WHY STATISTICS?

This course concerns a framework for thinking about oceanographic observations and, as they become more complex, ocean models. Much of the material involves "statistical" procedures and this material may appear foreign to students raised on a diet of "dynamics as boundary value problems". It is appropriate, therefore, to discuss why a "statistical" perspective is necessary in oceanography.

Without getting deeply into philosophy, let us accept the proposition that for most purposes the ocean can be regarded as exactly described by completely deterministic equations. There are three basic reasons why a statistical perspective is needed to describe this deterministic system:

a. the ocean is complex in the sense that it involves more dynamical degrees of freedom than can be observed or specified;

b. ocean dynamics are nonlinear so that groups of variables can not be studied in isolation;

c. ocean dynamics are often effectively unpredictable in the sense that the smallest changes in specifications lead quickly to O(1) changes.

The dynamical equations describing the ocean are a system of nonlinear partial differential equations which evolve in time and are formally specified only when complete initial and boundary data are prescribed. For the sake of argument, suppose that these systems are unique in the sense that only one solution exists for a given set of prescription data. Even with a perfect model which exactly reproduces the system's dynamics, predicting its evolution would require more prescription (initial and boundary) data than could possibly be observed, even in principle. The system then has more dynamical degrees of freedom than can be specified; it is complex. Because only part of the data can be prescribed, the dynamical system is underdetermined and, to us who know only part of the data, the evolution is not deterministic. Even if at two times all the observables are identical, the evolutions will differ because the unobservable degrees of freedom will be different. Although formally deterministic, the system outcome will appear to us as "random".

If the system dynamics were linear it would be possible to find observable subsystems, with finite degrees of freedom, whose evolution is isolated (determined only by itself) and therefore can be fully specified and made deterministic. The extreme of this is how surface waves can be divided individual components with distinct frequencies and wavenumbers. In the real world it is rarely possible to find subsystems that are strictly isolated and consequently fully observable (that is can be observed well enough that their dynamics are strictly deterministic). Oceanic sound propagation approaches this ideal more than any other ocean processes; although sound propagation depends weakly on most other ocean phenomena (currents, internal waves, surface waves etc.) it has little influence back on them. Most of what oceanographers and meteorologists call theories are actually attempts to isolate a particular range of time and space scales and determine the dynamics of the isolated system. While this is how we learn how things work, it is an approximation and the evolutions described can not be expected to be accurate for long nor can we expect to see these isolated dynamical systems in the ocean.

A system may be called unpredictable if infinitesimal changes in its prescription data lead in finite time to $O(1)$ changes in the outcome. Unstable linear systems are unpredictable since perturbations grow exponentially. Many of the nonlinear model systems used to describe parts of ocean dynamics are also unpredictable. If a complex system predictable, it may be possible to specify the important degrees of freedom with sufficient accuracy to make the system appear deterministic with a small added random component. But with an unpredictable system, even

if all prescription data is known, but with imperfect accuracy, the evolution is not deterministic, even with a perfect model. Not all ocean phenomena are unpredictable; sound waves, and to a lesser extent other wavelike motions, are reasonably predictable. But most phenomena are, as described in the classic paper by Lorenz (1969), "The predictability of a flow which possesses many scales of motion", Tellus XXI, 289-307, unpredictable and hence deterministic only for short times. The characteristics of complexity and unpredictability are not necessarily linked. Systems with few degrees of freedom can be highly unpredictable and infinite dimensional systems could be predictable so long as small changes always produce small effects.

In the case of underspecified or unpredictable systems the outcome can not be specified exactly, even with a perfect dynamical model. The unobserved degrees of freedom introduce uncertainty which we call random. It is still possible to learn things and this is where the statistical perspective comes in. The approach is to minimize randomness and explain as much as possible about the range of behavior under prescribed circumstances. Statistics are used to describe the typical characteristics of the system evolution when constrained by what is known.

As an example, consider the familiar procedure of interpolating between data points or fitting functions to data points. Despite its conceptual simplicity, this is actually an example of a complex system in the sense that we are interested in the properties of a system with more degrees of freedom than have been specified. Unless enough is known about the process that its behavior between observed points can be determined by calculation, the actual curve between points is, to us, random. It is then appropriate to approach interpolation as a statistical problem in which we try to interpolate with a curve that is typical of, or perhaps the most likely realization from, the process when it produces the observed points. Interpolation then depends on knowing something about the typical behavior, the statistics, of the particular process. When we fit a few functions to many data points we are admitting there are (random) things about the data we do not understand in detail and use the functions to describe what we hope are the features of the data that are not much influenced by these "noise" processes.

In a practical sense, the complexity or dynamical size of a system depends on how precisely we need to know it and that often depends on the size of the physical space in which it is to be examined. To demonstrate this, consider an ocean covered with a train of ideal linear surface waves. Viewed over a moderate range of time and space it might appear as a single sinusoid

$$\eta(x,t) = a\sin\left[k\cdot(x-x_0)+\omega t\right] \tag{1}$$

and the system's evolution could be predicted given the wave's amplitude a, frequency ω, wavenumber k, and the phase x_0. Given this information the parameters in η could be determined and the result advanced forward in time. Formally, this system is completely predictable in the sense that a small change in any parameter makes a small change in η so long as $x-x_0$ and t are $O(1)$. $O(\epsilon)$ errors in determining a and x_0 make $O(\epsilon)$ errors in η for all $x-x_0$ and t but an $O(\epsilon)\ll 1$ error in k or ω eventually leads to an $O(a)$ change in η because the error in the predicted phase grows with time until when $t>1/\epsilon$ it is effectively unpredictable. The converse of this is that given $O(\epsilon)$ observational uncertainty over a time T we cannot determine the frequency more accurately than $O(\epsilon/T)$. Thus over an $O(1)$ time range we could not tell the difference between a single sinusoid and a group of waves all having frequencies within a narrow band of width $\Delta\omega = O(\epsilon/T)$. Thus to describe behavior over the order one range we need only a simple 4-parameter system while a description over a longer range potentially involves many more parameters.

Systems with few degrees of freedom can sometimes be so unpredictable that they are better described with statistics of "typical" behavior than with analysis of detailed behavior. As an

example, Figures 1 and 2 show the behavior of the system

$$\begin{array}{rcccccc} dX/dt & = & -aX & +aY & & & \\ dY/dt & = & rX & -Y & & -XY & \\ dZ/dt & = & & & -bZ & +XY & \end{array} \tag{2}$$

with the parameters $a = 10$, $b = 8/3$ and $r = 28$. This system has three degrees of freedom because it is fully specified by three initial conditions. Lorenz pioneered the study of simple dynamical systems like this that have remarkably complicated chaotic behavior and developed this particular model as a distant dynamical relative of large-scale atmospheric waves.

Figure 1 shows time series for X and Z resulting from the initial conditions $X = Z = 0$ and $Y = 10$ (on the left) and $Y = 10.01$ on the right. The time series of Z looks a bit like a sinusoid with varying amplitude and frequency. X resembles a sinusoid that periodically jumps from one regime to another. There are no simple patterns (it has been proven) or periodicities. As comparison of the two realizations with initial conditions that differ by 0.1% shows, the system is unpredictable. While the time series is unpredictable and complex, it also has characteristics which distinguish it from others. The challenge to an observer would be to describe these characteristics more succinctly than to exhaustively plot the series.

Figures 2 shows the same time period as Figure 1 but portrayed as the system trajectory through X-Z space. Because we are looking at a three-degrees-of-freedom system in two dimensions the evolution of each cycle is along a different path. These trajectories are distinct but have a similar character. The figure shows why Z has pseudo sinusoidal behavior while X seems to exhibit two regimes, one associated with nearly circular orbits in X-Z space and the other associated with figure-eight trajectories. This figure also clarifies the discussion of how **randomness** is introduced by **missing or unobserved degrees of freedom.** Since the Lorenz model is a system of three first-order equations, each point in X-Y-Z space is on only one trajectory. But when projected onto the X-Z plane multiple trajectories pass through one point indicating variability and randomness. There is also a pattern to the random trajectories and a reasonable estimate of short time behavior could be made because certain patterns are more common, or more likely, than others. But exact predictions are impossible, even for short times. But if only the value of X were known it would be even more difficult to predict evolution. The analyst's challenge is to characterize the system with whatever fraction of its defining variables that can be observed.

In systems with many degrees of freedom, phase plane descriptions are usually not so useful and time series are a good deal more complex. One approach is then to separate the system into "signal plus noise". The separation is made according to what processes are to be studied. In one setting the focus may be on the general circulation with small scale processes (internal waves, turbulence, and perhaps mesoscale variability) considered "high frequency noise". In another setting the focus may be on surface waves, in which case the general circulation (sea-level variations, large-scale currents, etc.) might be regarded as noise which contaminates the signal with low frequency trends. When specific processes are examined, we are implicitly saying that certain degrees of freedom will not be treated deterministically but rather as random noise which hopefully only weakly affects the process being studied.

In large measure, the statistical view is forced by complexity. The aim is to deal with the essence of a process without dealing with it in detail. For example, the statement that penny flips produce 50% heads is descriptive and a lot more compact than describing the mechanics of coin ballistics and fingers. Economy is accomplished by ignoring some of the system's degrees of freedom (noise) and describing the typical behavior of a set of examined variables (signal) as the noise varies "randomly". (In Figure 2, X and Z are random signals and Y is noise). Statistics

are used to describe the typical behavior of the signal. Statistics are drawn from an **ensemble** of observations which differ from each other because the noise is different. In pure statistics and in simple systems it is easy to define the conditions leading to this ensemble but in the real world the most crucial and difficult aspect of statistical analysis is defining the conditions under which the ensemble was drawn, that is, in accurately defining the noise.

Scientific statistical analysis depends on three steps: (1) separating signal and noise; (2) defining the ensemble over which typical signal behavior is defined; (3) developing an economical and accurate statistical description of the signal over that ensemble. The first two are the more difficult; this course mainly deals with the third. There are no general rules for the hard parts; they require creativity and are judged by the essentially aesthetic criterion of the usefulness of the result. For example, to describe water temperature in San Diego we can choose to treat all days in the last 100 years as realizations of the process and define probabilities over this period. In this case we would have a single ensemble containing all the observations. Or we might separate into seasons, perhaps defining four ensembles, and produce a slightly more complex description which was, at the same time, more precise. We might go further and separate windy from calm days and years with El Niño from years without it. The ultimate in this process of adding variables to the signal category is a model in which all uncertainty is removed and the "ensembles" are so narrowly defined that all the dynamical system's specification data is determined and the outcome is deterministic.

The problem of defining the ensemble to which statistics apply is made more complicated because oceanography is an observational, rather than experimental, science. The ocean can be observed, but repeated experiments are essentially impossible. Our analog to the repeated experiment is typically defined by some observational time and place rather than being set by experimental control. Consequently, the influence of variables can not be eliminated by holding some conditions fixed; it is even difficult to know what "noise" variables are important. Because the observational ensemble can not be selected, it is critical to describe all the conditions that define it. In the San Diego water temperature example it would, for example, be important to be aware of El Niño events when measuring the "mean" temperature even if El Niños were not treated as signal.

Unfortunately, the hard parts of oceanographic analysis are not easy to study in a formal setting. They depend more on intelligence and creativity than on formal methodology. Experience is the best teacher, but to gain this one must first understand the essentials of the analysis methodologies. That will be the main purpose of the course. At the moment the foregoing discussion of the statistical perspective may seem a little windy and/or obvious. Nevertheless, most conceptual difficulties with the material to follow arises from not really understanding these seemingly simple ideas. It might be useful to reread this discussion again later. It will take on new meaning as your encyclopedia of examples expands.

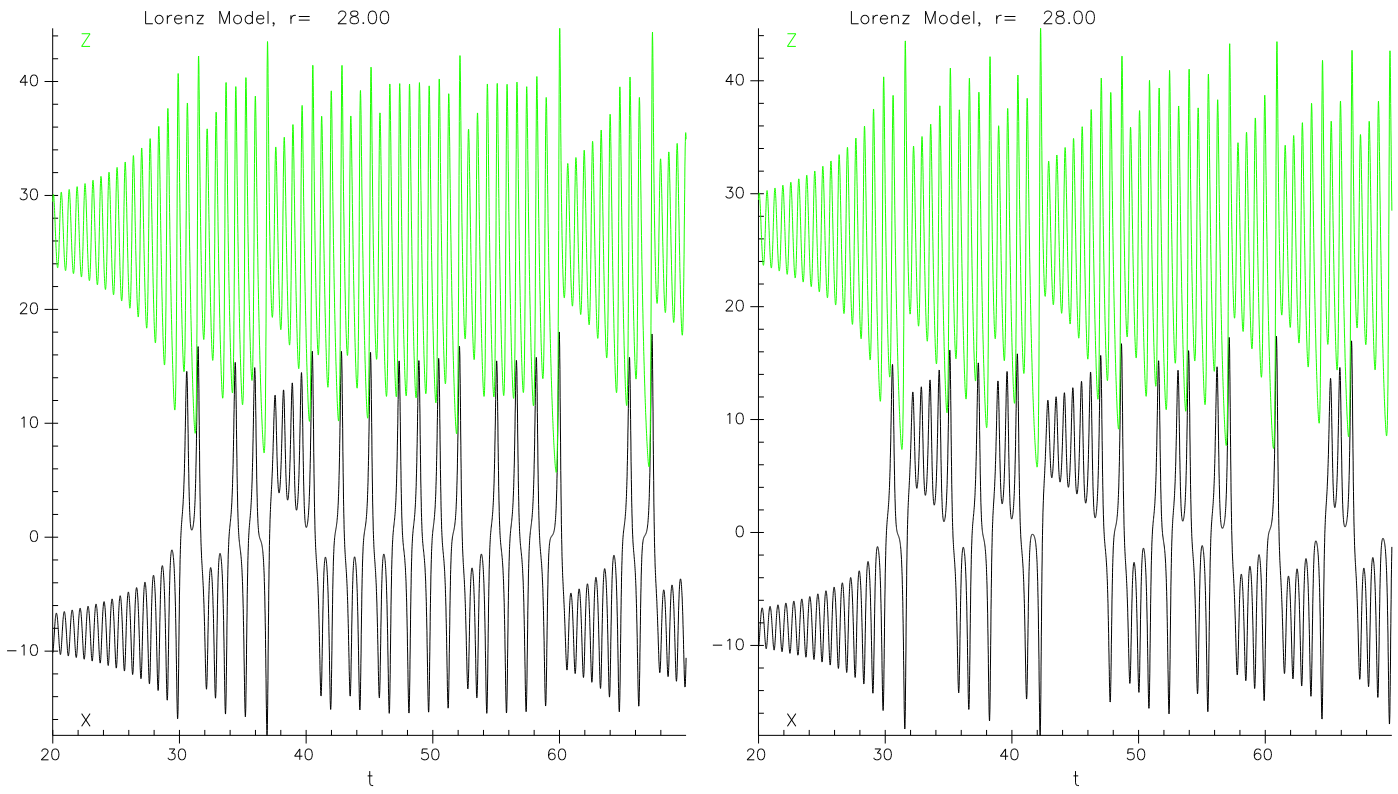

Figure 1: The variables X and Z from the initial evolution of the Lorenz model with $r = 28$. Both series start from $t = 0$ with initial conditions that differ by 0.1%. X shows that the evolution is completely different for $t > 31$. This is an example of a nonlinear, unpredictable system that is not complex (only 3 degrees of freedom).

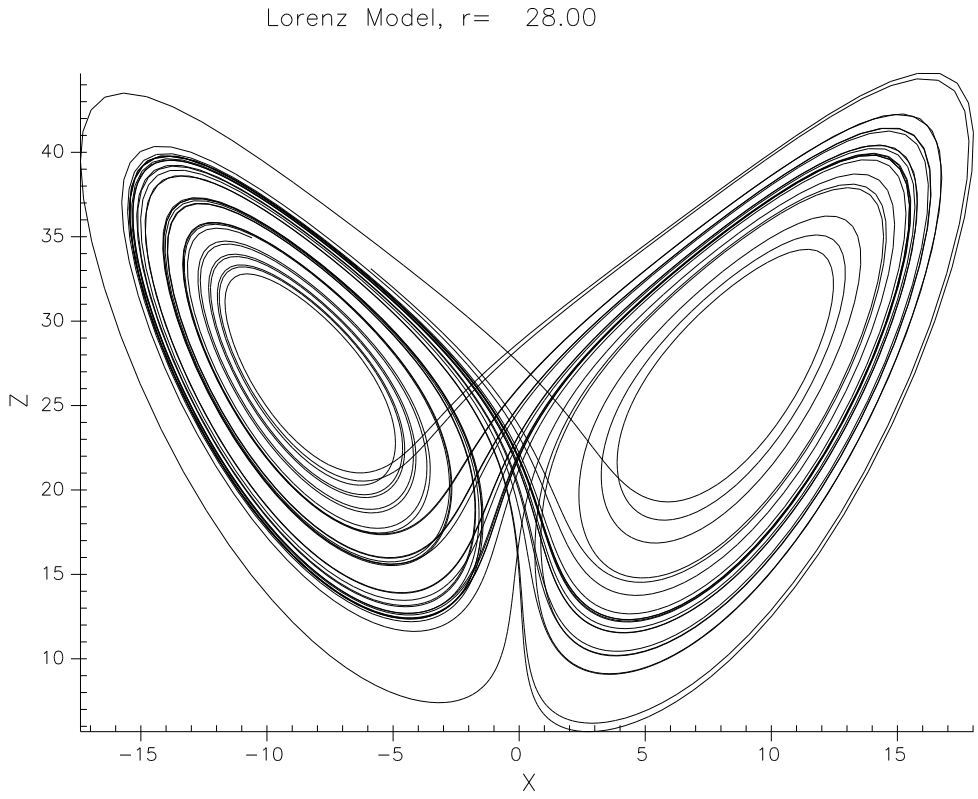

Figure 2: X-Z plane evolution of the Lorenz model shown in Figure 1. Data are taken from $t > 50$ when chaos has begun.