# 1
# Introduction

**About Chapter 1**

In this introductory chapter, we briefly go over the definitions of terms and tools we need for data analysis. Among the tools, MATLAB is the software package to use. The other tool is mathematics. Although much of the mathematics are not absolutely required before using this book, a person with a background in the relevant mathematics will always be better positioned with insight to learn the data analysis skills for real applications.

## 1.1 Some Definitions and Concepts

### *1.1.1 About Data*

In most parts of this book, by *data* we specifically mean a *time series* or a set of time series – numbers or vectors in a sequence obtained through time from measurements, numerical model simulations, or predictions using certain methods or *algorithms* from computation by some theory or formula. In other words, data here are generally a series of numbers (usually real numbers) or a series of vectors that are lined up in time, i.e. each element in the series (either a number or a vector) is associated with an independent *time stamp*. Occasionally, we discuss data without an explicit time stamp (e.g. coastline data). Here, a vector is a group of independent numbers or variables such as the velocity components $u, v, w$ defined in a three-dimensional space and time. A vector can also be a more abstract *collection of variables with different units*, e.g. the so-called *phase space* in statistical physics in which the instantaneous state of a system with $n$ particles is described by the coordinates of all $n$ particles and the velocities of the $n$ particles $(x_1, y_1, z_1, \ldots, x_n, y_n, z_n, u_1, v_1, w_1, \ldots, u_n, v_n, w_n)$. This kind of vector is often seen in *empirical orthogonal function* (EOF) analysis. When a time series is discussed, we either imply that there is a series of time stamps associated with a

series of numbers or vectors, or the data include explicitly the time stamps with a series of numbers or vectors.

Occasionally, data might be complex number time series with real and imaginary parts (the imaginary part is the product of a real number and $i = \sqrt{-1}$). In this book, we mostly work with real number time series. However, we will also work with complex number series in two subjects, especially the first subject: (1) *Fourier analysis* (Davis, 1963): when we do Fourier analysis and *fast Fourier Transform* for a time series, complex numbers are often introduced for operational computation purposes, although in theory that is not absolutely needed; (2) *rotary spectrum analysis*: in oceanography, for two-dimensional flow velocity vector time series, we sometimes use a technique called rotary spectrum analysis, which is a special kind of Fourier analysis for vector time series to examine the spectra of velocity in terms of the cyclonic and anticyclonic rotations. In this kind of analysis, the horizontal velocity components are expressed in a complex form for mathematical convenience ($w = u + iv, i = \sqrt{-1}$), taking advantage of the Fourier series in exponential form using the Euler formula $e^{ix} = \cos x + i \sin x$.

### 1.1.2 Function vs. Discrete Series of Numbers

Data are collected at discrete time instances and/or locations. For that reason, we sometimes use the phrases *digital data* and *discrete data*. Digital data is perhaps used more often to mean that data are saved in a "digital" form in a computer. In the context of discussion here, there is no difference between a data file saved in a computer and a data file with numbers written on a piece of paper. The point is that time series data are not "continuous" in the mathematical sense. However, we use a lot of theories and results from mathematical work on *continuous functions*. Of course, a continuous function is simply an idealized simplification and cannot be realized in real life, particularly in scientific data. No matter how fast one samples, there is always a finite sampling interval, and the data is discrete rather than continuous.

This intrinsic discontinuous nature of data does not really affect our application of mathematical results derived from studies on continuous functions. However, we sometimes do say something like "this time series has a discontinuity." That can be concluded by a quick eye-ball observation of the data file or a more sophisticated and objective computer-based scan of the data with certain criteria. This is, unfortunately, subjective. There may be an unusual outlier in the data that appears to be unreal due to instrument failure or some unexpected influence of other environmental factors. For example, say a water-level sensor is deployed at the bottom of an estuary, measuring the water level at an hourly interval. A storm arrives, and some strong waves unexpectedly push the instrument into a 5 m

deeper water despite the fact that the instrument may be heavy and would not move under normal conditions. The data file would record at the next hourly data point an increase of 5 m in water level. The sudden jump is obviously "discontinuous." Of course, if the instrument is securely installed, this jump may not happen. But the reality is not always ideal. An instrument normally considered to be securely mounted may not be secure anymore if there is a record-breaking storm. The judgment of whether there is a discontinuity in a data file depends on the nature of the data under study and the experience of the person who does the analysis. Usually, a quick QA/QC with defined rules would be able to find "discontinuity" points in the given time series data.

### *1.1.3 Time Series vs. Spatial Series*

Can numbers in space rather than in time be included in the analysis for either theoretical or application purposes? In general, yes. For instance, using the Fourier Transform as an example, instead of having variations expressed in time that can be converted into the *frequency domain*, a series of real numbers for different positions in space (i.e. *spatial series*) can be converted into the *wavenumber domain*. A spatial series can be in a one-dimensional, two-dimensional, or three-dimensional space, and thus the wavenumber can be either a scalar or a vector.

Our focus here in this book is time series, although at times we discuss the analysis of spatial data or space–time mixed data. In a sense, when one understands clearly time series analysis, it should become easier to understand spatial series analysis. Of course, these two types of data can be very different. One cannot simply use the time series analysis methods on spatial series without considering their differences. For one difference, a time series is ordered naturally from a start time (earlier) to an end time (later). Data in space, however, are not necessarily ordered "naturally." For example, for data defined on a rectangular grid, we could order the data according to the rows, or columns, or any other peculiar way. After all, spatial data may not be always given on a nice, rectangular grid. The data points could be totally random in distribution.

Full coverage of the analysis of spatial data is not within the scope of this book. In this book, we only include some aspects of the techniques dealing with data obtained in space, such as those obtained from a moving platform. Strictly speaking, they are not necessarily a pure spatial series – they are data mixed in time and space because a moving vessel's speed is limited.

It should be noted that we will discuss a method of analysis, the *empirical orthogonal function* or *EOF analysis*, which involves data mixed in time and space, or a collection of time series defined at different locations. As an example, a series of satellite images for the same region falls into this category – for each

image, it presents a spatial series, but for each position in the region, there is a time series provided by the sequence of the images.

### *1.1.4 The Time Stamps and Time Intervals*

When we are working with time series, the order of the sequence of numbers is important because that provides information about variations in time (e.g. an event or a trend) that determine the rate of change in time. We are interested in the variations, though the average values can be important as one parameter. How a measured or simulated variable would change with time (or space, for spatial series) is what we are often interested in.

To calculate the rate of change in time, we must have information about the unit of time and *time intervals* of the data. Therefore, a time series of temperature, for instance, is not just a series of temperature values but also a series of time values with a proper unit that goes side-by-side with the temperature values that are measured or defined at those times.

For time series data, the time intervals between data points can vary. It is more convenient in many applications that the time series data are obtained at constant time intervals. This is often the case with modern instruments that collect data automatically with a built-in *microprocessor* or mini-computer and *data logger*. In this case, and also in the case of model output of time series data, some data files might ignore explicit time stamps; however, information about time intervals and start time with proper units must be provided either inside the data file or separately, depending on the designer of the data collection device or the programmer of a computer model and nature of data (e.g. whether a uniformly spaced time series data or not).

Many of the data analysis techniques are directly applicable only to uniformly spaced data with constant time intervals. Otherwise, some treatment such as interpolations or resampling needs to be done first. There are some exceptions to this. For one example, the general *least squares method* does not require that data points are equally spaced in time. A specific example in oceanography is the *tidal harmonic analysis*, which does not require that the data have constant time intervals. Harmonic analysis is based on the least squares method. The following is an example of time series of velocity vector at 2-minute intervals.

Each line corresponds to one *record of data* from observations at a given time instance. Here the time stamps (the first six columns) are provided. The last two columns give the east and north components of velocity in cm/s. Alternatively, the data could have been provided without the first six columns, but the starting time

| Year | Month | Day | Hour | Minute | Second | VE (cm/s) | VN (cm/s) |
|---|---|---|---|---|---|---|---|
| 2016 | 04 | 05 | 21 | 32 | 56 | −1.1 | 1.1 |
| 2016 | 04 | 05 | 21 | 34 | 56 | −2.6 | 2.8 |
| 2016 | 04 | 05 | 21 | 36 | 56 | −5.1 | 4.0 |
| 2016 | 04 | 05 | 21 | 38 | 56 | −10.4 | 8.5 |
| 2016 | 04 | 05 | 21 | 40 | 56 | −8.3 | 7.9 |
| 2016 | 04 | 05 | 21 | 42 | 56 | −5.5 | 4.8 |
| 2016 | 04 | 05 | 21 | 44 | 56 | −5.0 | 3.7 |
| 2016 | 04 | 05 | 21 | 46 | 56 | −4.8 | 3.7 |
| 2016 | 04 | 05 | 21 | 48 | 56 | −13.2 | 9.8 |
| 2016 | 04 | 05 | 21 | 50 | 56 | −10.9 | 7.4 |
| 2016 | 04 | 05 | 21 | 52 | 56 | −6.7 | 3.9 |
| 2016 | 04 | 05 | 21 | 54 | 56 | −4.6 | 2.7 |
| 2016 | 04 | 05 | 21 | 56 | 56 | −3.0 | 1.1 |
| 2016 | 04 | 05 | 21 | 58 | 56 | −2.7 | 0.4 |
| ... | | | | | | | |

and time intervals have to be provided for the time series data to be meaningful. For observational data, it is probably better to have the time stamps included explicitly for each record to avoid mistakes because real observations, especially the raw data files, often have gaps for various reasons. If the time stamp is not explicitly included in a data file for each record (or each sample), a single gap can mess up the entire dataset by introducing misalignment in time. With proper time stamps, after reliable quality control of the data, the gaps could be filled. For constant interval data, inclusion of the time stamps will make the data file larger. If data file size is to be minimized, the time stamps can be omitted, given that the start time and time interval are provided in the data file or with the data file. This is often practiced for saving numerical model output files, which can be extremely large.

### 1.1.5 Oceanographic and Other Data

By *oceanographic data*, we intend to limit our discussion to data of oceanic origin, i.e. those obtained from the ocean or coastal and estuarine waters. This, however, should not be taken too narrowly as far as the data analysis techniques are concerned. Generally speaking, the methods discussed in this book can be applied to data from other disciplines. For instance, the fundamental methods of analysis for atmospheric data should be very similar, if not the same. There would be no

essential differences. The atmospheric and oceanic time series data sometimes need to be analyzed together, e.g. for storm surge problems, in which the atmosphere provides the forcing (air pressure, wind stress, precipitation), while ocean and coastal waters respond to the forcing. Oceanographic data do indeed have some unique aspects that are pertinent mainly, if not only, to the ocean dynamics. For example, tides occur in the ocean, and we will discuss the tidal analysis (i.e. the abovementioned harmonic analysis). Although there is a tidal signal in the atmosphere known as the *atmospheric tide*, harmonic analysis is usually meant for the ocean only. The technique, however, is applicable to the atmospheric tide as well.

### *1.1.6 The Tool We Use: MATLAB*

MATLAB is a commercial software package that is suitable for calculations and related visualizations of *vectors* (one-dimensional), *matrices* (two-dimensional), and *arrays* (one-, two-, or multidimensional). Although the methods of analysis are independent of the computer programs that implement them, we use MATLAB exclusively in this book. It has many choices of mathematical tools. For example, it has specialized tools for Signal Processing and Wavelet Analysis Toolboxes, which we may need to use at some point in this book. MATLAB is easy to learn and use, and yet has lots of powerful capabilities. Alternative computer languages and or software packages also exist, e.g. IDL, R, C, C++, FORTRAN, Python, etc. Many resources on these computer languages are available online or in publications, but we only use MATLAB throughout this book for consistency and simplicity.

## 1.2 Background Knowledge

To learn the materials in this book well, some background knowledge in mathematics and physics and related basic skills are preferred, though a high-level skill of any of these subjects is not necessarily required. The subjects of study, particularly in terms of the techniques in this book, are quite broad. Background knowledge includes *linear algebra*, *calculus*, *Fourier theorem*, *numerical analysis*, and some basics of statistics. In addition, oceanographic data analysis is often aimed at the resolution of dynamical processes in the ocean. Therefore, some background knowledge in e.g. physics, fluid dynamics, and tidal theory would also help to understand some of the techniques (e.g. rotary spectrum analysis and tidal harmonic analysis). Assumptions are made here that the readers have some basic background knowledge of these subjects, but efforts have been made to provide enough information as standalone materials. Some selected basic

information and review of background theories are provided in the book for convenience. Intuitive interpretations may be provided for a better understanding when some background information is discussed.

### 1.2.1 Linear Algebra

Linear algebra is a branch of mathematics dealing with arrays of numbers. A time series of a quantity is itself a special case of an array. Linear algebra includes theories and methods of solving linear sets of equations. This provides the basis for many data analysis techniques, such as linear regression, Fourier Transform, and harmonic analysis. Knowledge in linear algebra makes it much easier to work with matrix operations, which are the major backbone of MATLAB. MATLAB is best suited for working with vector, matrices, and arrays. Linear algebra with matrix operations greatly simplifies the mathematical expression of concepts. These simplifications are implemented by the design of MATLAB language, while inside MATLAB the computer is directed to do the heavy lifting for detailed calculations. With a combination of the two (simplified concepts and MATLAB), our brain power can be reserved for interpretation of the results instead of the details of the actual calculations.

### 1.2.2 Calculus

Calculus is an important branch of mathematics about the rate of change of functions (differentiation) and the inverse calculations of differentiation (integration). It is the mathematical backbone of almost all disciplines in modern physics. Calculus is fundamental to many theories and techniques of analysis and number crunching, in many ways. This can be seen mainly in two examples in this book: one is Fourier Transform, in which the basic knowledge of integration and *linear independence* of *base function*s will help greatly. Similarly, the least squares method is also based on the theory of calculus. The optimal solution of the *Fourier coefficients* or *best fit* in the least squares method is a great product of calculus. Another example of the application of calculus is *Taylor series expansion*, which is very useful in laboratory experiments for *error estimations*. The beauty of Taylor series expansion is that it can approximate an arbitrary differentiable function to any accuracy (at least in theory) by a polynomial, which is one of the simplest general functions one can find. In addition, the theories for the empirical orthogonal function and wavelet analysis are also based on calculus. Without calculus, none of these techniques would have been invented.

### 1.2.3 Fourier Theorem

*Fourier theorem* is the basis for the technique we discuss extensively in this book: the Fourier analysis. Although we will provide relatively complete, albeit brief, coverage of the theory, it will help if the readers have learned the subject before. The importance of Fourier theorem to time series analysis can be compared to that of Newton's Second Law to mechanics. This may be an overstatement, but the point is, Fourier theorem provides a great leap in understanding the characteristics of a function – in our case, a function of time. The theorem basically guarantees that, except for a peculiar function with infinite discontinuity points, a general (arbitrary) piece-wise continuous function that only has a finite number of finite-range jumps can be expressed (decomposed) in terms of a series of sine and cosine functions of different scales. Here, scales are either in terms of frequency for time series or wavenumber for spatial series.

In the case of a time function, if the length of time of the function is finite (which is usually the case, because no real observations can be infinitely long, except in idealized thought experiments), the series of sine and cosine functions generally are discrete in terms of the scales (discrete frequencies or wavenumbers), although there are infinite numbers of them. These (infinite number of ) discrete sine and cosine functions are called the base functions. The series itself is called the *Fourier series*. The Fourier theorem tells us that such an infinite Fourier series converges to the original function in an averaged sense. Here, "average" means the average of the right and left limits of the function at one point, obtained by approaching from the right of the point and that the left of the point. In other words, at any point (or time, for time function) where the function has a finite discontinuity, the Fourier series converges to the average of the two points on both sides of the jump – i.e. the middle of the jump. If the function at the point is continuous, these two limits are the same. That is, if the function is continuous, the Fourier series will converge to the function at that point.

The Fourier theorem also helps us to understand the effect of finite sampling, such as the highest frequency that can be resolved (i.e. the *Nyquist frequency*), the frequency resolution, and the artificial oscillations or *Gibbs Effect*. The properties of the Fourier series allow a wide applicability and an extension into the *Fourier Transform*.

### 1.2.4 Numerical Analysis

Numerical analysis deals with some practical methods and techniques of calculations using computers, such as root finding from a nonlinear equation, *interpolation* and techniques using polynomials, numerical differentiation,

numerical integration, calculations in linear algebra, and numerical solutions for ordinary differential equations. An example of relevance to time series analysis in this book is an understanding of interpolations, although we will not discuss the theories of interpolations; we will, rather, rely on just MATLAB's built-in functions for interpolations. Interpolations are oftentimes among the first steps of data processing, before conducting a Fourier Transform or spectrum analysis for a time series obtained from actual measurements from the ocean (or anywhere else). As mentioned earlier, observations usually contain gaps. Interpolations are useful for filling "small" *data gaps*, in which the gaps between adjacent data points are much smaller than the minimum useful *time scale*s. For large data gaps, interpolation may introduce significant errors. How do we determine if a gap is really "small"? This depends on the problem under discussion. It depends on the significant time scales contained in the signal, time intervals, number of gaps, and length of the gaps. Although actual situations can have an infinite number of possibilities, the decision is usually based on common sense. For instance, for a problem with tidal variations of water level, we have a time series with 10-minute time intervals. For this time series, if there are six consecutive *missing data* points in a row within a tidal cycle (12-hour period for a semi-diurnal, tide-dominated case), the gap would not pose a major problem; a simple interpolation would fill the data gap without altering the information in any significant way, meaning no major influence on the spectrum or temporal properties. However, if the time interval is 1 hour and there are six consecutive data points missing (e.g. a 6-hour data gap), the time series property and spectrum after an interpolation for that 6-hour gap would be highly questionable. That, of course, also depends on how we do the interpolation: Are we using a linear interpolation, or are we using some functions for the interpolation? The problem with a function is that for real world observations, if data are missing, there is no way we know exactly what happened during the time when valid data were not recorded. Any function used for interpolation is subjective and can never be verified, because what we do not know is exactly whether there was any "event" during the time with missing data. If we have additional information, e.g. observations from an adjacent position, then the story will be completely different, as then we might be able to assume a certain relationship for the interpolation based on the additional information.

### *1.2.5 Statistics*

Statistics is needed here and there in the book. Statistics, however, can arguably be a double sword in this context. On the one hand, statistics is often needed for analysis on real data because of the intrinsic random nature of observations: real data are functions affected by *random processes*, at least to some extent. The random

processes include intrinsic characteristics of the physical processes, such as chaos in the dynamics or random errors in measurements and systematic errors in instruments. On the other hand, many of the mathematical tools are developed first, without considering the random nature of numbers at all, i.e. only deterministic functions are considered. As we all know, calculus is presented originally and mostly with the differentiations and integrations of *deterministic functions* rather than random functions. Fourier Transform was also originally developed without any consideration of the randomness of functions. When these mathematical operations are applied to real data, however, the effects of randomness may have to be considered. This will make the problem a little more complicated, but we have to admit that there are some uncertainties, and any estimate will need to take that into account. Having said these, this book is not a statistics book. Some of the analysis will be easier to understand without considering randomness in the beginning. The statistical aspects can be better understood after one has grasped the basics of the techniques developed for deterministic functions.

## 1.3 About Wind, Current, and Wave Directions in 2-D

In oceanographic data analysis, we often work with two-dimensional vector time series data such as wind, current, and wave, when the vertical components of wind, current, and wave are not considered. Here we discuss and contrast the definitions of directions of two-dimensional wind, current, and ocean waves. It can be confusing as to what the definitions of directions of the two-dimensional wind, current, and wave are. Although two-dimensional wind and currents are vectors that can be expressed in two components (for the horizontal two-dimensional wind and current velocities), they are routinely reported with a magnitude and direction. The direction of wind and that of currents are, however, opposite in definition. It can be confusing and therefore is important that the definitions of these directions are well understood by the data users.

*Wind velocity* is a vector which tells us where an air parcel at a given location and time is moving to. In theory, wind velocity should have three components in a Cartesian coordinate: the *vertical wind velocity component* and two horizontal components. The vertical component, however, is not routinely reported unless for special studies of cloud physics, precipitation mechanism, and small-scale atmospheric processes in the troposphere. For this reason, when we talk about the *wind direction*, we are referring only to that of the *horizontal wind velocity*. This is the same for current velocity and waves. When the ocean current direction is mentioned, we usually mean that of the horizontal velocity. The direction of propagation of internal waves can be at an angle to the surface. Here we are only referring to the direction of surface waves.