Big Data in Time: Progress and Challenges from Oceanography

J. M. Lilly, NorthWest Research Associates

Physical oceanography—the study of the physics of the ocean currents—depends in part upon spatial maps collected by satellites, but also to a large degree upon the venerable practice of going out on a boat and throwing instruments overboard. This leads to a sampling of the ocean circulation that is extremely heterogeneous in its spatial and temporal resolution. On top of this, the ocean is varying at a vast range of scales, some of which can be resolved and some of which will probably never be resolved. From an observational standpoint, it is like putting together a jigsaw puzzle in which one only has one percent of the pieces.

Here we review aspects of Big Data in oceanography that arise from analyzing timeseries recorded by freely drifting instruments, known as Lagrangian instruments. There are currently about 20,000 available timeseries of position and other variables from such instruments having temporal resolutions of one day or better, comprising about 30 million data points. Moving beyond first- and second-order statistics in order to access the detailed physical information contained within this data poses a considerable challenge.

Two main topics have been the focus of the author and collaborators. The first is the treatment of so-called 'coherent eddies', organized structures within ocean turbulence that manifest as oscillations in position or velocity data. These structures are believed to be important in the climate system, but their signals are notoriously tedious to extract on account of two features: firstly, they are quasi-periodic (or nonstationary) rather than strictly periodic; and secondly, they are embedded within a 'noise' field associated with fluctuations of the background currents. A relatively well-known method for analyzing oscillatory features in noise is the so-called 'wavelet ridge' method. To extract nonstationary oscillations from tens of thousands of noisy timeseries required substantial advances in the theory and practice of wavelet ridge analysis, which will be discussed during this talk. Based on this experience, we believe that this method is a powerful tool that could find application in other areas.

Most other aspects of the structure of these Lagrangian timeseries are better represented using stochastic modeling, which leads us to the second topic. A general problem in Big Data is finding suitable summary statistics that can take the place of large amounts of data. In this case, using a combination of physical reasoning and explorative analysis, we find that Lagrangian velocity spectra can be well modeled as arising from the composite effects of several simple stochastic processes. A maximum likelihood method is utilized to infer model parameters, which can be shown to have illuminating physical interpretations.

From these experiences we can point to two main challenges. Firstly, both of these methods are relatively slow, and can take days to implement on our full dataset even with a modern 12-core machine. It would therefore be desirable to have rigorous methods that can approximate or accelerate the maximum likelihood and wavelet approaches. This is particularly the case if one wishes to analyze numerical models in the same manner, as then one may be dealing with millions of timeseries rather than tens of thousands; such data volumes are currently beyond our capacity. Secondly, these analyses are based only on individual trajectories, and it would be highly desirable to have meaningful ways of accessing *joint* time/frequency/spatial structure from instruments that are nearby to one another in space and time.