**Machine Learning Approaches to Ensure Statistical Reproducibility of E3SM**

Effective utilization of novel hybrid architectures of pre-exascale and exascale machines requires transformations to global climate modeling systems that may not reproduce the original model solution bit-for-bit. Round-off level differences grow rapidly in these non-linear and chaotic systems. This makes it difficult to isolate bugs/errors from innocuous growth expected from round-off level differences. Following an implementation for the atmosphere model within the E3SM testing framework, we apply some classical and modern multivariate two sample equality of distribution tests to evaluate statistical reproducibility of the ocean model component of US DOE’s Energy Exascale Earth System Model (E3SM). A baseline 1-yr simulation ensemble is compared to modified ensembles – after a known non-bit-for-bit change in a model component is introduced – to evaluate the null hypothesis that the two ensembles are statistically indistinguishable. To quantify the false negative rates of these tests, we conduct a formal power analysis using targeted suites of short simulation ensembles. Each such suite contains several perturbed ensembles, each with a progressively different climate than the baseline ensemble - obtained by perturbing the magnitude of a single model tuning parameter in a controlled manner. The null hypothesis is evaluated for each of perturbed ensembles using these tests. This broad power analyses provides a framework to quantify the degree of differences that can be detected confidently by these tests for a given ensemble size (sample size), allowing model developers to make an informed decision when accepting/rejecting an unintentional non-bit-for-bit change to the model solution.