CHAPTER SUMMARIES

Chapter 1, *Data Analysis with MatLab*, is a brief introduction to *MatLab* as a data analysis environment and scripting language. It is meant to teach *barely enough* to enable the reader to understand the *MatLab* scripts in the book and to begin to start using and modifying them. While *MatLab* is a fully-featured programming language, *Environmental Data Analysis with MatLab* is not a book on computer programming. It teaches scripting mainly by example and avoids long discussions programming theory.

Chapter 2, *A First Look at Data*, leads students through the steps that, in our view, should be taken when first confronted with a new dataset. Time plots, scatter plots and histograms, as well as simple calculations, are used to examine the data with the aim both of understanding its general character and spotting problems. We take the position that *all data*sets have problems – errors, data gaps, inconvenient units of measurement, and so forth. Such problems should not scare a person away from data analysis! The chapter champions the use of the *reality check* - checking that observations of a particular parameter really have the properties that we know it must possess. Several case study datasets are introduced, including a hydrograph from the Neuse River (North Carolina, USA), air temperature from Black Rock Forest (New York) and chemical composition from the floor of the Atlantic Ocean.

Chapter 3, *Probability and what it has to do with Data Analysis*, is a review of probability theory. It develops the techniques that are needed to understand, quantify and propagate measurement error. Two key themes introduced in this chapter and further developed throughout the book are that error is an unavoidable part of the measurement process and that error in measurement propagates through the analysis to affect the conclusions. Bayesian inference is introduced in this chapter as a way of assessing how new measurements improve our state of knowledge about the world.

Chapter 4, *The Power of Linear Models*, develops the theme that making inferences from data occurs when the data are distilled down to a few parameters in a quantitative model of a physical process. An integral part of the process of analyzing data is developing an appropriate quantitative model. Such a model links to the questions that one aspires to answer to the parameters upon which the model depends, and, ultimately, to the data. We show that many quantitative models are *linear* in form and, thus, are very easy to formulate and manipulate using the techniques of linear algebra. The method of least squares, which provides a means of estimating model parameters from data, and a rule for propagating error are introduced in this chapter.

Chapter 5, *Quantifying Preconceptions*, argues that we usually know things about the systems that we are studying that can be used to supplement actual observations. Temperatures often lie in specific ranges governed by freezing and boiling points. Chemical gradients often vary smoothly in space, owing to the process of diffusion. Energy and momentum obey conservation laws. The methodology through which this prior information can be incorporated into the models is developed in this chapter. Called generalized least squares, it is applied to several substantial examples in which prior information is used to fill in data gaps in datasets.

Chapter 6, *Detecting Periodicities*, is about spectral analysis, the procedures used to represent data as a superposition of harmonically-varying components and to detect periodicities. The key concept is the Fourier series, a type of linear model in which the data are represented by a mixture of sinusoidally-varying components. The chapter works to make the student completely comfortable with the Discrete Fourier Transform (DTF), the key algorithm used in studying periodicities. Theoretical analysis and a practical discussion of *MatLab's* DFT function are closely interwoven.

Chapter 7, *The Past Influences the Present*, focuses on using past behavior to predict the future. The key concept is the *filter*, a type of linear model that connects the past and present states of a system. Filters can be used both to quantify the physical processes that connect two related sets of measurements and to predict their future behavior. We develop the *prediction error filter* and apply it to hydrographic data, in order explore the degree to which stream flow can be predicted. We show that the filter has many uses in addition to prediction; for instance, it can be used to explore the underlying processes that connect two related types of data.

Chapter 8, *Patterns Suggested by Data*, explores linear models that characterize data as a mixture of a few significant patterns, whose properties are determined by the data, themselves (as contrasted to being imposed by the analyst). The advantage to this approach is that the patterns are a distillation of the data that bring out features that reflect the physical processes of the system. The methodology, which goes by the names, *factor analysis* and *empirical orthogonal function (EOF)* analysis, is applied to a wide range of data types, including chemical analyses and images of sea surface temperature (SST). In the SST case, the strongest pattern is the El Niño climate oscillation, which brings immediate attention to an important instability in the ocean-atmosphere system.

Chapter 9, *Detecting Correlations among Data*, develops techniques for quantifying correlations within data sets, and especially within and among time series. Several different manifestations of correlation are explored and linked together: from probability theory, covariance; from time series analysis, cross-correlation and from spectral analysis, coherence. The effect of smoothing and band-pass filtering on the statistical properties of the data and its spectra is also discussed.

Chapter 10, *Filling in Missing Data*, discusses the interpolation of one and two dimensional data. Interpolation is shown to be yet another special case of the linear model. The relationship between interpolation and the gap-filling techniques developed in Chapter 5 are shown to be related to different approaches for implement prior information about the properties of the data. Linear and spline interpolation, as well as kriging, are developed. Two-dimensional interpolation and Delaunay triangulation, a critical technique for organizing two-dimensional data, are explained. Two dimensional Fourier transforms, which are also important in many two-dimensional data analysis scenarios, are also discussed.

***Chapter 11**, ‘Approximate’ is not a pejorative word,* explores the use of approximations to make data analysis simpler, faster and more adaptable. Taylor’s theorem is derived and used to linearize nonlinear functions. The method is used to create small-number approximations, to estimate the variance of nonlinear functions and to solve nonlinear least squares problems. The gradient of the error is identified as the key quantity controlling the solutions and its further application to the gradient method is explored. The lookup table is introduced as an approximate method of evaluating functions of one or two variables and applied to speeding up iterative calculations. Finally, the artificial neural network is developed as an approximation technique that shares the adaptability of a lookup table while providing smoothness and enhanced flexibility. The properties of several simple network designs are illustrated and the back-propagation algorithm for training them is derived. The power of a neural network is demonstrated by predicting the non-linear response of a river network to precipitation.

Chapter 12, *Are My Results Significant?,* returns to the issue of measurement error, now in terms of *hypothesis testing*. It concentrates on four important and very widely applicable statistical tests – those associated with the statistics, Z, χ^2, t and F. Familiarity with them provides a very broad base for developing the practice of *always* assessing the significance of *any* inference made during a data analysis project. We also show how empirical distributions created by *bootstrapping* can be used to test the significance of results in more complicated cases.

Chapter 12, *Notes* is a collection of technical notes that supplement the discussion in the main text.