Computational Challenges with Big Environmental Data

Marc G. Genton
King Abdullah University of Science and Technology, Thuwal, Saudi Arabia – marc.genton@kaust.edu.sa

In this talk, we discuss two types of computational challenges arising from big environmental data. The first type occurs with multivariate or spatial extremes. Indeed, inference for max-stable processes observed at a large collection of locations is among the most challenging problems in computational statistics, and current approaches typically rely on less expensive composite likelihoods constructed from small subsets of data. We explore the limits of modern state-of-the-art computational facilities to perform full likelihood inference and to efficiently evaluate high-order composite likelihoods. With extensive simulations, we assess the loss of information of composite likelihood estimators with respect to a full likelihood approach for some widely-used multivariate or spatial extreme models, we discuss how to choose composite likelihood truncation to improve the efficiency, and we also provide recommendations for practitioners. The second type of challenges occurs with the emulation of climate model outputs. We consider fitting a statistical model to 1 billion global 3D spatio-temporal temperature data using a distributed computing approach. The statistical model exploits the gridded geometry of the data and parallelization across processors. It is therefore computationally convenient and allows to fit a non-trivial model to a data set with a covariance matrix comprising of 10^{18} entries. The talk is based on joint work with Stefano Castruccio and Raphael Huser.

Keywords: big data; climate model output; computational statistics; spatial extremes.