surface pressure 1001. This image is from the US National Weather Service Southern Region Headquarters.

The meteorological use of geopotential height is often the Z value reaching a certain atmospheric pressure, $Z = f(\phi, \theta, p, t)$. For a given pressure, what is the geopotential height? Thus, $Z = f(\phi, \theta, 500mb, t)$ is used to describe the surface of a given pressure, say, 500 mb. See Fig. 7.1 for a general steering pattern of moisture from Gulf of Mexico to the southern US in the 2013 peak hurricane season August-October.

The 500 mb pressure surface is 588 meters high over the Gulf of Mexico, but it is only 572 meters over New England. The higher the 500 mb surface, the higher the land or ocean surface pressure. Thus, geopotential height is an intuitive way to indicate the atmospheric pressure field.

The geopotential height and pressure are related. In the case of geostrophic flow of no air acceleration, small vertical velocity, and no friction, the pressure difference between the point in atmosphere (x, y, z) and the corresponding sea level point $(x, y, 0)$ is the integration of the total amount of the gravitational force in $[0, Z]$:

$$p_0(x, y) - p(x, y, Z, t) = I[\rho g, 0, Z] \tag{7.3}$$

where ρ is the air density, and 0 in the integral means sea level.

For a given pressure, say $p(x, y, z, t) = 500$ mb, Z must vary with respect to location and time (x, y, t), which forms the geopential height surface $z = Z(x, y, t)$ evolving with time. An average height in a period of time is shown in Fig. 7.1.

The gradient of geopotential height ∇Z and that of atmospheric pressure $\nabla p(x, y, Z)$ are in the same direction, along the opposite side of the PGF. Due to Coriolis force, the wind goes counter clockwise around the lower pressure center and clockwise around the high pressure center over northern hemisphere. Thus, Figure 7.1's wind from

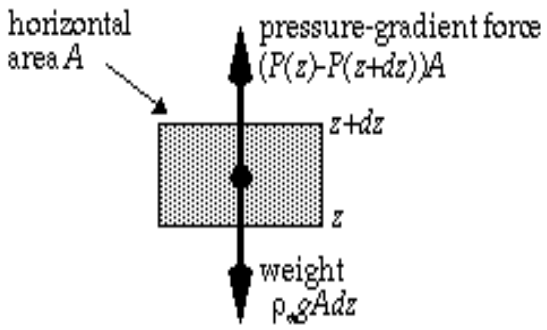


Figure 7.2 A small pressure decrease due to a small altitude increment.

ocean to land is in the direction shown by the green arrow, along which the ocean moisture is transported to the land and provides the water source of storms over the land in the southern US.

From the wind and moisture transport shown in Fig. 7.1, we can conclude that the geopotential, geopotential height, and atmospheric pressure are critical functions a weather forecaster should know. Indeed, various kinds of maps of geopotential, geopotential height, and atmospheric pressure are plotted in weather forecaster centers.

7.2 Exponential decrease of pressure in an isothermal layer of atmosphere

If in a layer of atmosphere temperature changes with altitude is negligible, we treat this layer of atmosphere as an isothermal atmosphere. Under the condition of ideal gas assumption for the air in this layer, the pressure can be proven to decay exponentially with altitude.

A small increase of altitude is denoted by dz, which results in a small pressure change, denoted by dp. The change is caused by the gravitational force on the air mass in this altitude increment dz. Figure 7.2 illustrates this relation: the vertical pressure gradient force balanced by the weight (i.e., the gravitational force). Thus,

$$dp = p(z + dz) - p(z) = -\rho g dz. \tag{7.4}$$

The ideal gas law $pV = nR_dT$ can be written as $p/T = (n/V)R_d$ which becomes

$$\rho R_d = \frac{p}{T}, \tag{7.5}$$

where T is the isothermal temperature of this layer, n is the amount of substance of gas in the volume V, and R_d is the ideal gas constant for dry air, also called universal gas constant, which is $8.314462[J(mol)^{-1}K^{-1})]$. This number is the product of Botlzmann constant $1.380649 \times 10^{-23}[JK^{-1}]$ and Avogadro constant $6.022141 \times 10^{23}[(mol)^{-1}]$.

Substituting this relation into equation (7.4) yields

$$dp = -\frac{p}{TR_d} g dz, \tag{7.6}$$

$$\frac{dp}{p} = -\frac{gdz}{TR_d}. \tag{7.7}$$

Integrate both sides of the equation (7.7) assuming that g, T and R_d are constant in this isothermal layer of atmosphere from its bottom z_1, corresponding to p_1, to its top z_2, corresponding to p_2. We have the following

$$I[\frac{dp}{p}, p_1, p_2] = I[-\frac{gdz}{TR_d}, z_1, z_2]. \tag{7.8}$$

The anti-derivative of $1/p$ is $\ln p$. Thus,

$$\ln p_2 - \ln p_1 = -\frac{g}{TR_d}(z_2 - z_1), \tag{7.9}$$

or

$$\ln(p_2/p_1) = -\frac{g}{TR_d}(z_2 - z_1), \tag{7.10}$$

or

$$p_2 = p_1 \exp\left(-\frac{g}{TR_d}(z_2 - z_1)\right). \tag{7.11}$$

This shows the exponential decay of pressure with respect to altitude.

7.3 Work done by an air mass in expansion

Suppose an air mass with volume V_1 is expanded to V_2. How can one express the work done by the air mass to its surrounding?

Suppose that the air mass is in a cylindrical shape with a cross section area A and the expansion is through the extension of cylinder's length. When the cylinder is expanded by a small increment dx in height, the force is pA and the distance of the force is dx. Thus, the small amount of work done by the air system is

$$dW = pAdx. \tag{7.12}$$

Here Adx can be regarded as a small volume increment dV. Thus, the work is

$$dW = p(V)dV, \tag{7.13}$$

where pressure is a function of volume. The total amount of work done by the air system to its surrounding environment when the volume is increased from V_1 to V_2 is

$$W = I[p, V_1, V_2] \tag{7.14}$$

If it is compression, i.e., $V_2 < V_1$, then the work is negative, meaning the environment has done the work to the system, which gains potential.

When a system goes through a cycle of expanding from V_1 to $V_2 > V_1$ via a pressure path $p_{12}(V)$ and compressing from V_2 to V_1 via another path $p_{21}(V)$, the work done by the system is not zero when the two paths are not the same:

$$\Delta W = I[p_{12}(V), V_1, V_2] - I[p_{21}(V), V_2, V_1] = \oint p(V)dV \neq 0. \tag{7.15}$$

Here, $\oint$ denotes an integral along a closed path or called closed contour.

7.4 Internal energy and enthalpy

When it is heated from temperature T_1 to T_2, a system of dry air gains potential. If the system's volume is kept the same, the gain is the system's internal energy from u_1 to u_2. The gain is an integration of specific heat capacity with respect to temperature:

$$u_2 - u_1 = I[c_v(T), T_1, T_2], \tag{7.16}$$

where c_v is the specific heat capacity of dry air at the state of constant volume, and the internal energy's units is [joule/unit mass].

When the system's pressure is kept the same, the gain is enthalpy from h_1 to h_2, which includes both internal energy gain and the work done to the system because the volume changes under a pressure means the work done to the system. The gain is also an integration of specific heat capacity with respect to temperature:

$$h_2 - h_1 = I[c_p(T), T_1, T_2], \tag{7.17}$$

where c_p is the specific heat capacity of dry air at the state of constant pressure, and the enthalpy's units is also [joule/unit mass].

Intuition suggests that for the same heat, $c_p(T)$ is more efficient to allow absorption of more heat energy into the system due to the volume expansion, while $c_v(T)$ is comparably less efficient. Thus,

$$c_p(T) \geq c_v(T). \tag{7.18}$$

Their difference is an important property of a system, called specific gas constant, denoted by R:

$$R = c_p - c_v. \tag{7.19}$$

Surprisingly, this R does not change with temperature T and is an intrinsic property of the gas. Different gas have different R values. For example, standard air has $R = 287[JK^{-1}Kg^{-1}]$, water vapor $R = 462$, carbon dioxide $R = 189$, methane $R = 310$, and. hydrogen $R = 4124[JK^{-1}Kg^{-1}]$.

For the ideal gas, this R is the same as the universal gas constant in the ideal gas law

$$PV = nR_dT \tag{7.20}$$

where the universal gas constant $R_d = 8.3145[Jmol^{-1}K^{-1}]$, and n is the gas' amount with units [mol]. Since one mole of dry air's weight is $0.02896[Kg \cdot mol^{-1}]$, we have

$$\frac{8.3145[Jmol^{-1}K^{-1}]}{0.02896[Kg \cdot mol^{-1}]} = 287[JK^{-1}Kg^{-1}]. \tag{7.21}$$

The heat capacity we commonly refer to is usually the specific heat capacity of fixed volume. Water has much larger heat capacity than air. Per unit mass, water's heat capacity is $4.18[Jg^{-1}K^{-1}$, about four times of that of dry air $1.01[Jg^{-1}K^{-1}]$. The total amount of water on the Earth surface is about $1,350 \times 10^{18}[kg]$, which is about 260 times of the Earth's air mass, about $5.1 \times 10^{18}[kg]$. Thus to heat up the Earth's air uniformly by 1°C, it needs $5,151 \times 10^{18}[J]$ heat. To heat up the Earth's surface water by the same 1°C, it needs about 1,000 times more heat, $10^3 \times 5,643 \times 10^{18}[J]$. Thus, the global ocean can store a lot of heat. The global climate change, particularly

the global warming research, should also consider the roles of ocean in addition to the atmospheric circulations and the land observations.

Enthalpy is a very useful parameter in thermodynamics that measures both internal energy and the work done to a system. The general definition of enthalpy is

$$H = U + pV, \tag{7.22}$$

where U is the internal energy, p pressure, and V volume, and H has the same units as U.

The differential of H means a small increment of H when all the relevant independent variables are also subject to a small change. This statement is denoted by the differential of the above equation:

$$dH = dU + pdV + Vdp. \tag{7.23}$$

Note that the differential of a product $d(uv) = udv + vdu$ is a sum of two differentials, increment in one parameter while another is fixed. This is called the product rule of differentiation, or product rule of derivative.

The first two terms of the right hand side is the system's total energy increment: the internal energy increment dU plus the mechanical energy pdV. These energies come from the heat dQ put into the system. Thus,

$$dH = dQ + Vdp. \tag{7.24}$$

A system may expand its volume from v_1 to v_2 in many ways. An important way is to keep the system having the same temperature, which is called the isothermal process. Another is to keep no work from being done to the system, which is called an adiabatic process. See Figure 7.3 for these two processes. Various kinds of relationships among p, v, and T can be derived based on these processes and using integration of differentials.

7.5 Entropy and information loss

The specific heat capacity c_p was defined by enthalpy:

$$dh = c_p dT \tag{7.25}$$

for per unit mass. The first law of thermodynamics for energy conservation equation (7.24) means that

$$c_p dT = dq + vdp \tag{7.26}$$

per unit volume. This equation can be rewritten as

$$dq = c_p dT - vdp. \tag{7.27}$$

Dividing both sides by T and using the ideal gas law $pv = RT$ per unit mass yields

$$\frac{dq}{T} = c_p \frac{dT}{T} - R\frac{dp}{p} = d\ln(T^{c_p}) - \ln(p^R) = d\ln\left(\frac{T^{c_p}}{p^R}\right). \tag{7.28}$$

This is an exact differential of the function

$$\ln\left(\frac{T^{c_p}}{p^R}\right). \tag{7.29}$$

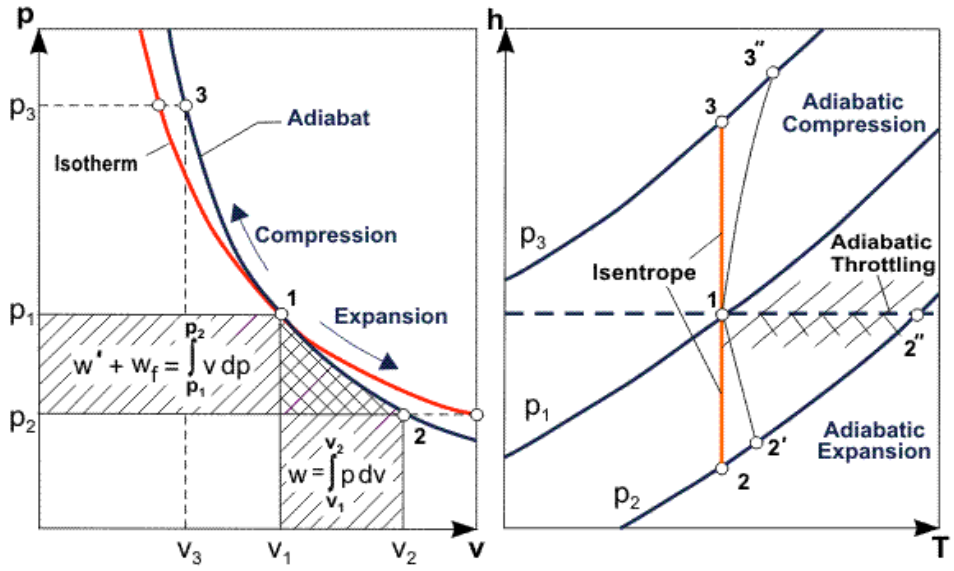


Figure 7.3 Adiabatic and isothermal expansion or compression.

Because of exact differential, the closed contour integral should be zero, i.e.,

$$\oint \frac{dq}{T} = \oint d \ln \left(\frac{T^{c_p}}{p^R} \right) = 0. \tag{7.30}$$

However, the second law of thermodynamics says that this equality should be an inequality

$$\oint \frac{dq}{T} \leq 0, \tag{7.31}$$

which is called the Clausius' inequality.

We also define

$$\Delta\eta = -I[dq/T, A, B] \tag{7.32}$$

as the entropy change from the state A to state B. The second law of thermodynamics means that this change must be non-negative.

In general, the entropy of a probability distribution of temperature T is

$$S = -k_B \sum_j P_j(T) \ln(P_j(T)) \tag{7.33}$$

where k_B is the Boltzmann constant. One can prove that the increment of S due to the input of energy into the system is non-negative. If $-S$ is considered information, then information is lost in any change of a system. The nature tends to be more chaotic. It is natural for kids to make a mess in a house. Only the external interference can restore a house in an order.

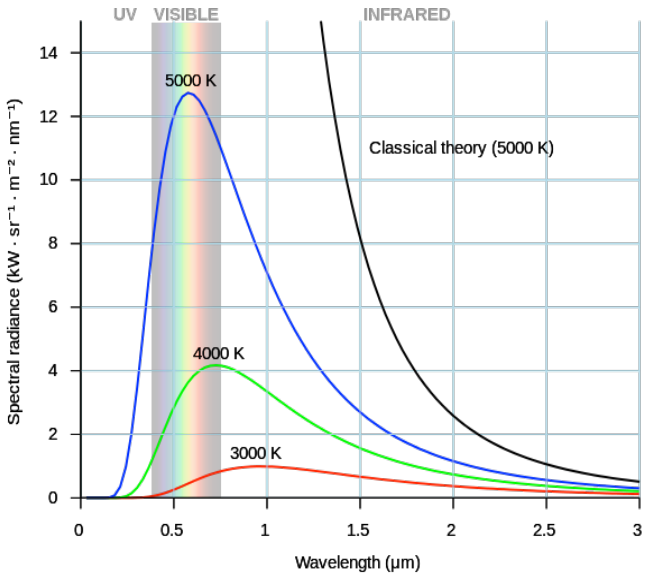


Figure 7.4 Planck's law of black body radiation and the ultraviolet catastrophe (black curve).

7.6 Derivation of Stefan-Boltzmann's black body radiation from Planck's law of radiation

Planck's law of radiation quantifies the radiation emitted by a blackbody in thermal equilibrium at a given temperature. Max K. Planck (1858-1947) proposed the law of spectral radiance for a blackbody at a given temperature in 1900. The radiation energy per unit wavelength is

$$B_\lambda(\lambda, T) = \frac{2hc^2}{\lambda^5} \times \frac{1}{\exp[hc/(k_b \lambda T)] - 1}, \tag{7.34}$$

where $h = 6.626070040(81) \times 10^{-34}[J \cdot sec]$ is the Planck constant, $k_b = 1.3806488(13) \times 10^{-23}[J \cdot K^{-1}]$ is Stefan-Boltzmann constant, $c = 300,000,000[m \cdot sec^{-1}]$ is the light speed in vacuum, $T[^\circ K]$ is temperature, and $\lambda[\mu m]$ is wavenumber. The SI units of B_λ is $W \cdot sr^{-1} \cdot m^{-3}$, where sr stands for steradian, or called square radian, SI unit of solid angle on a sphere.

Figure 7.4 shows the plots of $B_\lambda(\lambda, T)$ as a function of wave length λ for given temperatures T. The figure shows that the peak radiation moves to the lower wave number zone (i.e., the higher frequency zone) when the temperature increases. This agrees with our intuition. A very high temperature body, such as a burning arc welding rod (around 6,000 °C), shows bright purple color, which has higher frequency than liquid iron that emits bright red light (around 1,200°C).

For a given T, the total amount of energy emitted by a blackbody throughout the entire range of wavelength is an integration of $B_\lambda(\lambda, T)$ with respect to wavelength λ:

$$L = \int_0^\infty \frac{2hc^2}{\lambda^5} \times \frac{1}{\exp[hc/(k_b \lambda T)] - 1} d\lambda. \tag{7.35}$$

This is equivalent to integrate with respect to frequency ν from zero to infinity.

$$L = \int_0^\infty \frac{2h\nu^3}{c^2} \times \frac{1}{\exp[h\nu/(k_bT)] - 1} d\nu, \tag{7.36}$$

since $\nu \times \lambda = c$.

Let

$$x = h\nu/(k_bT), \tag{7.37}$$

then

$$dx = hd\nu/(k_bT), \tag{7.38}$$

or

$$d\nu = (k_bT/h)dx. \tag{7.39}$$

Substituting the above into the L formula

$$L = 2\frac{k_b^4T^4}{h^3c^2} \int_0^\infty \frac{x^3}{e^x - 1} dx. \tag{7.40}$$

Let us consider this integration by series

$$\begin{aligned} I &= \int_0^\infty \frac{x^3}{e^x - 1} dx \\ &= \int_0^\infty \frac{x^3e^{-x}}{1 - e^{-x}} dx. \end{aligned} \tag{7.41}$$

The integrand can be expanded into a convergent series

$$\frac{e^{-x}}{1 - e^{-x}} = \sum_{n=1}^\infty e^{-nx}. \tag{7.42}$$

Integration by parts for the following interal

$$\int_0^\infty x^3e^{-nx} dx \tag{7.43}$$

leads to

$$\int_0^\infty x^3e^{-nx} dx = \sum_{n=1}^\infty \frac{6}{n^4} = \frac{\pi^4}{15}. \tag{7.44}$$

Thus,

$$L = 2\frac{k_b^4T^4}{h^3c^2}\frac{\pi^4}{15} = \frac{2\pi^4k_b^4}{15h^3c^2}T^4. \tag{7.45}$$

Its SI units is $Wm^{-2}sr^{-1}$.

The total radiated energy is an integration of the energy emitted from a hemisphere

$$M = \int_0^{2\pi} d\theta \int_0^{\pi/2} d\phi \;\; L(\nu, T)\cos\theta\sin\theta = \pi L(\nu, T). \tag{7.46}$$

This leads to the Stefan-Boltzmann radiation law

$$M = \alpha T^4 \; [Units : Wm^{-2}], \tag{7.47}$$

where

$$\alpha = \frac{2\pi^5k_b^4}{15h^3c^2} = 5.670373 \times 10^{-8} \;\; [W \cdot m^{-2}K^{-4}]. \tag{7.48}$$

EXERCISES

7.1 Suppose that the greenhouse effect results in a net energy gain of the Earth surface by 1.0 $[Wm^{-2}]$. If the gained heat is all used to heat the Earth's atmosphere, how many years does it need to warm the Earth's entire atmosphere by $1.0°$C? If the heat is all used to heat the Earth's surface water, including the water in the oceans, lakes, and rivers, how many years does it take to warm the water by $1.0°$C? Make comments about this study's implication on global warming.

Hint: You can find the relevant information about the Earth's atmosphere and water in this book or from internet.

7.2 A piece of material with mass m_1, specific heat c_1 and temperature T_1 is put in contact with another piece of material with mass m_2, specific heat c_2 and temperature T_2. Without any loss of energy, the temperatures of the two pieces eventually become the same, T, due to heat conduct. (a) Find T from the given conditions: T_1, T_2, c_1, c_2, m_1 and m_2. (b) Use weighted average to explain your result. (c) Discuss some special cases, such as two very different masses, the two same masses, two very different specific heats, and the two same specific heat.

Hint: Use the energy conservation law: energy before contact = energy after a long time contact, i.e.,

$$c_1 m_1 T_1 + c_2 m_2 T_2 = c_1 m_1 T + c_2 m_2 T. \tag{7.49}$$

7.3 (a) Under the assumptions of isothermal layer and ideal gas, derive the equation of exponential decay of pressure with respect to altitude, i.e., eq. (7.11) using the calculus method: cut a small piece of air of thickness equal to dz and base area equal to A, as shown in Figure 7.2, and then integrate all the small pieces together. Start with the balance of forces on this small piece of air and derive the final equation (7.11). (b) Suppose that (i) an isothermal layer has an average temperature $253°$K, (ii) the layer's bottom is at the sea level with $z_1 = 0$ and $p_1 = 1000[mb]$, and (iii) the isothermal layer's top pressure is $p_2 = 500[mb]$. Calculate the isothermal layer's top coordinate z_2 using the formula derived in (a) and using a calculator or R.

Hint: Pay attention to the units of the universal gas constant. Search internet and find out how many grams of air are equal to one [mol] of air. If certain conditions are attached to the units conversion, then discuss the conditions and the numerical results.

CHAPTER 8

CONSERVATION LAWS IN CLIMATE DYNAMICS

Climate models are a set of equations of conserved quantities (e.g., mass, momentum, energy) and various kinds of relationships among the climate parameters, such as the algebraic equation of ideal gas law, criteria of the formation of rain drops and snow flakes, probability distribution functions of cloud fraction, and Stefan-Boltzmann radiation law or Budyko law of radiation. These mathematical and statistical equations together with initial and boundary conditions form a climate model. Numerical solutions of the model yields numerical simulations of the Earth's climate. This chapter provides the detailed mathematics for the conservations commonly used in climate science.

The climate dynamical models are built upon the conservation of mass, momentum, and energy for a small volume of mass. This small volume means a cut, an application of the calculus method. Differentials and derivatives will be involved, and the conservations will thus lead to a set of equations of derivatives or differentials, called differential equations (DE). If more than one independent variables are involved, such as the air temperature's dependence on 3D position (x, y, z) and time t, i.e., $T = T(x, y, x, t)$, then the partial derivatives $\frac{\partial T}{\partial t}, \frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z}$ will be involved. The equations involve partial derivatives are called partial differential equations (PDE). An equation that involves only ordinary derivative is called ordinary differential equation (DOE). PDEs are often more complex than ODEs. Because of the calculus method of cut, the numerical solutions of the the climate model can only represent the average of the climate parameter in this small volume, which is often a

latitude-longtitude grid box horizontally, plus the global air's thickness of each layer, and

8.1 Conservation of mass

8.1.1 Lagrangian and Eulerian observers, and mass conservation in the Lagrangian framework

The fluid mass contained inside a small box with volume V is conserved, meaning that the mass does not change with time. The average density of this small box of fluid is ρ. Thus, the mass in the small volume is $m = \rho V$. The conservation of mass means that the total mass in this box does not change with time:

$$\frac{dm}{dt} = 0, \tag{8.1}$$

i.e.,

$$\frac{d}{dt}(\rho V) = 0. \tag{8.2}$$

This formula means that one follows the box and sees no leaking of fluid from the box even though the box can deform since it is a fluid body. Because of the box deformation, the volume and the average density can change with time. However, the box is closed: no mass goes out and no mass comes in. It is like a sealed plastic tube full of fluid.

The observer following the fluid volume is called a Lagrangian frame: the fluid flow properties, such as density and velocity, are function of time and the initial position, i.e., $\rho(x_0, y_0, z_0; t)$. For a given initial position, the fluid box is only a function of time. The Lagrangian observer acts as if a dog walker while the small fluid mass may be considered as the dog being walked. We often denote the derivative in a Lagrangian frame as

$$\frac{D}{Dt}. \tag{8.3}$$

Thus, the mass conservation equation is

$$\frac{D}{Dt}(\rho V) = 0, \tag{8.4}$$

or

$$\frac{Dm}{Dt} = 0, \tag{8.5}$$

However, it is extremely difficult, if not possible, to follow a small fluid box. In most cases, we observe the fluid flow at a fix place, or place an instrument inside a flow at a fix place. When we experience wind, we do not follow the wind, rather we stay where we are, feel the wind, and sense the wind speed and direction. Thus, we are an instrument measuring the flow of atmosphere passing by us. The above fluid box can pass us and be sensed, or observed, by us.

Another example is our watch of water flow over a river. We observe the waves and rapid of water within the range of our vision. We do not trace the individual water particle or water mass.

This fixed-position observer is an Eulerian framework, who experiences air flowing by, feels the velocity, and density at the point of observation, and watch the water flowing down the river. Still another example is our watch of the ocean waves coming onshore. We see a wave coming from 100 meters away from the shore, moving rapidly to the shore and arriving at the shore within a short time, say 10 seconds. The wave has arrived to the shore, traveling at a rapid speed $10[m/s]$, running straight toward the shore. However, a small ocean water box does not go directly to the shore. It goes up and down and circulates in the neighborhood of the original position of 100 meters away from the shore. The Lagrangian observer tags the box and wonders around this position. The Eulerian observer cares about the waves coming to the shore.

8.1.2 Total derivative

Consider a fluid density change $\Delta\rho$ in a Lagrangian box of fluid observed by a Lagrangian observer from time t to $t + \Delta t$. The Lagrangian box's position is given by coordinates (x, y, z), which are functions of t. Thus, $(x(t), y(t), z(t))$ from t_1 to t_2 gives trajectory of the small Lagrangian fluid box in the time interval $[t_1, t_2]$. When t is increased to $t+\Delta t$, the small Lagrangian box is changed $(x+\Delta x, y+\Delta y, z+\Delta z)$. The change of the density from $(x(t), y(t), z(t))$, to $(x(t)+\Delta x, y(t)+\Delta y, z(t)+\Delta z)$ due to the time change from t to $t + \Delta t$ is

$$\Delta\rho = \rho((x + \Delta x, y + \Delta y, z + \Delta z, t + \Delta t) - \rho(x, y, z, t). \tag{8.6}$$

The linear approximation of this expression around (x, y, z, t) yields

$$\Delta\rho \approx \frac{\partial\rho}{\partial t}\Delta t + \frac{\partial\rho}{\partial x}\Delta x + \frac{\partial\rho}{\partial y}\Delta y + \frac{\partial\rho}{\partial z}\Delta z. \tag{8.7}$$

Dividing both sides of the above by the time increment Δt leads to

$$\frac{\Delta\rho}{\Delta t} \approx \frac{\partial\rho}{\partial t} + \frac{\partial\rho}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial\rho}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial\rho}{\partial z}\frac{\Delta z}{\Delta t}. \tag{8.8}$$

Here, the partial derivatives are taken at point (x, y, z, t). For the Lagrangian observer,

$$\frac{\Delta x}{\Delta t} \tag{8.9}$$

is the x-direction velocity u of the small Lagrangian box at time t when Δt approaches zero. The same can be said about $\frac{\Delta y}{\Delta t}$ and $\frac{\Delta z}{\Delta t}$.

Thus, as Δt approaches zero,

$$\frac{\Delta\rho}{\Delta t} \tag{8.10}$$

represents the total change rate of density at t and is regarded as the total derivative

$$\frac{D\rho}{Dt}. \tag{8.11}$$

When Δt approaches zero, the equation (8.8) thus will become

$$\frac{D\rho}{Dt} = \frac{\partial\rho}{\partial t} + \frac{\partial\rho}{\partial x}u + \frac{\partial\rho}{\partial y}v + \frac{\partial\rho}{\partial z}w. \tag{8.12}$$

This means that a total derivative with respect to time has two parts. The first is the change rate directly due to time as if the fluid is at rest, and is called the local derivative. The second part is the density change rate due to the motion of the fluid, and is called the advection part of the total derivative, or called density advection.

In the above total derivative expression, the velocity field (u, v, w) are the same to both Lagrangian and Euler observers. Thus, the total derivative provides a mathematical link between the two observers, when coming to the rate of change.

8.1.3 Mass conservation in the Eulerian framework

The Eulerian observer has his own small box for fluid. His box is a fixed observation device that allow fluid to freely come in and to freely go out. It is a virtual box, or a conceptual box. The Eulerian observer records the amount of fluid goes in and out. This is in contrast to the Lagrangian small box, which is a plastic tube flowing with the fluid and does not allow fluid to go in or out, but allows the tube to deform: stretch, compression or distortion. The Eulerian small box does not change its shape and stays at a fixed position with instrument.

In the Eulerian framework, the above Lagrangian derivative can be written into two terms. One is the mass increase with respect to time in a virtual fixed box. The increase mass must come from the mass flux from outside of the box to inside of the box. Thus,

$$\frac{\partial(\rho V)}{\partial t} = -\oint_{\partial V} \rho \mathbf{u} \cdot \mathbf{n} dS, \tag{8.13}$$

where ∂V stands for the entire surface of V, dS a small surface area whose unit normal vector is $\mathbf{n}$, and the vector $\mathbf{u}$ is the fluid velocity observed by a fixed instrument. The negative sign of the right hand side is due to that $\mathbf{u} \cdot \mathbf{n} dS$ indicates the flow from inside to outside of the box through the area dS. The reverse flow from outside to inside needs a negative sign.

The divergence theorem for a 3D surface integral transforms the right hand side of the equation to

$$-\oint_{\partial V} \rho \mathbf{u} \cdot \mathbf{n} dS = -\iiint_V \nabla \cdot (\rho \mathbf{u}) dV, \tag{8.14}$$

where

$$\nabla \cdot (\rho \mathbf{u}) = \frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y} + \frac{\partial(\rho w)}{\partial z} \tag{8.15}$$

is called the divergence of vector $\mathbf{u}$, dV means the small volume cut inside V, and u is the velocity vector $\mathbf{u}$'s projection component in x-axis, v in y-axis, and w in z-axis.

When V is small,

$$\iiint_V \nabla \cdot (\rho \mathbf{u}) dV = V \nabla \cdot (\rho \mathbf{u}) \tag{8.16}$$

where $\nabla \cdot \mathbf{u}$ is the divergence at any point inside V. This is an approximation under the condition that both ρ and $\cdot \mathbf{u}$ are continuous in the small V. Thus, V cannot be across a shock, which produces discontinuities.

Therefore, the law of conservation of mass for a fixed observer, also called Eulerian coordinates, is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0. \tag{8.17}$$

This equation is called continuity equation of fluid mechanics, in the Eulerian coordinates. For numerical computing, we write this equation into the xyz-coordinate form

$$\rho_t + u\rho_x + v\rho_y + w\rho_z + \rho(u_x + v_y + w_z) = 0, \tag{8.18}$$

where the subscripts of x, y, z, t means partial derivative with respect to a given variable. This is an equation in the Eulerian framework and the xyz-coordinate system.

Now, the Lagrangian derivative $\frac{D(V\rho)}{Dt}$ is changed to partial derivatives in Eulerian coordinates. An Eulerian observer places his instrument at the point (x, y, z) and observes at time t. Thus, a climate parameter in the Eulerian coordinates is a function of space and time, and is a function of time only under the Lagrangian coordinates for a particular small box to start with. The Lagrangian coordinate frame is moving with the box, like a dog (i.e., the fluid box) and its walker (i.e., the Lagrangian observer).

Most cases of fluid dynamics use Eulerian coordinates.

Thus, the Lagrangian coordinates are a dog walker, and the Eulerian coordinates are a street on-looker watching the dog and its walker passing by.

EXAMPLE 8.1

A 1D flow in $x \in (1, 4)$ and $t \in (0, \infty)$ is defined by

$$\rho = 0.5(1 + 2t)x, \tag{8.19}$$

$$u = \frac{-x}{1 + 2t} \tag{8.20}$$

satisfies the mass conservation equation $\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0$, since

$$\frac{\partial \rho}{\partial t} = x, \tag{8.21}$$

$$\rho u = -0.5x^2, \tag{8.22}$$

$$\frac{\partial}{\partial x}(\rho u) = -x, \tag{8.23}$$

and

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = x + (-x) = 0. \tag{8.24}$$

This is a pipe flow that is compressed from right to the left.

The divergence theorem was used above as given, although the explanation of flux is sufficient for a physicist. One can be more mathematically rigorous in deriving the divergence theorem in 1D case and then extend it to 3D case as a formality. Again, this 3D extension is not a proof. Rigorously speaking, a theorem holds in 1D may not hold in 3D.

One may wish to watch the classical film of John Lumley's lecture on "Eulerian and Lagrangian Descriptions in Fluid Mechanics" produced in 1968 and supported by the National Science Foundation via Youtube

`https://www.youtube.com/watch?v=mdN8OOkx2ko`

or a library.

In the 1D case, consider two people holding a rubber band at each end. The rubber band's length is h. The rubber band's linear density [Units: Kg/m] decreases if the

band gets stretched. A way to stretch the band is to have the right end move forward faster than the left end. We then have the following balance of mass. The band density decrease rate is $-\partial\rho/\partial t$ where ρ is the linear density of the band and is a function of both location and time: $\rho(x,t)$. The front end of the band stretches, causing the mass decrease at the rate $\rho(x+h,t)u(x+h,t)$ in $(x,x+h)$, while the rear end's compression causes the mass increase in $(x,x+h)$ at the rate $\rho(x,t)u(x,t)$. The mass conservation is

$$\begin{aligned} & [\texttt{mass/decrease}]\ (-\partial(h\rho)/\partial t) \\ = \ & \texttt{mass/decrease/from/the/front/end}]\ (\rho(x+h,t)u(x+h,t)) \\ & -[\texttt{mass/increase/from/the/rear/end}]\ (\rho(x,t)u(x,t)). \end{aligned} \tag{8.25}$$

This can be simplified as

$$\frac{\partial\rho}{\partial t}+\frac{\rho(x+h,t)u(x+h,t)-\rho(x,t)u(x,t)}{h}=0. \tag{8.26}$$

The linear approximations of $\rho(x+h,t)=\rho(x,t)+\rho_x(x,t)h$ and $u(x+h,t)=u(x,t)+u_x(x,t)h$ for a small h lead to the 1D mass conservation equation, when h^2 term is ignored:

$$\frac{\partial\rho}{\partial t}+\rho\frac{\partial u}{\partial x}+u\frac{\partial\rho}{\partial x}=0, \tag{8.27}$$

or

$$\frac{\partial\rho}{\partial t}+\frac{\partial(\rho u)}{\partial x}=0, \tag{8.28}$$

In the above, $u_x=\partial u/\partial x$ is another notation of partial derivatives.

Extensions of this equation to 2D in xy-plane and to 3D in xyz-space are below

$$\frac{\partial\rho}{\partial t}+\frac{\partial(\rho u)}{\partial x}+\frac{\partial(\rho v)}{\partial y} = 0 \tag{8.29}$$

$$\frac{\partial\rho}{\partial t}+\frac{\partial(\rho u)}{\partial x}+\frac{\partial(\rho v)}{\partial y}+\frac{(\rho w)}{\partial z} = 0. \tag{8.30}$$

The three equations above are the mass conservation equations in Eulerian coordinates. They have a compact form

$$\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\mathbf{u})=0, \tag{8.31}$$

which is the previous mass conservation equation (8.17). This procedure gives an approach to the rigorous mathematically proof for the 3D mass conservation equation, although our statement above is not really a proof, strictly speaking in the mathematical sense.

The two terms in the mass conservation equation have their physical meanings. The first term $\partial\rho/\partial t$ means the density increase in time, expansion or compression right at the location of the Eulerian observer. A traffic light Eulerian observer sees a density increase in front of its red signal and density decrease in front of its green signal following a red signal. This is called the local derivative: $\partial\rho/\partial t$.

The second term is caused by the speed differences in a fluid flow. If the flow ahead is slower than the flow behind, then the fluid piles up and increases the density.

Otherwise, the density decreases. This part of mass change is accomplished by fluid motion and is called mass advection.

In general, a Lagrangian derivative is equal to local derivative plus the mass advection

$$\frac{D\rho}{Dt}, \tag{8.32}$$

or called material derivative, or substantial derivative. It is a derivative in Lagrangian coordinates.

A material derivative in a flow always has two terms: local derivative and mass convection. Thus, the following formula is universal, disregard the conservation:

$$\frac{D\rho}{Dt} = \frac{\partial \rho}{\partial t} + (\mathbf{u} \cdot \nabla)\rho. \tag{8.33}$$

The mass conservation equation, also called continuity equation, (8.17), can be rewritten as

$$\begin{aligned} & \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) \\ = & \left[\frac{\partial \rho}{\partial t} + (\mathbf{u} \cdot \nabla)\rho\right] + \rho \nabla \cdot \mathbf{u} \\ = & \frac{D\rho}{Dt} + \rho \nabla \cdot \mathbf{u} = 0. \end{aligned} \tag{8.34}$$

The part in the square bracket is the Lagrangian derivative of ρ, and the second term is the divergence times density.

When a fluid is incompressible, the density of a box fluid does not change from the beginning to the end, thus

$$\frac{D\rho}{Dt} = 0. \tag{8.35}$$

Equivalently,

$$\nabla \cdot \mathbf{u} = 0, \tag{8.36}$$

i.e., a divergence free fluid flow. The divergence theorem for an incompressible fluid flow implies that the flow flux (not the mass flux) through any closed surface must be zero:

$$\oint_{\partial\Omega} \mathbf{u} \cdot \mathbf{n} dS = 0. \tag{8.37}$$

EXERCISES

8.1 A pipe flow shown in Figure 8.1 is considered a 1D flow. The flow speed, density, and cross-section area in section i is u_i, ρ_i and A_i for $i = 1, 2$. Find a relationship between u_1 and u_2.

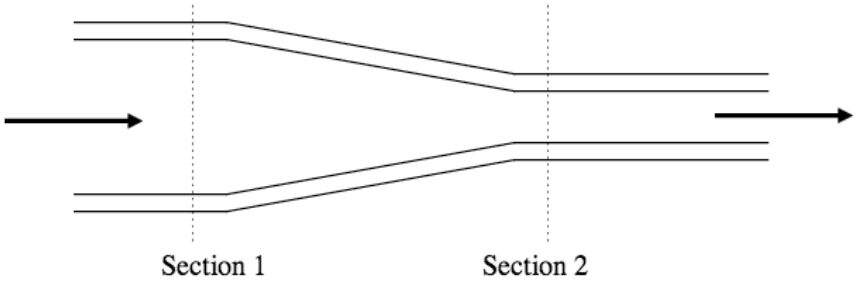


Figure 8.1 A stationary time-independent 1D flow.

8.2 Conservation of momentum over a grid box: $\mathbf{F} = m\mathbf{a}$

The momentum equation $\mathbf{F} = m\mathbf{a}$ can be applied to a small Lagrangian fluid box of volume V and average density ρ:

$$m\frac{D\mathbf{u}}{Dt} = \mathbf{F}_c + \mathbf{F}_p + \mathbf{F}_g + \mathbf{F}_f, \tag{8.38}$$

where

$$m = \rho V \tag{8.39}$$

is the mass of the Lagrangian fluid box,

$$\frac{D\mathbf{u}}{Dt} = \mathbf{a} \tag{8.40}$$

is the Lagrangian box's acceleration, and $\mathbf{F}_c, \mathbf{F}_p, \mathbf{F}_g, \mathbf{F}_f$ are Coriolis force (CF), pressure gradient force (PGF), Earth gravity force (EGF), and friction force (FF).

The Coriolis force is a force in a rotational platform, which is Earth in our case and deflects the direction of a particle on this rotating platform. The force is perpendicular to both the angular velocity of the rotating platform and the particle's velocity. This force can be expressed by cross product of two vectors:

$$\mathbf{F}_c = -2m\mathbf{\Omega} \times \mathbf{u}, \tag{8.41}$$

where $\mathbf{\Omega}$ is the angular velocity of the platform at the location of interest. The Earth's Coriolis force varies at different latitude. On the Northern Hemisphere, the Coriolis force drags a flow to the right of the main flow direction. For example, a centrifugal force on a rotating plate draws a ball on the plate outward. The Coriolis force deflects the outward moving ball to the right when one faces the outward direction, and thus result on a non-straight line trajectory.

The pressure gradient force was discussed in the last chapter and points from higher pressure region to a lower pressure region:

$$\mathbf{F}_p = -V\nabla p, \tag{8.42}$$

where p is the pressure field.

In the large scale of atmospheric motion of a hurricane, the balance of the PGF and CF plays critical role in steering the wind direction spinning counterclockwise around a low pressure center on NH, and clockwise around a high pressure center.

The gravitational force is due to the Earth's gravity and is equal to

$$\mathbf{F}_g = m\mathbf{g}, \tag{8.43}$$

where $\mathbf{g}$ is the gravitational acceleration and points toward the Earth center and whose magnitude is $9.8[m/s^2]$.

The friction force can be due to boundary, such as that water flows slows at the bottom of a river, and can also be due to fluid's internal friction, called viscosity. Oil has a large viscosity, while the viscosities for atmosphere and ocean water are usually small. We often ignore the viscosity in the atmospheric and oceanic dynamics.

Ignoring the friction force $\mathbf{F}_f$, the Newton's second law of motion now becomes

$$m\frac{D\mathbf{u}}{Dt} = -2m\mathbf{\Omega} \times \mathbf{u} - V\nabla p + m\mathbf{g}. \tag{8.44}$$

Dividing both sides by m, we have the acceleration equation

$$\frac{D\mathbf{u}}{Dt} = -2\mathbf{\Omega} \times \mathbf{u} - \frac{\nabla p}{\rho} + \mathbf{g}. \tag{8.45}$$

When we expand the material derivative $\frac{D\mathbf{u}}{Dt}$ and express the above only in the Eulerian framework, we have the following momentum equation in the form of $a = F/m$ $[dimension : LT^{-2}]$:

$$\frac{\partial \mathbf{u}}{\partial t} = -(\mathbf{u} \cdot \nabla)\mathbf{u} - 2\mathbf{\Omega} \times \mathbf{u} - \frac{\nabla p}{\rho} + \mathbf{g}. \tag{8.46}$$

The first term on the right hand side is the adjective acceleration, or called velocity advection which behaves as an acceleration.

8.3 The momentum equation in x, y, z coordinates

As Eulerian observers on Earth to measure the atmospheric and oceanic flows, we often set our Cartesian coordinates's z pointing directly to the sky (see Figure 8.2), a vector normal to the Earth surface, or perpendicular to the tangent plane of the spherical Earth at our observation location. Thus, the z directional vector is in the direction pointing from the Earth center outward to our location. The x-coordinate is along the zonal direction pointing from west to east. The y-coordinate points along the meridional line toward the north pole.

Under this xyz-coordinate system, the Earth angular velocity $\mathbf{\Omega}$ in the momentum equation (8.46) has three components, which can be calculated from $\mathbf{\Omega}$'s projections on the three axes.

$$(0, \Omega\cos\phi, \Omega\sin\phi), \tag{8.47}$$

where $\Omega = ||\mathbf{\Omega}||$ is the magnitude of $\mathbf{\Omega}$, and ϕ is the latitude.

The Earth rotation speed is a constant with

$$\begin{aligned} \Omega &= (7.2921150 \pm 0.0000001) \times 10^{-5}\ [radians/second] \\ &= (72.921150 \pm 0.000001)\ [microradians/second]. \end{aligned} \tag{8.48}$$

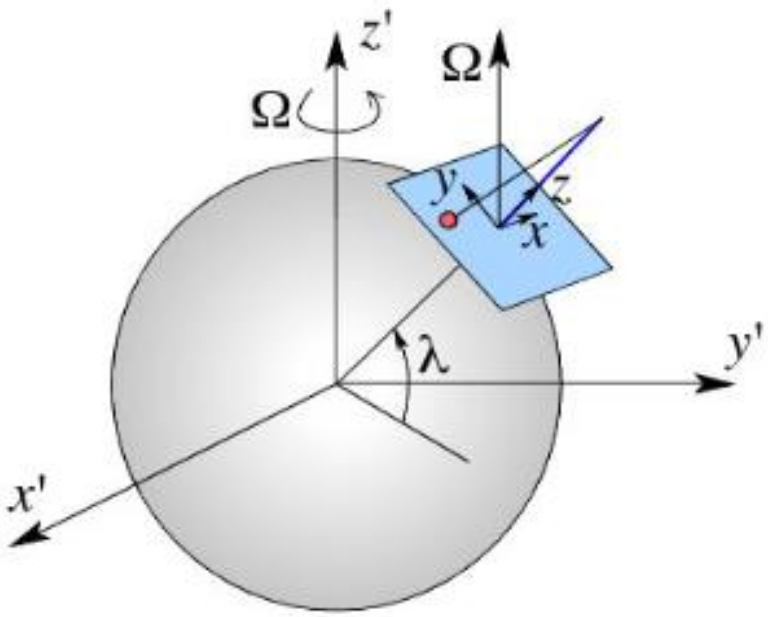


Figure 8.2 The Cartesian coordinate x, y, z-system on the northern hemisphere surface.

The major spins of atmosphere are rotating around the z-direction, such as hurricanes, tornados, and typhoons. The z-component $\Omega \sin\phi$ of the angular velocity thus determines the rotation direction of the atmosphere. Because $\sin\phi$ changes sign from positive in NH to negative in SH, the hurricanes on NH and SH have opposite spin directions: counterclockwise on NH, and clockwise on SH.

The Coriolis force can be calculated from its definition of cross product of the Earth angular velocity $\mathbf{\Omega}$ and the flow velocity $\mathbf{u}$:

$$\mathbf{F}_c = -2\Omega \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & \cos\phi & \sin\phi \\ u & v & w \end{bmatrix} \tag{8.49}$$

This determinant yields the following x, y, z-component expression of the Coriolis force per unit mass:

$$\mathbf{F}_c/m = 2\Omega \left[\mathbf{i}(-w\cos\phi + v\sin\phi) - \mathbf{j}u\sin\phi + \mathbf{k}u\cos\phi\right] \tag{8.50}$$

The pressure gradient force (PGF) per unit mass can also be written explicitly in terms of x, y, z-components:

$$\mathbf{F}_p/m = -\frac{1}{\rho}\left[\mathbf{i}p_x + \mathbf{j}p_y + \mathbf{k}p_z\right]. \tag{8.51}$$

The gravitational force per unit mass is

$$\mathbf{F}_g/m = -\mathbf{k}g. \tag{8.52}$$

The friction force is complicated and can be due to boundary friction or internal friction of viscosity. This book does not discuss it in detail. We just keep the generic notation:

$$\mathbf{F}_f/m = \mathbf{i}f_{(rx)} + \mathbf{j}f_{(ry)} + \mathbf{k}f_{(rz)}. \tag{8.53}$$

Finally, the momentum equation in Eulerian frame and in xyz-components is below:

$$\begin{aligned} u_t &= -(uu_x + vu_y + wu_z) - 2\Omega(w\cos\phi - v\sin\phi) - p_x/\rho + f_{(rx)}, && (8.54)\\ v_t &= -(uv_x + vv_y + wv_z) - 2\Omega u\sin\phi) - p_y/\rho + f_{(ry)}, && (8.55)\\ w_t &= -(uw_x + vw_y + ww_z) - 2\Omega u\cos\phi - p_z/\rho - g + f_{(rz)}, && (8.56) \end{aligned}$$

where the subscripts of x, y, z, t means the local partial derivative with the assigned variable.

Eulerian observers can in theory measure every quantity in the above equations, except the friction force. The nonlinearity of the equations are from the advective velocities, the quantities in the round brackets.

These momentum equations and the continuity equation (8.18) form the four basic equations to describe the atmospheric and oceanic flows. Due to the numerous varieties of atmospheric and oceanic flows in different spatiotemporal scales, solving the four partial differential equations (PDEs) is a very challenging task. Users then turn to special cases with given conditions.

Two special cases can yield analytic solutions, which can explain many atmospheric and oceanic phenomena:

(a). Geostrophic wind as a balance of the pressure gradient force and the Coriolis force, and

(b). Large-scale oceanic gyres and coastal currents due to the conservation of vorticity.

The following two sections will describe the special approximations of the above four equations, the analytic solution of the approximate equations, and applications to explaining and quantifying the above two atmospheric and oceanic phenomena.

8.4 Geostrophic approximation of the momentum equations, and geostrophic wind

Geostrophic approximation is a simplification of the momentum equations under the following assumptions

(i). The acceleration $D\mathbf{u}/Dt$ is small and can be ignored, and thus it is for a relatively stationary climate system and is not in a fast transition state;

(ii). The vertical velocity w is much smaller than the horizontal velocity and can be ignored;

(iii). The friction force is small and ignored; and

(iv). The gravity force is balanced by the vertical pressure gradient and the entire equation for the vertical acceleration can be ignored.

Under these assumptions, the momentum equations, the three momentum equations (8.54)-(8.56) are reduced to the following two equations

$$2\Omega v \sin\phi - p_x/\rho = 0, \tag{8.57}$$
$$-2\Omega u \sin\phi - p_y/\rho = 0. \tag{8.58}$$

These become the balance between the Coriolis force and the pressure gradient force. Let

$$f = 2\Omega \sin\phi = 146 \sin\phi, \tag{8.59}$$

which is twice the projection of the Earth self-spin's angular velocity on the local vertical coordinate z at latitude ϕ. Equations (8.57) and (8.58) lead to the expression of the horizontal wind velocity in terms of pressure gradient and Coriolis force.

$$u = -p_y/(\rho f), \tag{8.60}$$
$$v = p_x/(\rho f). \tag{8.61}$$

Thus, the wind direction (u, v) and the pressure gradient direction (p_x, p_y) are perpendicular since

$$(u, v) \cdot (p_x, p_y) = (1/(\rho f))(-p_y, p_x) \cdot (p_x, p_y) = (1/(\rho f))(p_x p_y - p_y p_x) = 0. \tag{8.62}$$

This is counterintuitive. The wind direction is not along the PGF, instead perpendicular to PGF. It is because of this property of atmospheric motion that leads to the large scale rotations of atmosphere, such as hurricanes and tornadoes.

EXAMPLE 8.2

Consider the pressure field shown in Figure 6.5a. The pressure has no gradient in x-direction: $p_x = 0$. The pressure gradient in y-direction is negative, thus PGF points to north. Thus

$$u = -p_y > 0, v = p_x = 0. \tag{8.63}$$

So the wind is along the x-axis. The Coriolis' force $F_c = -2\Omega \times \mathbf{u}$ is thus perpendicular to the wind velocity $\mathbf{u}$ and points to south and balance the PGF.

This example explains the global scale trade wind direction due to the sub-tropical high pressure around the latitude band of 30-40○N and the ITCZ tropical low pressure. This sub-tropical high to the ITCZ low PGF causes the Northeast Trade Winds, which blows from the east to the west globally between equator and the northern Hemisphere sub-tropical zone.

For the same reason, the North Pole's high pressure and the polar front low pressure zone around 60-70○N form the polar Easterlies, which blows from the west to east.

The interaction of the polar front low and the sub-tropical high forms a PDF pointing to north, which results in a wind flows from the west to the east. This is the westerlies in the latitude zone of 40-70○N. This wind slows down the international airliners flying from San Francisco to Beijing via the Alaska route because the westerlies along the flight route are the head wind for an air plane travels from the east (San Francisco) to the west (Beijing). The return flight from Beijing to San Francisco has the westerlies as the tail wind and hence can flow faster. The Beijing-San Francisco flight (11.7 hours) is one hour shorter than the San Francisco-Beijing flight (12.5 hours) for Boeing 747.

8.5 Western boundary current vorticity, vorticity conservation, and Ekman layer

8.5.1 Ocean boundary current and its vorticity

When the two ends of a rod move at a different speed, the rod will spin. The spin in atmospheric or oceanic mass parcels is atmospheric and oceanic vorticities, caused by the shearing flow speed. The large scale vorticities have two sources: Eath's rotation, and the speed difference. The latter is called local vorticity and is mathematically expressed as

$$\zeta = \nabla \times \mathbf{u}. \tag{8.64}$$

Here, $\mathbf{u} = (u, v, w)$ is the velocity vector. The cross product is defined as

$$\nabla \times \mathbf{u} = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ u & v & w \end{bmatrix} \tag{8.65}$$

Around the vertical z-axis that points from the Earth's center to the sky, the local vorticity component is

$$\zeta^{(z)} = v_x - u_y. \tag{8.66}$$

The tree leaves in a river and near the river bank may spin very fast due to the flow speed difference: the water flows faster in the middle of the river, and slower near the river banks. Figure 8.3 For this ideal model of river flow, we have

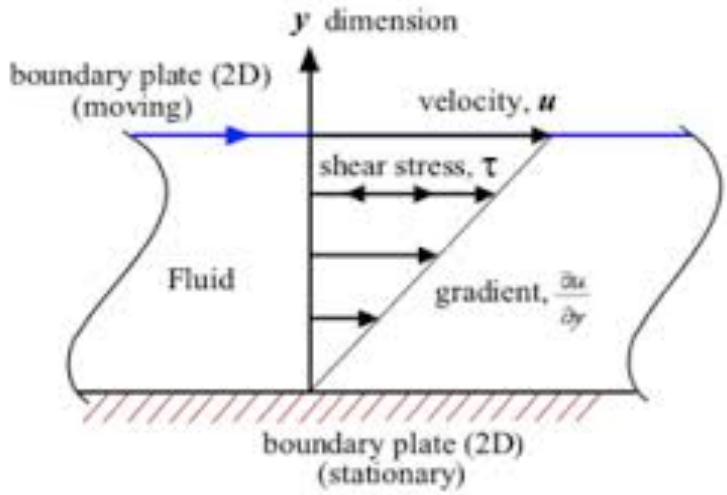


Figure 8.3 River flow with shear speed (This figure needs to be replotted with a leave).

$$u = u_0(1 + y/b), \tag{8.67}$$

$$v = 0 \tag{8.68}$$

where b is the width of the river.

The local vertical vorticity is

$$\zeta^{(z)} = v_x - u_y = -u_0/b, \tag{8.69}$$

is a clockwise spin when u_0 is positive. The river's opposite side will support a leave that spins in the opposite direction: a counterclockwise local vorticity.

For this localized small scale flow, the Earth rotation plays almost no role. Coriolis force does not need to be considered. However, for a large scale oceanic current, such as the California Coastal Current (CCC) and the North Pacific Gyre, the Earth rotation and Coriolis force play a critical role. Figure 8.1.2 shows the Florida warm current and the Gulf stream along the east coast of the United States. Cape Hatteras (35.3°N, 75.5°W), is on the North Carolina coast with the Gulf steam current observations: The flow speed is increased by 1.0 m/sec for every 100 [km] away from the coast. When

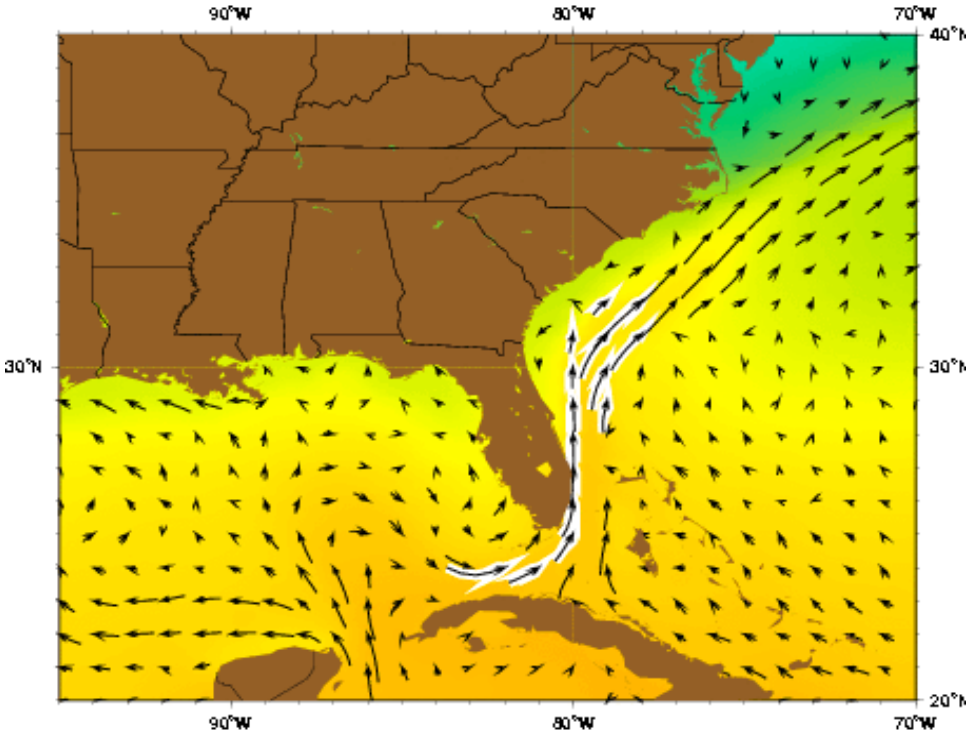


Figure 8.4 The Atlantic Gulf stream near the United States' east coast. (This figure needs to be revised with more features).

we set the y-axis pointing north along the meridional direction, and x-axis pointing away from the coast to the east in the Atlantic ocean along the zonal direction, the velocity field is approximately as a function of x:

$$u = u(x)\ (very\ small), \quad v = v(x) = v_0(1 + x/L), \tag{8.70}$$

where v_0 is the coast current speed and L is the length scale associated with this current. The observational data of the flow speed increment by 1.0 m/sec for every 100 [km] for v determines

$$v_0/L = \frac{1}{100 \times 1000}[(m/s)m^{-1} = 1 \times 10^{-5}\ [1/s]. \tag{8.71}$$

The local vorticity is

$$\zeta^{(z)} = v_x - u_y = v_0/L = 1 \times 10^{-5}\ [radian/sec]. \tag{8.72}$$

Again, radian measures angle and is dimensionless: $[radian] = [1]$.

The Earth rotation effect at Cape Hatteras (35.3°N) is

$$f = 2\Omega \sin\phi = 2 \times 7.3 \times \times 10^{-5} \times \sin(35.3^\circ) = 8.4 \times 10^{-5}\ [radian/sec]. \tag{8.73}$$

In the Eulerian xyz-coordinate frame, the most important spin is the rotation around the z-axis, which forms a vortex spinning around the z-axis like a hurricane. Section 8.3 shows that the Earth's self-rotation has a projection to the z-axis, which is

$$\Omega_z = \Omega \sin\phi, \tag{8.74}$$

where ϕ is the latitude, and Ω is the Earth's angular speed equal to $73[microradian/sec]$. It is the Coriolis force produced by this component of the Earth rotation that causes the right turn of the wind from a higher pressure to a lower pressure on NH. This right turn causes the geostrophic wind, which travels around the low pressure center counterclockwise on NH, and forms a counterclockwise vortex. Section 8.4 shows that the geostrophic wind speed inversely proportional to $2\Omega \sin\phi$ or $2\Omega_z$. This quantity is usually called the planetary vorticity f:

$$f = 2\Omega \sin\phi = 73 \times 10^{-6} \sin\phi\ [radian/sec] = 73 \sin\phi\ [microradian/sec]. \tag{8.75}$$

This is also called an initial angular momentum. The pressure gradient is pushing this initial momentum and producing vorticity like a hurricane spin.

Thus, the ratio of the planetary vorticity to the local vorticity at Cape Hettera is

$$f/\zeta = 8. \tag{8.76}$$

The local vorticity of the Gulf stream is only about 12% of the global Coriolis effect at Cape Hatteras. So, the Gulf stream is dominated by the global Coriolis effect, which forms the North Atlantic Gyre (see Figure 8.3).

The absolute vorticity is the sum of the local vorticity and the inertial angular momentum:

$$\zeta_a = \zeta + f, \tag{8.77}$$

which is dominated by f except where ocean current's velocity gradient is very strong, such as the collision at the coast of South Africa of the cold Benguela Current from the polar Atlantic and the warm Agulhas Stream from Indian ocean. The collision of this current causes complex and large spatial changes of speed and hence large local vorticities, which generate rogue waves or called freak waves that are navigation hazards of wreaking ships. The current speed can change from 5-8 m/s to near zero m/s in a short range of 30-60 km near the Cape of Good Hope (34.3581°S, 18.4719°E).

Thus, the maximum vorticity can reach $8/(30 \times 1000) = 27 \times 10^{-5}[radian/sec]$. The planetary vorticity is

$$f = 2\Omega \sin\phi = 2 \times 7.3 \times 10^{-5} \times \sin(-34.4^{\circ}) = 8.2 \times 10^{-5}\ [radian/sec]. \quad (8.78)$$

Thus, the local vorticity $27 \times 10^{-5}[radian/sec]$ is 3 times larger than the planetary vorticity $8.2 \times 10^{-5}\ [radian/sec]$. The local vorticity at the Cape of Good Hope, South Africa is 27 times of that $1.0 \times 10^{-5}\ [radian/sec]$ of Cape Hetteras, North Carolina. These are huge local oceanic vorticity anomalies, causing many notorious disasters in the navigation history.

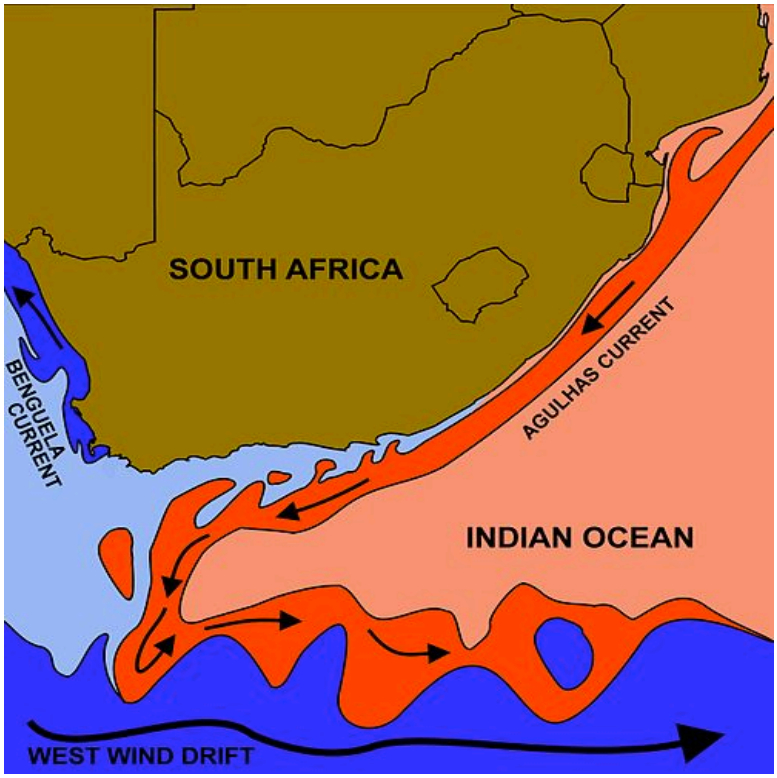


Figure 8.5 Extremely large local vorticity at the Cape of Good Hope, South Africa due to the collision of the cold Benguela Current from the polar Atlantic and the warm Agulhas Stream from Indian ocean.

8.5.2 Vorticity conservation for oceanic current

The vorticity conservation principle of ocean flows states that the potential vorticity

$$\zeta_p = \frac{f + \zeta}{H} \quad (8.79)$$

is a constant in the path of the ocean water mass with depth H and absolute vorticity $f + \zeta$.

If a water mass moves from location P_1 to another location P_2, then we have $\zeta_{p1} = \zeta_{p2}$:

$$\frac{f_1 + \zeta_1}{H_1} = \frac{f_2 + \zeta_2}{H_2}. \tag{8.80}$$

EXAMPLE 8.3

A water mass moves from middle of an ocean west to a content shelf: Consider the North Equatorial Current from the middle of the Tropical Pacific directly west to South China Sea, and then to the coast of Vietnam (see Figure 8-7). During this process of current motion, the depth changes from 5,000 meters around East Mariana Basin to 1,000 meters in the Philippine Basin, and even shallower near the coast of Vietnam. The local vorticity in the middle of Pacific is small and can be regarded as zero. Since it is along the same latitude, the inertial angular momentum does not change. We thus have the following data:

$$\begin{aligned} &H_1 = 5,000[m], \quad H_2 = 1,000[m], \ \ \zeta_1 = 0[1/s], \\ &f_1 = f_2 = 2 \times 7.3 \times 10^{-5} \times \sin(10^\circ) = 2.5 \times 10^{-5}[1/s]. \end{aligned} \tag{8.81}$$

The local vorticity over the Philippine Basin will be

$$\begin{aligned} \zeta_2 &= \frac{H_2}{H_1}(f_1 + \zeta_1) - f_2 = (r_h - 1)f \\ &= -0.8f = -0.8 \times 2.5 \times 10^{-5}[1/s] = -2.0 \times 10^{-5}[1/s], \end{aligned} \tag{8.82}$$

where

$$r_h = \frac{H_2}{H_1} = 1/5 = 0.2 \tag{8.83}$$

is the ratio of the water depths. This negative local vorticity reduces the absolute vorticity, and helps reduce the navigation accidents.

Since

$$\zeta_2 = v_x - u_y = -2.0 \times 10^{-5}[1/s]. \tag{8.84}$$

and since v_x must be positive because the meridional speed is larger when further away from the coast, the negative local vorticity value must be contributed from $-u_y < 0$, and further $u_y = 2.0 \times 10^{-5}[1/s] + v_x$ and $u_y > v_x \geq 0$. This means that the zonal speed (which is already negative because of moving to the west) is an increasing function of y.

EXAMPLE 8.4

A water mass moves along the Gulf Current: Consider a water mass moving along the Gulf Current over the west boundary of the North Atlantic (see Figure 8-8). The depth of the west coast of the Atlantic varies from zero at the shore to 5,000 meters in the mid-Atlantic. The Gulf Current is mainly along the depth of 2,000-3,000 meters from a southern point P_3 to a northern point P_4. The vorticity conservation is

$$\frac{f_3 + \zeta_3}{H} = \frac{f_4 + \zeta_4}{H}. \tag{8.85}$$

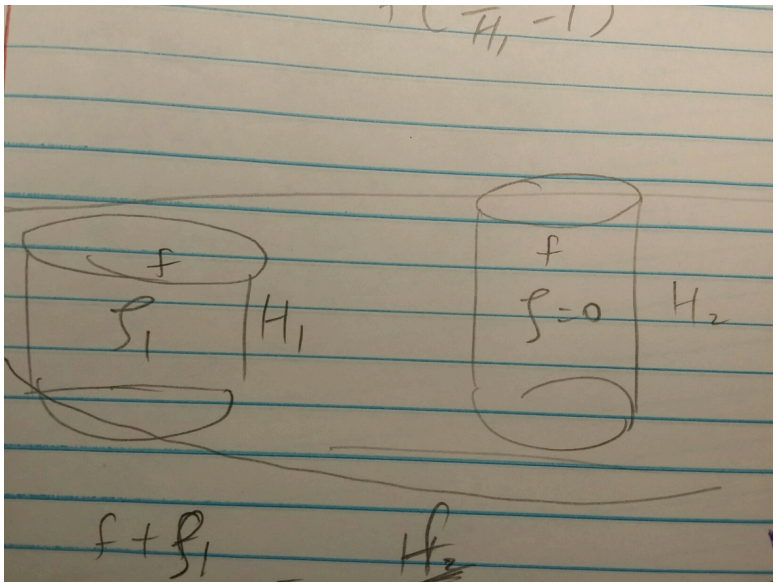


Figure 8.6 Conservation of ocean flow vortices as a water mass moves to middle Pacific to the western Pacific boundary.

The local vorticity at a northern location is

$$\zeta_4 = (f_3 - f_4) + \zeta_3 = \zeta_3 + 14.6 \times 10^{-5} \times (\sin\phi_3 - \sin\phi_4). \tag{8.86}$$

Since $\phi_4 > \phi_3$ due to the move to north, thus

$$\sin\phi_3 - \sin\phi_4 < 0 \tag{8.87}$$

and

$$\zeta_4 < \zeta_3. \tag{8.88}$$

Thus, the local vorticity decreases as the water mass moves to the north along the Gulf Current. When the u-speed is close to zero, the v-speed's gradient is decreasing as the current moves to the north.

See Figure 8.10 for more vortex conservation examples.

The geostrophic approximation of geophysical fluid dynamics equations based on the four assumptions described in Section 8.4 is very important in physical oceanography and can help explain many atmospheric and oceanic motions, such as the directions of trade wind and hurricane wind.

The second most important approximation following the geostrophic approximation is the above vorticity approximation, which helps explain many important oceanic flow phenomena, such as the two examples above, boundary current, and Ekman layer. This simple and beautiful vorticity conservation equation was due to Carl-Gustaf Rossby (1898-1957).

The following will provide assumptions and derivation of the vorticity conservation. The assumptions are:

(i). The friction force is small and ignored;

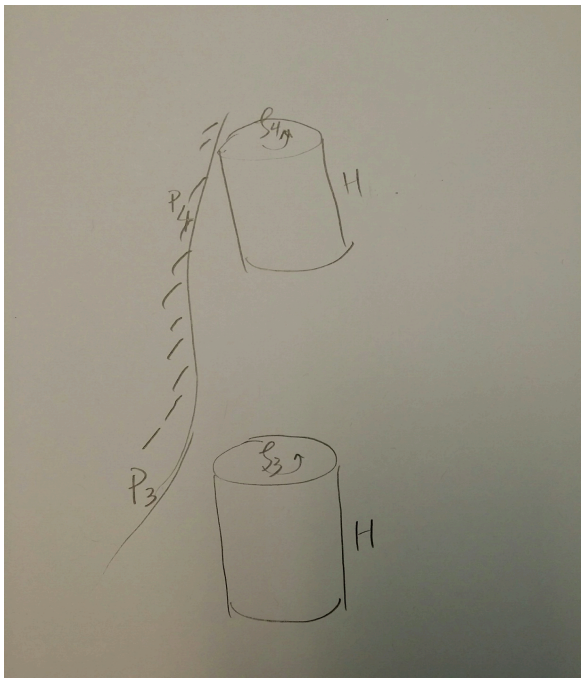


Figure 8.7 Conservation of ocean flow vortices as a water mass moves north along the Gulf Current over the western boundary of Atlantic.

(ii). The gravity force is balanced by the vertical pressure gradient, the entire equation for the vertical acceleration Dw/Dt can be ignored, and thus the momentum equation will be only on the horizontal plane;

(iii) The vertical velocity w is small compared with u and v and can be ignored when appears in a formula with u and v (but w cannot be ignored if not together with u and v);

(iv). However, in the continuity equation, the term $\partial w/\partial z$ cannot be ignored.

Therefore, the assumptions for the vorticity conservations are different from that of geostrophic approximation in (a) the horizontal acceleration needs to be included, and (b) 3-Dim continuity equation needs to be used.

The above assumptions imply the follow momentum equations

$$u_t = -(uu_x + vu_y) - fv - p_x/\rho, , \tag{8.89}$$
$$v_t = -(uv_x + vv_y) + fu - p_y/\rho. \tag{8.90}$$

These can be written as

$$u_t + (uu_x + vu_y) + fv = -p_x/\rho, , \tag{8.91}$$
$$v_t + (uv_x + vv_y) - fu = -p_y/\rho. \tag{8.92}$$

From these two equations, we can compute vorticity: $\zeta^{(z)} = v_x - u_y$. Taking $\partial/\partial x$ for equation (8.92) minus $\partial/\partial y$ for equation (8.91) leads to the following

$$(v_x - u_y)_t + u(v_x - u_y)_x + v(v_x - u_y)_y + v_x(u_x + v_y) - u_y(u_x + v_y) + f(u_x + v_y) = 0. \tag{8.93}$$

This simplifies to

$$\zeta_t + u\zeta_x + v\zeta_y + (\zeta + f)(u_x + v_y) = 0. \tag{8.94}$$

The continuity equation for impressible ocean water flow

$$u_x + v_y + w_z = 0 \tag{8.95}$$

leads to

$$\zeta_t + u\zeta_x + v\zeta_y - (\zeta + f)w_z = 0, \tag{8.96}$$

or

$$\frac{D}{Dt}(\zeta + f) - (\zeta + f)w_z = 0. \tag{8.97}$$

The z derivative hints us to integrate this equation in the ocean water domain shown in Figure 8-9 from the bottom $z = b(x, y)$ to the surface $z = b(x, y) + H(x, y, t)$. This integration yields

$$\frac{D}{Dt}(\zeta + f)H + (\zeta + f)\left[w(x, y, b, t) - w(x, y, z, b + H, t)\right] = 0. \tag{8.98}$$

In the integration, we have assumed that the absolute vorticity $\zeta + f$ does not depend on z and is only a function of x, y, t.

The kinematic conditions for the bottom and the free surface are that the fluid particle at the boundary remains at the boundary:

$$\frac{D}{Dt}(z - b) = 0 \quad (ocean\ bottom) \tag{8.99}$$

$$\frac{D}{Dt}[z - (b + H)] = 0 \quad (sea\ surface). \tag{8.100}$$

By definition of material derivative, we have

$$\frac{\partial}{\partial t}(z - b) + u\frac{\partial}{\partial x}(z - b) + v\frac{\partial}{\partial y}(z - b) = 0 \tag{8.101}$$

$$\frac{\partial}{\partial t}[z - (b + H)] + u\frac{\partial}{\partial x}[z - (b + H)] + v\frac{\partial}{\partial y}[z - (b + H)] = 0. \tag{8.102}$$

These can be simplified to

$$w(x, y, b, t) = ub_x + vb_y \tag{8.103}$$

$$w(x, y, b + H, t) = H_t + u(b + H)_x + v(b + H)_y, \tag{8.104}$$

where $\partial z/\partial t = w, \partial b/\partial t = 0, \partial z/\partial x = 0, \partial z/\partial y = 0$. The first equation above minus the second yields

$$\begin{aligned} & w(x, y, b + H, t) - w(x, y, b, t) \\ = \ & H_t + u(b + H)_x + v(b + H)_y - (ub_x + vb_y) \\ = \ & -(H_t + uH_x + vH_y) \\ = \ & -\frac{DH}{Dt}. \end{aligned} \tag{8.105}$$

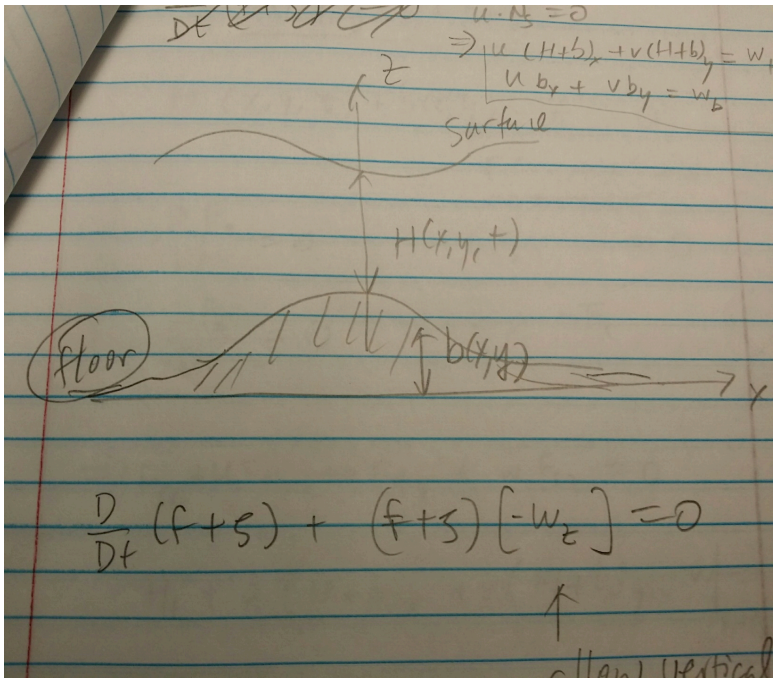


Figure 8.8 The ocean water domain for the derivation of vorticity conservation equation: from the 3-dim ocean bottom to the 3-dim free surface.

Let

$$\zeta^{(a)} = \zeta + f \tag{8.106}$$

denote the absolute vorticity, or called total vorticity.

From the above two formulas, equation (8.98) becomes

$$\frac{D\zeta^{(a)}}{Dt} H - \zeta^{(a)} \frac{DH}{Dt} = 0. \tag{8.107}$$

The quotient rule of differentiation

$$\left(\frac{f}{g}\right)_x = \frac{f_x g - g_x f}{g^2} \tag{8.108}$$

is applicable to material derivative D/Dt. Dividing equation (8.107) by H^2 leads to

$$\frac{\frac{D\zeta^{(a)}}{Dt} H - \zeta^{(a)} \frac{DH}{Dt}}{H^2} = \frac{D}{Dt}\left(\frac{\zeta^{(a)}}{H}\right) = 0. \tag{8.109}$$

Define

$$\zeta^{(p)} = \frac{\zeta^{(a)}}{H} \tag{8.110}$$

as the potential vortiticity. Then the vorticity conservation theorem is that

$$\frac{D\zeta^{(p)}}{Dt} = 0 \tag{8.111}$$

Namely, the potential vorticity of a water parcel is a constant along its path of motion.

In the two examples we listed as the beginning of this section, the water parcels are a cylinder column of the ocean water.

Since the local vorticity's units is $[radian/sec]$ (dimension T^{-1}), the potential vorticity's units is thus $[radian/(sec \cdot meter)]$ (dimension $L^{-1}T^{-1}$).

However, the material derivative's quotient rule is not obvious and needs a proof, which is below:

$$\begin{aligned}
\frac{D}{Dt}\left(\frac{\zeta^{(a)}}{H}\right) &= \left(\frac{\zeta^{(a)}}{H}\right)_t + u\left(\frac{\zeta^{(a)}}{H}\right)_x + v\left(\frac{\zeta^{(a)}}{H}\right)_y \\
&= \frac{\zeta_t^{(a)}H - H_t\zeta^{(a)}}{H^2} + u\frac{\zeta_t^{(a)}xH - H_x\zeta^{(a)}}{H^2} + v\frac{\zeta_y^{(a)}H - H_y\zeta^{(a)}}{H^2} \\
&= \frac{1}{H^2}\left[(\zeta_t^{(a)} + u\zeta_x^{(a)} + v\zeta_y^{(a)})H - (H_t + uH_x + vH_y)\zeta^{(a)}\right] \\
&= \frac{\left[H\frac{D\zeta^{(a)}}{Dt} - \zeta^{(a)}\frac{DH}{Dt}\right]}{H^2}. \qquad (8.112)
\end{aligned}$$

Robert Stewart's book on his Texas A&M University's website has a nice proof of this theorem: Chapter 12: Vorticity in the Ocean,

http://oceanworld.tamu.edu/resources/ocng_textbook/chapter12/chapter12_01.htm

Lynn Talley's physical oceanography's lecture notes at the Scripps Institution of Oceanography lists many important applications of this theorem:

http://www-pord.ucsd.edu/~ltalley/sio210/dynamics_sverdrup/

8.5.3 Western boundary currents and oceanic gyres.

Spanish navigators in the 16th century noticed strong northward currents along the Florida coast that seemed to be unrelated to the wind. How can this happen? And, why are strong currents found offshore of east coast of the US but the US west coast?

Let us examine the North Pacific Gyre shown in Figure 8.9. The tropical trade winds blow westward (from eastern side of Pacific to the western side. The mid-latitude westerlies blow eastward. This wind pattern applies a stress to the subtropical ocean surface with negative curl in the North Pacific.

A theory by Vagn W. Ekman (1874-1954) helps explain ocean water circulation driven only by the transfer of momentum from the wind through friction. It is thus a balance between a wind friction force balanced by Coriolis force. Surface currents flow at a 45 angle to the wind due to a balance between the Coriolis force and the drags generated by the wind and the water. If the ocean is divided vertically into thin layers, the magnitude of the velocity (the speed) decreases from a maximum at the surface until it dissipates. The direction also shifts slightly across each subsequent layer, right for the North Pacific Gyre. This is called the Ekman spiral. The layer of water from the surface to the deep end of this spiral is known as the Ekman layer. If all flow over the Ekman layer is integrated, the net transportation is at 90 to the right of the surface wind in North Pacific. See

http://oceanworld.tamu.edu/resources/ocng_textbook/chapter09/chapter09_02.htm

for detailed mathematical derivations.

Harald Sverdrup (1947) showed that the circulation in the upper kilometer or so of the ocean is directly related to the curl of the wind stress. Henry Stommel (1948) showed that the circulation in oceanic gyres is asymmetric because the Coriolis force varies with latitude. Finally, Walter Munk (1950) added eddy viscosity and calculated the circulation of the upper layers of the Pacific. These three papers help explain different ocean gyres and circulations.

Different from Ekman theory which is derived from the balance of wind friction and Coriolis force, Sverdrup theory of transport includes pressure gradient force into account in addition to the friction and Coriolis force. See

`http://oceanworld.tamu.edu/resources/ocng_textbook/chapter11/chapter11_01.htm`

for detailed mathematical derivations.

As a result, the deep ocean Sverdrup transport is equatorward (i.e., southward) while the Ekman pumping transport on surface is downward and northward (See Figure 8.10). Because of conservation of mass and potential vorticity conservation, that transport is balanced by a narrow, intense poleward current, which flows along the western boundary of the ocean basin, allowing the vorticity introduced by coastal friction to balance the vorticity input of the wind.

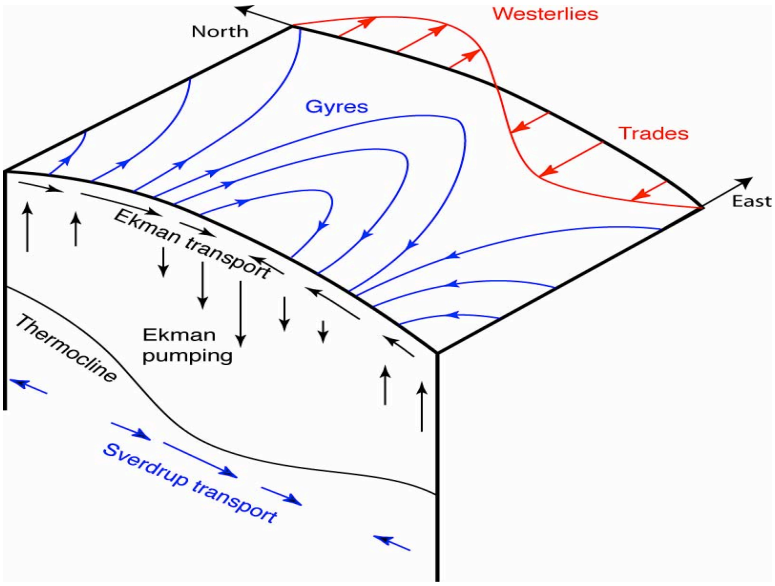


Figure 8.9 Sverdrup transport in the North Pacific Gyre.

The western intensification makes the currents on the western boundary current strong, such as the Kurishio Current in the Pacific and the Gulf Stream in the Atlantic. The eastern boundary current, such as the California Coastal Current on the eastern side of the Pacific, is relatively weaker.

Western intensification also occurs for all the oceanic gyres shown in Figures 8.10 and 8.12. The Western intensification phenomenon was first explained by the American oceanographer Henry Stommel.

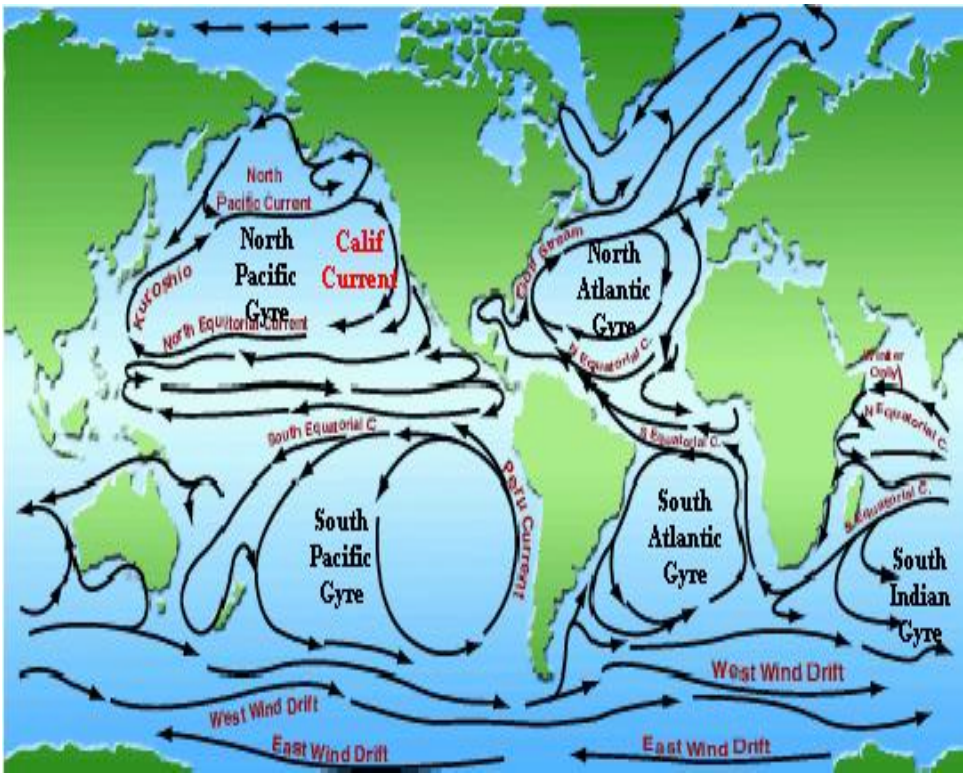


Figure 8.10 Coastal currents and and oceanic gyres.

Because of equatorial upwelling, the thermocline is shallower over the equatorial zone as shown in the above figure (Figure 8.14). Figure **??** shows the observed ocean temperature climatology at the 5-meter surface layer and the 200-meter layer below the equatorial thermocline.

In classroom show the movie of ocean temperature climatology of 33 layers down to 5,500 meters depth, produced in 2013 by SDSU, NASA JPL, and University of Hawaii.

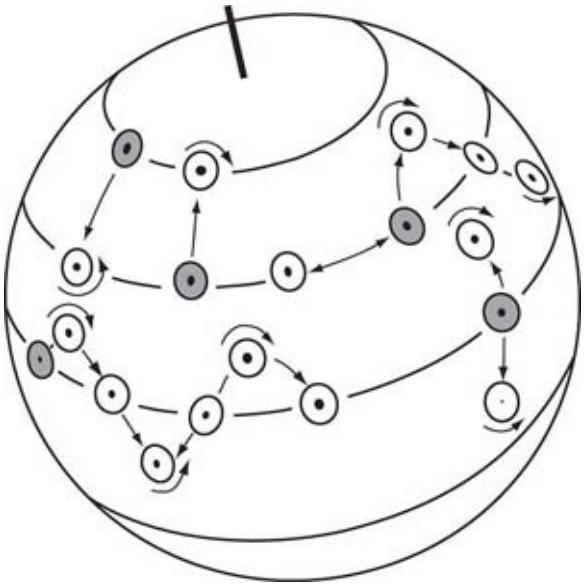


Figure 8.11 Vortex travels north and south according to the conservation of potential vorticity.

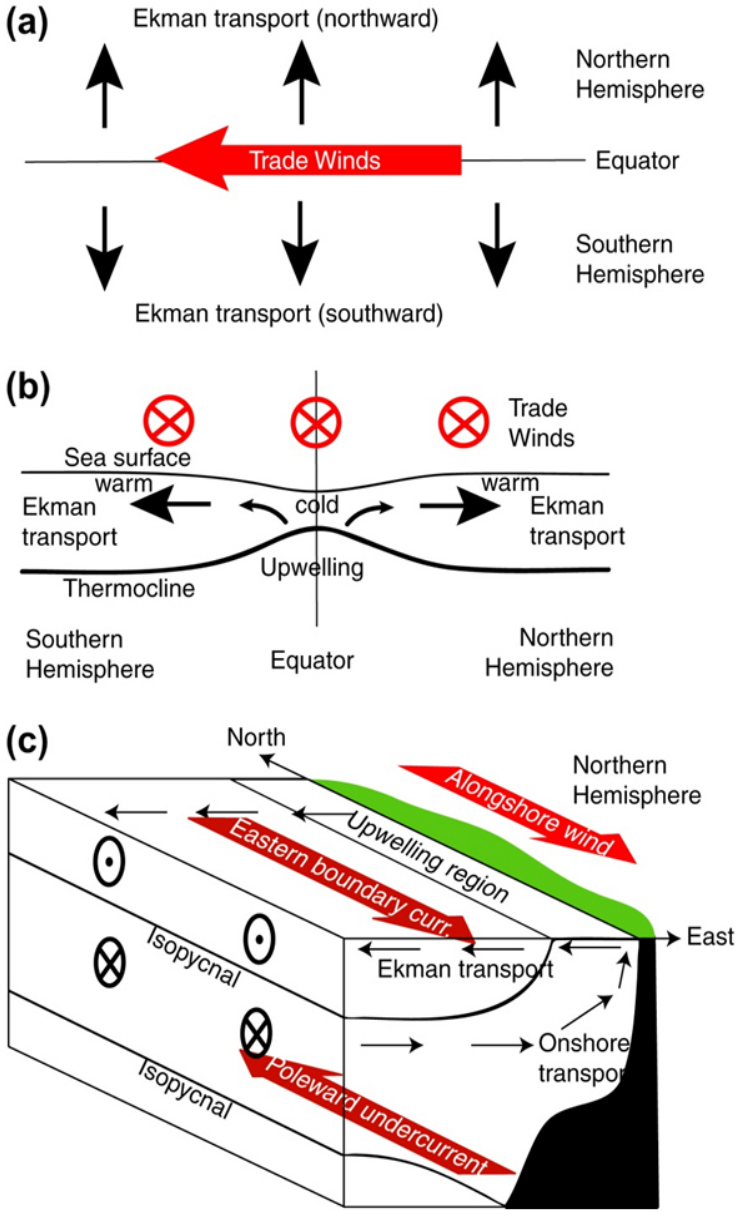


Figure 8.12 Ekman transport of water from equator to north and south.

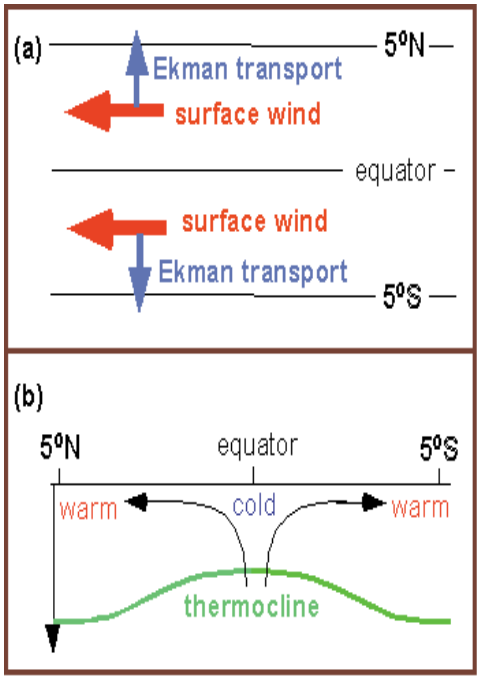


Figure 8.13 Upwelling of cold water in the tropical ocean. About 150 meters deep, the ocean water is not the hottest in the equator, but around zonal band near 20°N and 20°S.

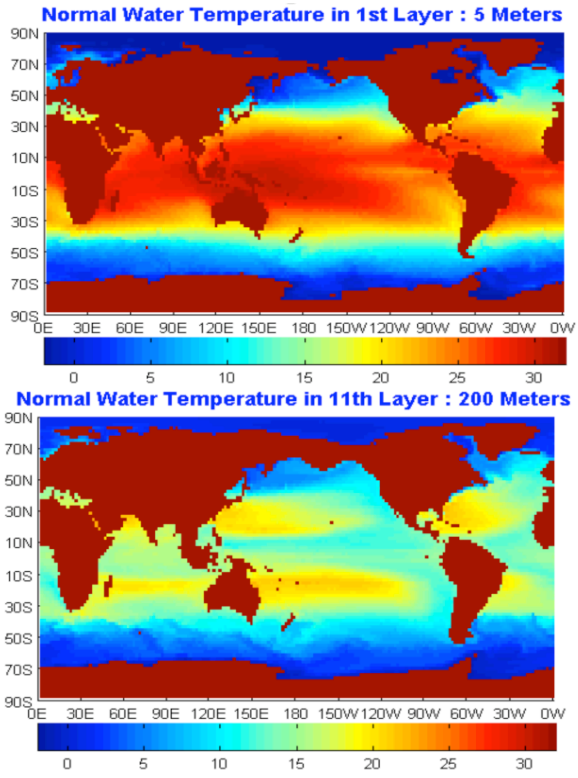


Figure 8.14 Ocean water climatology at the 5-meter surface layer and at the 200-meter layer.

APPENDIX A

DOT PRODUCT AND CROSS PRODUCT OF VECTORS

The multiplication of two scaler is still a scaler, such as $4.2 \times 2 = 8.4$. However, vector multiplication can have two kinds. One is the dot product which yields a scaler, and another is the cross product which yields a vector. The work done by a force $\mathbf{F}$ which makes $\mathbf{D}$ displacement is a dot product of the two vectors $W = \mathbf{F} \cdot \mathbf{D}$. The Coriolis force is a vector and is equal to a cross product of a velocity vector with the angular velocity vector of the rotation frame $\mathbf{f_c} = 2\mathbf{u} \times \mathbf{\Omega}$.

Without introducing the coordinate frames, the two products are computed by the following formulas:

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta, \tag{A.1}$$

$$\mathbf{a} \times \mathbf{b} = (|\mathbf{a}||\mathbf{b}| \sin\theta)\, \mathbf{c}, \tag{A.2}$$

where $|\mathbf{a}|$ is the length of vector $\mathbf{a}$, $|\mathbf{b}|$ is the length of vector $\mathbf{b}$, θ is the angle between the two vectors $\mathbf{a}$ and $\mathbf{b}$, and $\mathbf{c}$ is a unit vector whose length is one and whose direction is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$ and determined by the right hand principle shown in Figure A.1.

The above formulas and Figure A.1 show that the vector length of a cross product is equal to the area of the parallelogram formed by the two vectors, and the dot product is equal to the area of a rectangle whose length $|\mathbf{a}|$ and whose width is the projection of $\mathbf{b}$ on $\mathbf{a}$: $|\mathbf{b}| \cos\theta$.

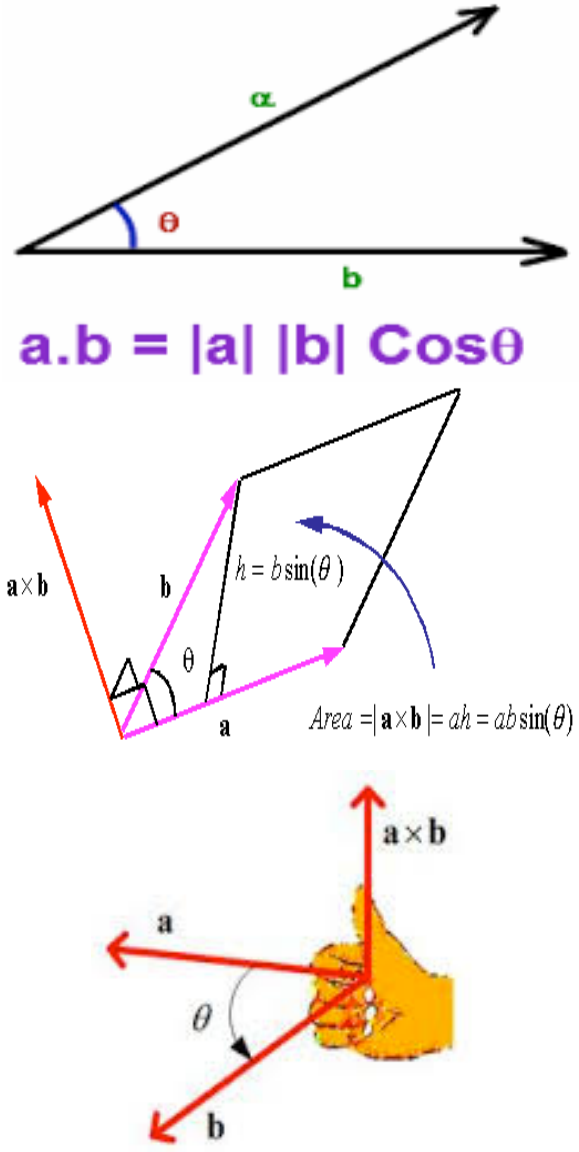


Figure A.1 a. Dot product, b. cross product, and c. the right hand principle.

The above gives a good and convenient conceptual definitions of the two products. In practice, vectors are represented by coordinate components. The formulas for the above two products using coordinate components are below.

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}, \tag{A.3}$$

$$\mathbf{b} = b_x\mathbf{i} + b_y\mathbf{j} + b_z\mathbf{k}, \tag{A.4}$$

$$\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y + a_z b_z, \tag{A.5}$$

$$\mathbf{a} \times \mathbf{b} = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{bmatrix}, \tag{A.6}$$

or

$$\mathbf{a} \times \mathbf{b} = (a_y b_z - b_y a_z)\mathbf{i} + (b_x a_z - a_x b_z)\mathbf{j} + (a_x b_y - b_x a_y)\mathbf{k}. \tag{A.7}$$

EXAMPLE A.1

$$\mathbf{a} = \mathbf{i} - \mathbf{j} + 2\mathbf{k}, \tag{A.8}$$

$$\mathbf{b} = -\mathbf{i} + \mathbf{j} + 1\mathbf{k}, \tag{A.9}$$

$$\mathbf{a} \cdot \mathbf{b} = 1 \times (-1) + (-1) \times 1 + 2 \times 1 = 2, \tag{A.10}$$

$$\begin{aligned} & \mathbf{a} \times \mathbf{b} \\ = \; & [(-1) \times 1 - 1 \times 2)]\mathbf{i} + [(-1) \times 2 - 1 \times 1]\mathbf{j} + [1 \times 1 - (-1) \times (-1)]\mathbf{k} \\ = \; & -3\mathbf{i} - 3\mathbf{j}. \end{aligned} \tag{A.11}$$

Because these two vectors $\mathbf{a}$ and $\mathbf{a}$ are on the same plane perpendicular to the xy-plane, whose normal vector must be perpendicular to the z-axis.

APPENDIX B

TRIGONOMETRIC FUNCTIONS

Some students may need to refresh their knowledge of trigonometric functions which are used often for wave modeling. This appendix chapter will provide a one-hour lecture materials to describe the trigonometry on a unit circle, trigonometric identities without memorization, trigonometric functions, and their graphs, derivatives and integrals.

The main concepts to be learned are frequency, amplitude, phase, and wave superposition.

Trigonometric functions can perhaps be best described on a unit circle, whose radius is one (see Figure B.1)

When a point is moving on the circle, the functions vary according to the angle θ. Because the point makes the circular motion and returns to the same point after a complete circle, the trigonometric functions all repeat their values after each cycle of 2π radian, or 360°. When functions repeat themselves after a certain interval, these functions are called periodic functions, and the minimum interval for the cycle to repeat is called the period of the function.

Given Figure B.1, the $x - y$ coordinates of the points A, B, C, D, E are functions of the angle θ between the x-axis and the unit radius OB. Using Pythagorean theorem, one can easily compute the trigonometric function values of the special angles at $0, 30, 45, 60, 90, 180, 270^\circ$, which are graphically shown in Figure B.2.

One can plot the graphs of the trigonometric function, such as the sine function shown in Figure B.4.

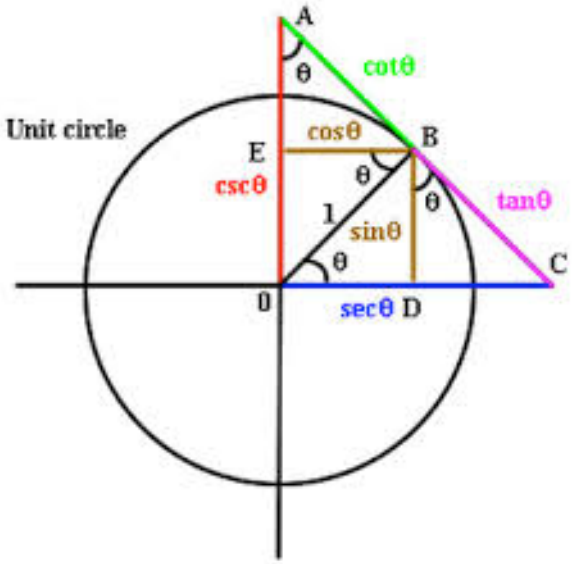


Figure B.1 Trigonometric functions defined by unit circle.

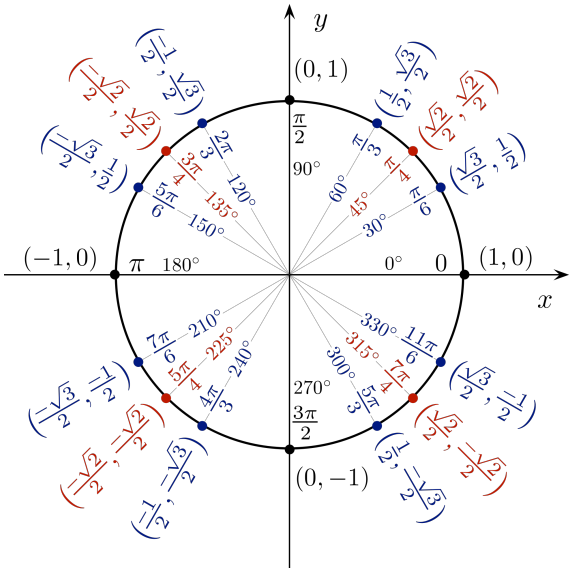


Figure B.2 Trigonometric function values for special angles.

degrees	0°	30°	45°	60°	90°
radians	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$
sin x	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1
cos x	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0
tan x	0	$\frac{1}{\sqrt{3}}$	1	$\sqrt{3}$	–

Figure B.3 Table B.1: Trigonometric function values for special angles.

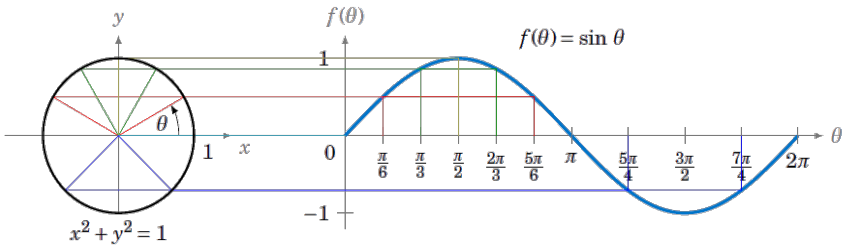


Figure B.4 Table B.1: Trigonometric function values for special angles.

APPENDIX C

POLAR COORDINATES AND CARTESIAN COORDINATES ON A UNIT CIRCLE

Climate Mathematics,
First Edition.

APPENDIX D

SPHERICAL COORDINATES AND CARTESIAN COORDINATES ON A UNIT SPHERE

Large scale atmospheric and oceanic circulations involve fluid dynamics on a rotating sphere. Coriolis force is very important. The spherical coordinates and Cartesian coordinates on a unit sphere often need to be converted from one to another.

Concepts: Spherical geometry, great circle, gradient, Coriolis force, wind direction, circulation patterns

Skills: Conversion formulas, climate model PDEs in Cartesian and spherical coordinates.

Climate Mathematics,
First Edition.

APPENDIX E

2015 MIDTERM EXAM OF UCSD SIOC290S: CLIMATE MATHEMATICS

Climate Mathematics,
First Edition.

SIOC290S Midterm Exam
September 4, 2015, Friday, 11:00am-12:15pm

Student Name:

Student ID Nnumber:

1. [20 $points$] Use integral to describe the accumulated degree days (ADD) [units: °C Day] of a growing season for a farm. You may use the integral of daily mean temperature [units: °C] with respect to time [units: day] to describe ADD. Discuss the ADD values in a given period of time during a growing season, relating the ADD values to the growth of a crop, say corn.

   Requirements: You must use at least one figure and at least one formula. The English text must be longer than 50 words.

2. [20 $points$] The spatial average annual mean surface air temperature over a country is 14°C. The temperature is assumed to be normally distributed. A group of 36 samples was taken in the same country. The sample data have a mean equal to 14°C and a standard deviation of 0.3°C. Find the confidence interval of this group of samples at 95% confidence level.

3. [20 $points$] A ball is tossed up straight at 20.0 $[m/s]$ at an initial height $1.0[m]$. How long will it take the ball to reach the maximum point?

   Requirement: Take derivative at least once.

   Hint: Choose the gravitational acceleration to be $9.8[m/s^2]$, and find an approximate answer. The general formula for the height position of the ball is $h = -gt^2/2 + v_0 t + h_0$, where v_0 is the initial velocity, and h_0 is the initial position.

   {2 bonus points: What is the speed when the ball returns to the ground, i.e., when $h = 0$? You have to use the h function, derivative, and detailed calculations to justify your answer. }

4. [20 $points$] Find the linear approximation of $f(x) = x^2 - 1$ around $x = 2$. If this $x_0 = 2$ is used as the first guess, find the next approximate root x_1 of $f(x) = 0$ by Newton's method.

5. [20 $points$] The following is the SVD result of a matrix called mat:

```
mat
      [,1]   [,2]
[1,]    1     1
[2,]    1    -1
```

```
svd(mat)
$d
[1] sqrt(2)   sqrt(2)

$u
           [,1]         [,2]
[1,] -sqrt(2)/2    -sqrt(2)/2
[2,] -sqrt(2)/2    sqrt(2)/2

$v
     [,1]    [,2]
[1,]   -1    0
[2,]    0   -1
```

Use A=UDV' to recover the second column of

```
mat
     [,1]   [,2]
[1,]    1    1
[2,]    1   -1
```

Requirements: Show detailed calculations of all the relevant matrices and vectors. Use your calculation results and at least 20 words to describe the space-time decomposition of a space-time data matrix.

{2 bonus points: Assume that the data matrix's spatial locations are identified by rows and the time is counted by columns. Use (a) figure(s), (b) mathematical formula(s) and calculation, and (c) English text to describe the two spatial modes determined by the U matrix. Comment on the variance, or energy, corresponding to each mode. }