Coupled Systems I Hierarchical Testing Framework Recommendations

Developed during DTC Community UFS Test Plan and Metrics Workshop July 31, 2018. This document should be considered as guidance from the workshop participants only.

- Consider a hierarchical testing framework such as the one described in the Figure below, which was designed by DTC for evaluation of physical parameterizations. Note that DA is data assimilation and LR/MR/HR refers to low-, medium- and high-resolution.

.

- Also consider a hierarchical testing such as the CESM one, that allows turning on/off individual components (add figure).
- Is hierarchical testing necessary for the UFS evaluation? How does hierarchical testing help improve model development and the transition from research to operations?
- Which tiers should be considered? Do the tiers depend on which application is being evaluated? if so, how.
  - For each of these applications, please provide specific hierarchical tests that could be conducted: GFS, GEFS, S2S, CAM, physics
- What can we learn[1] from the simplified tiers. For example,
  - What can we learn from Single Column Models? When is it appropriate to use this tool and when is it not appropriate?
  - Running models in lower resolution can save lots of computational resources, What can be learned from testing global models in lower-resolution? What are the dangers in testing at lower resolution?
  - Is there value in testing NWP models (GFS, CAM etc.) in cold-start mode (without data assimilation)? If so, under what circumstances?
  - What is the tradeoff in resources? If we can learn a little that is of only marginal value for operational prediction, should such testing should be mandatory?
- What tests are essential for advancing prediction capability? What tests are scientifically interesting?
- Are there simplified tests not covered[2] in this diagram that should be considered? For example, single-components tests (ocean-only, sea-ice only etc.)? What can be gained from such configurations?
- How can we separate evaluation of DA from model evaluation? How does one evaluate DA schemes and quantify improvements in the DA system.
- Gather Recommendations
- Be prepared to populate 1-slide for the Outbrief (panel) Discussion on Wednesday morning
- Attend Wednesday morning Break-out Group prep to synthesize commonalities amongst the groups

---

Another aspect: what is the tradeoff in resources? If we can learn a little that is of only marginal value for operational prediction, how much such testing should be mandatory?
What tests are essential for advancing prediction capability? What tests are scientifically interesting?