Statistical Compression for Climate Model Output (IN11B-0035 AGU'17)

Joseph Guinness[1], Dorit Hammerling[2], and Yun Joon Soh[3]

[1]North Carolina State University [2]National Center for Atmospheric Research and [3]Stony Brook University

Introduction

High resolution climate simulation runs typically executed on high performance computing systems produce massive datasets. Scientists use large ensembles of model runs and perform intercomparison studies of climate models, which raises data storage concerns. For example, CMIP 6 is expected to produce more than 10 Petabytes of data, and half of the supercomputer budget for The National Center for Atmospheric Research (NCAR) has now been spent on storage hardware. Storing data is becoming a bottleneck for climate researchers that are dependent on numerical simulations. Compressing data is a natural solution. We propose a lossy statistical compression method that saves a subset of summary statistics and a statistical model, which is then used for decompression.

Compression Methods

There are two types of compression methods: lossless and lossy. Suppose the original data is X and compressed data is C. We annotate the compression process as following: $X \to C$. Furthermore, we annotate the decompression process as $C \to \bar{X}$ where $\bar{X}$ is the output of decompression. Lossless compression implies that $X = \bar{X}$, whereas lossy compression implies $X \approx \bar{X}$.

Data

We use daily mean temperature data from the CESM-LE project, which has a grid of 190×288 resolution in latitude $\times$ longitude. This gives us about 20 million floating point numbers per one year period.

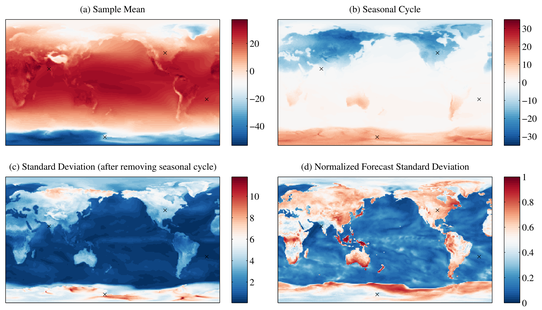

Figure 1: Maps of (a) sample mean, (b) seasonal cycle, (c) standard deviation of deseasonalized data (all in Celsius), and (d) normalized forecast standard deviation (unitless). The black crosses indicate the locations of the time series plotted in Figure 6.

Contact Information

- National Center for Atmospheric Research: dorith@ucar.edu

Statistical Compression Overview

Suppose X is the data vector (i.e. 20 million temperature values). In statistical compression, we partition the data vector as $X = (X_1, X_2)$, and we store $C = (X_1, P(X_2|X_1))$, where $P(X_2|X_1)$ is a conditional probability distribution for X_2 given X_1. The cost of storing a probability distribution is equal to the cost of storing the parameters needed to characterize the probability distribution.

Thus, in statistical compression there are two related challenges, (1) partitioning the data into the stored and unstored parts X_1 and X_2, and (2) picking a reasonable probability distribution $P(X_2|X_1)$.

At decompression, we compute $\bar{X}_2$ as either the conditional expectation $\bar{X}_2 = E(X_2|X_1)$ or as a simulation from the probability distribution $\bar{X}_2 \sim P(X_2|X_1)$, giving $\bar{X} = (X_1, \bar{X}_2)$. Thus in statistical compression, some features X_1 are preserved exactly, while other features X_2 are approximated or simulated according to a conditional probability distribution.

Data Transformation

Suppose $Y(x,t)$ is temperature at pixel x and day t. The discrete Fourier transform (DFT) over time is

$$\mathcal{Y}(x,\omega) = \frac{1}{T}\sum_{t=1}^{T} Y(x,t)\exp(-i\omega t).$$

The data $Y(x,t)$ can be recovered via an inverse DFT, and so we can represent the full dataset as X being the set of all Fourier coefficients at all locations x and frequencies ω, and X_1 is the subset of these Fourier coefficients that we preserve exactly.

Probability Distribution

A Fourier representation for the data is convenient because the coefficients from different frequencies ω_1 and ω_2 are approximately uncorrelated under a stationary model. This makes it theoretically convenient when defining $P(X_2|X_1)$, and computationally convenient when computing $\bar{X}_2$.

We model $\mathcal{Y}(x,\omega)$ as a Gaussian process in space x, and uncorrelated across ω. The Gaussian processes are nonstationary and have parameters that depend on frequency ω, intended to capture that fact that low frequency coefficients are more spatially correlated than high frequency coefficients.

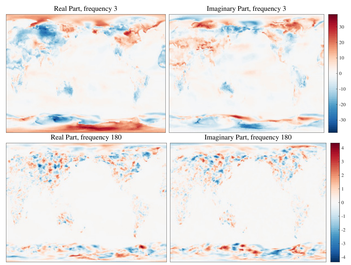

Figure 2: Maps of pixelwise Fourier coefficient values for frequency 3 and 180. The maps are shown for the real and imaginary part separately.

Results

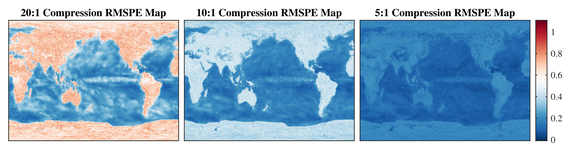

Figure 3: Maps of pixelwise RMSPE. Units are degrees Celsius.

Selection algorithm	Comp. Ratio	land	ocean	all	runtime
sequential	20:1	0.6864	0.2780	0.4367	4.33
	10:1	0.4203	0.1888	0.2762	8.30
	5:1	0.1991	0.1150	0.1444	14.13
distributed	20:1	0.6937	0.3239	0.4618	1.15
	10:1	0.4202	0.2153	0.2896	1.63
	5:1	0.1971	0.1225	0.1479	2.32

Table 1: RMSPEs and run times for three different compression ratios and the two different greedy selection algorithms. Means are pixel area weighted averages. Runtimes are shown in hours.

Selected Summary Statistics

Given a compression ratio, there is a fixed budget for the number of Fourier coefficients that can be stored. Selecting optimal subset is an intractable problem. Instead, we propose two versions of greedy selection algorithm: sequential and parallel. After initial interpolation using the selected coefficients, we repeatedly add a handful amount of Fourier coefficients in the order of largest discrepancy between the interpolated value and the original value until the storage allocated for summary statistics is full. Interpolated values are updated after each selection. The main difference between the sequential and parallel version is that in sequential version selection is considered across every frequency while in parallel version each node performs selection for its assigned frequency.

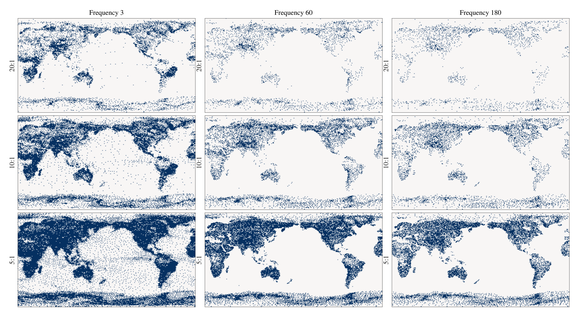

Figure 4: Each of the map indicates selected coefficients for the corresponding compression ratio and frequency.

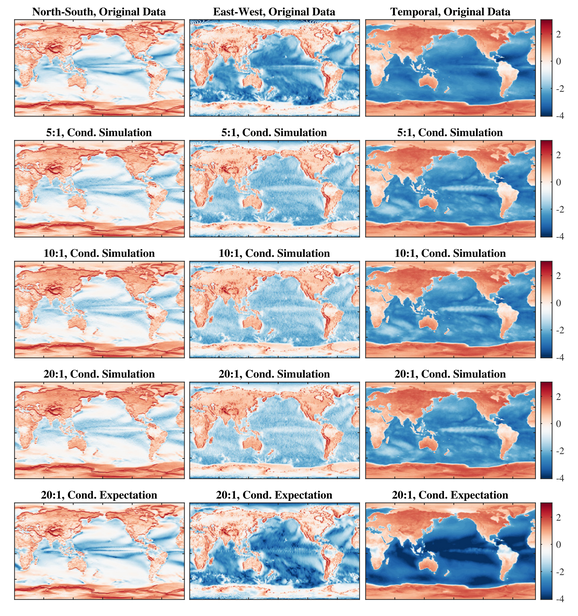

Figure 5: Maps of log contrast variances. First column are average North-South contrast variances, middle column are east-west contrast variances, and third column are one-step temporal contrast variances. First row is computed from the original data, second through fourth rows from conditionally simulated data at the three compression levels, and the last row is from conditional expectation data at 20:1 compression ratio.

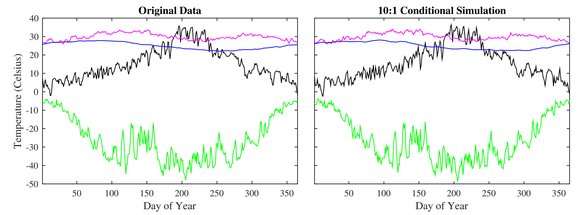

Figure 6: Original and 10:1 decompressed time series plots for Chicago (black), Mumbai (magenta), south Atlantic Ocean (blue), and Ross Island, Antarctica (green).

References

[1] Joseph Guinness and Dorit Hammerling.
Compression and conditional emulation of climate model output.
Journal of the American Statistical Association, 05 2016.

[2] Khalid Sayood.
Introduction to data compression.
Newnes, 2012.

[3] Michael L Stein.
Statistical methods for regular monitoring data.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(5):667–687, 2005.