Climate Reproducibility: Updates

Salil Mahajan, Joe Kennedy, Michael Kelleher, Xylar Asay-Davis
Oak Ridge National Laboratory
Los Alamos National Laboratory

E3SM All Hands Fall Meeting 2019
CMDV-SM Reproducibility Tests (EAM) on Master

- **Nightly** tests run on Cori
  - Time step convergence test
  - Perturbation growth test
  - KS testing framework

- **Departures**: Joe Kennedy
- **Arrivals**: Michael Kelleher

- On CDASH under E3SM_Customs_Tests
  - [https://my.cdash.org/index.php?project=A CME_Climat]e

- All runs archived:
  - Large ne4 1yr F1850C5 ensemble available (>1000)
Power Analysis: Atmosphere tests

• Expand on Power Analysis:
  • More tuning parameters
    • ice_sed_ai
    • sol_factb_interstitial
    • sol_factic_interstitial
    • cldfrc_dp1
    • zm_conv_lnd
    • dcs
    • zm_conv_ocn
    • zm_conv_dmpdz

• KS testing framework most powerful:
  • detects changes of smaller magnitudes confidently
  • compared to Kernel and Energy test.

Example of Power Analysis. Probability of correctly rejecting a false null hypothesis (Power) of the test in detecting changes to a EAM tuning parameter from a control case (dcs = 400) for different short simulation (1yr) ensemble sizes (N).
Cori vs. Edison

Evaluate if E3SMv1 DECK simulations on Edison can be reproduced on Cori

- Conducted short simulation (1yr) ensembles on both Edison and Cori:
  - F1850C5-CMIP6 compset
  - ne4 (100 ensemble members)
  - ne30 (30 ensemble members)

- All three - TSC (Hui), perturbation growth (Balwinder), and KS - climate reproducibility tests passed.

- Implications: Cori can be confidently used for remaining DECK simulations
MPAS-O Reproducibility tests: Ensembles

• Testing approaches to generate ensembles

  1. Initial conditions from a long control run:
     • Conducted a 120 year long run CMPAS-O-NYF comp-set QU240 resolution
     • Still transient, non stationary: not useful as initial conditions for ensembles
     • Run longer?

  2. Multi-instance approach (work in progress):
     • Perturb initial condition to machine order precision:
       • Add perturbations to 3D temperature field initial condition
       • Save perturbed initial condition files
     • Use multi-instance (or create_clone) to generate ensembles:
       • each run reading a different perturbed initial condition file

  3. Pertlim capability for MPAS-O (near future):
     • Replicate capability within EAM to MPAS-O
     • Automatically perturb initial conditions
     • Generate ensembles by tweaking a namelist parameter.

Machine Precision Perturbations to $T$ at each grid point, $j$

$$T'_j = (1 + x')T_j$$

$x'$ is a uniform random number transformed to range from $(-10^{-14}, 10^{-14})$
MPAS-O Reproducibility tests: Approach

**Larger Null Hypothesis:** Control and perturbed ensembles belong to the same population

- Generate control and perturbed ensembles at QU240 resolution
- Evaluate 5 prognostic variables (Baker et al. 2016)
  - SSH, T, U, V, Salinity
- Ocean variability is spatially very heterogenous (as compared to the atmosphere):
  - Evaluate at each grid point.
- Conduct fine-grained null hypothesis tests at each grid point:
  - Two sample KS test: Popular non-parametric test
  - Cucconi test: Better power, rank based non-parametric test.
MPAS-O Reproducibility Tests: Approach

Correct for simultaneous multiple null hypothesis tests ($M$ grid points)

False Discovery Rate (FDR) approach (Wilks et al. 2006, Ventura et al. 2004):

- For single test, null hypothesis is rejected if:
  - Test statistic p-value ($p$) is less than a critical value, $\alpha$ (say 0.05): $p \leq \alpha$
  - For $M$ tests, $\alpha M$ would be rejected for true null hypotheses just by chance

- For multiple tests, FDR constrains critical value ($\alpha_{FDR}$) for local hypothesis tests ($H_0$):
  $$\alpha_{FDR} = \max_{j=1,2,...,M} \{ p_j : p_j \leq \alpha(j/M) \}$$
  $p_j$ are sorted p-values of $M$ tests

- Global Null Hypothesis Test ($G_0$): Reject if $p_j \leq \alpha_{FDR}$ at any grid point.

- Robust for correlated tests – e.g. spatial correlations (Wilks et al. 2006, Renard et al. 2008).
- Used in testing field significance
MPAS-O Reproducibility Tests: Preliminary Results

**Known climate changing test cases:**
- Segments of QU240 CMPASO-NYF 120 yr run (still transient)
- EAM runs as test cases

**Examples:**

<table>
<thead>
<tr>
<th>Ensemble A</th>
<th>Ensemble B</th>
<th>Ens. size</th>
<th>Variable</th>
<th>$\alpha_{FDR}$ for $\alpha=0.05$ Cucconi Test</th>
<th>Grid Points Rejecting H$_0$</th>
<th>Global Null Hypothesis Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMPAS-NYF Yrs 31-60</td>
<td>CMPAS-NYF Yrs 61-90</td>
<td>30</td>
<td>SSH</td>
<td>0.01</td>
<td>100%</td>
<td>Reject</td>
</tr>
<tr>
<td>CMPASO-NYF Yrs 31-60</td>
<td>CMPASO-NYF Yrs 41-70</td>
<td>30</td>
<td>SSH</td>
<td>0.011</td>
<td>100%</td>
<td>Reject</td>
</tr>
<tr>
<td>F1850C5 Ice_sed_ai = 705</td>
<td>F1850C5 Ice_sed_ai = 1000</td>
<td>100</td>
<td>FSNT</td>
<td>0.036</td>
<td>80%</td>
<td>Reject</td>
</tr>
<tr>
<td>F1850C5 zm_c0_ocn =0.007</td>
<td>F1850C5 zm_c0_ocn = 0.045</td>
<td>100</td>
<td>FSNT</td>
<td>0.042</td>
<td>85%</td>
<td>Reject</td>
</tr>
</tbody>
</table>
### MPAS-O Reproducibility Tests: Preliminary Results

#### Known non-climate changing test cases

**Examples:**

<table>
<thead>
<tr>
<th>Ensemble A</th>
<th>Ensemble B</th>
<th>Ens. size</th>
<th>Variable</th>
<th>$\alpha_{FDR}$ for $\alpha=0.05$ Cucconi Test</th>
<th>Grid Points Rejecting $H_0$</th>
<th>Global Null Hypothesis Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1850C5 ne4 Ens. 1-30</td>
<td>F1850C5 ne4 Ens. 31-60</td>
<td>30</td>
<td>FSNT</td>
<td>0.00</td>
<td>0</td>
<td>Accept</td>
</tr>
<tr>
<td>F1850C5 ne4 Ens. 1-30</td>
<td>F1850C5 ne4 Ens. 61-90</td>
<td>30</td>
<td>FSNT</td>
<td>0.00</td>
<td>0</td>
<td>Accept</td>
</tr>
<tr>
<td>F1850C5 ne4 Ens. 31-60</td>
<td>F1850C5 ne4 Ens. 61-90</td>
<td>30</td>
<td>FSNT</td>
<td>0.00</td>
<td>0</td>
<td>Accept</td>
</tr>
</tbody>
</table>
Planned near future work

- Generate proper **MPAS-O ensembles**
  - Known **climate-changing** and **non-climate-changing** changes

- **Power Analysis** with controlled changes to tuning parameters
  - Determine length of short simulations
  - Determine no. of ensemble members needed

- **Parallelize** python code implementation

- Evaluate other approaches for multiple simultaneous tests - resampling, etc.

- Evaluate other univariate and multivariate tests – kernel test, energy test, Lepage test, etc.

- **Identify software kernel in EAM** to target for applying **ensemble testing**:
  - RRTMG? – stochasticity from sub-columns
  - CLUBB? – stochasticity from joint pdfs of sub-grid vertical velocity, T and Q
  - MG2? – stochasticity from pdfs of mass mixing ratio, number concentration of cloud droplets and ice
  - SCM? – stochasticity from full model physics, albeit at a single column
FDR Approach: Illustration

\[
\alpha_{FDR} = \max_{j=1,2,\ldots,M} \{p_j : p_j \leq \alpha(j/M)\}
\]

Fig. 2. Illustration of the traditional FPR and FDR procedures on a stylized example, with \( q = \alpha = 20\% \). The ordered \( p \)-values, \( p_{(i)} \), are plotted against \( i/n, i = 1, \ldots, n \), and are circled and crossed to indicate that they are rejected by the FPR and FDR procedures, respectively.

Ventura et al. 2004
Cucconi Test

- Test Statistic:
  \[ CUC = \frac{U^2 + V^2 - 2\rho UV}{2(1 - \rho^2)} \]
  
  - \( U \): squared sum of ranks of samples in Ensemble A in the two sample pool of Ensembles A and B
  - \( V \): squared sum of contrary-ranks of samples in Ensemble A in the pool.
  - \( \rho \): Correlation coefficient between \( U \) and \( V \)

- Larger test-statistic indicates that Ensemble A and B come from different populations.
- Popular in other fields like hydrology, quality control, etc. (e.g. Mukherjee and Marozzi et al. 2014)