Hierarchical Testing Framework

Coupled Systems - I & II

PARTICIPANTS - I	**Arun Chawla**	**Christa Peters-Lidard**
Mike Ek, Facilitator	**Jason Levit**	**Tim Schneider**
Suranjana Saha, Facilitator	**Louisa Nance**	**and others ...**

PARTICIPANTS - II	**Kevin Burris**	**Malaquias Pena**
Jim Kinter, Facilitator	**Tressa Fowler**	**Lutz Rastaetter**
Tara Jensen, Scribe	**Bob Grumbine**	**Perry Shafran**
Ligia Bernardet, Scribe	**Martin Janousek**	**Dave Turner**

What is a Model Hierarchy?

Definition from NRC (2012): A model hierarchy allows key mechanisms of climate [and NWP] response in complex models to be related to underlying physical principles through simpler models, hence giving complex models credibility. Model hierarchies also allow one to proceed in this manner by slotting out (either conceptually or in software) a model component, which could be a physics parameterization, a hierarchy of fluid dynamical solvers, a dynamical framework (e.g., regional versus global), or a complete model component (e.g., land-surface models) for a simpler or more computationally efficient one. One could even return to a box model with a handful of degrees of freedom (see Held [2005] for a good discussion of simulation versus understanding in a model hierarchy).

Finding 3.4: A hierarchy of models will be a requirement for scientific progress in order to maintain the ability to (1) systematically evaluate model components one at a time and (2) use simpler models to understand complex model behaviors and underlying mechanisms. (NOTE: It will be valuable to determine if there are operational prediction centers outside the US that employ hierarchical testing.)

Recommendation 3.2: To support a national linked hierarchy of models, the US should nurture a common modeling infrastructure and a shared model development process, allowing modeling groups to efficiently share advances while preserving scientific freedom and creativity by fostering model diversity where needed.

-- “Strategies for Developing Climate Models:Model Hierarchy, Resolution, and Complexity." National Research Council. 2012. A National Strategy for Advancing Climate Modeling. Washington, DC: The National Academies Press. doi: 10.17226/13430.

Findings and Recommendations

- *Finding*: Hierarchical coupled model testing is needed for R&D at universities and partner agencies so that testing of innovations is relevant for potential future R2O. Good for process studies; exploring parameter space; sanity checks; idealized testing.
- *Recommendation*: Every component (atmos, ocean) etc. should be tested individually with “data” components for all other parts of the coupled model. Full coupled system testing should also be done as early in the development cycle as possible.
    - Example: Space Weather community has a test harness that allows for 1 to many components to be turned on at a time providing an ensemble of solutions that allow the developers to see which combinations work well
    - Example: Feedback/process withholding experiments to understand behavior
- *Finding*: There are two strategies for hierarchical testing of the coupled model:
    - Coupling the fast-physics components first (atmosphere-land, atmosphere-waves), followed by slower feedbacks, in order to optimize the impact on the forecast
    - Coupling the slowest components first in order to achieve balanced state and minimize drift of the solution
    - NB: Some configurations may be overly sensitive to forcing data and provide misleading results

Findings and Recommendations

- *Recommendation*: In addition to a single column version of the atmosphere, reduced-complexity versions of other components are needed:
    - Ocean: mixed layer model (no dynamics)
    - Land surface: bucket (or leaky, dirty bucket)
    - Sea ice: thermodynamics (no motion)
    - Atmosphere: small radius aquaplanet
    - All: cold start (without DA) capability
- *Finding*: behavior of the coupled system is similar across a range of model resolutions, so testing at lower resolution may be suitable, depending on the application (scale-aware capability?)
- *Recommendation*: A statistical expert should be involved in experimental design to ensure that the hypothesis being tested can really be verified/falsified against benchmarks and to determine how much computation and which metrics are needed for a given test (not all tests have same value)

HIERARCHY, COUPLED MODELING

Q1: Do we need a model hierarchy?

1. Definition from: NRC (2012): A model hierarchy allows key mechanisms of climate response in complex models to be related to underlying physical principles through simpler models, hence giving complex models credibility. Model hierarchies also allow one to proceed in this manner by slotting out (either conceptually or in software) a model component, which could be a physics parameterization, a hierarchy of fluid dynamical solvers, a dynamical framework (e.g., regional versus global), or a complete model component (e.g., land-surface models) for a simpler or more computationally efficient one. One could even return to a box model with a handful of degrees of freedom (see Held [2005] for a good discussion of simulation versus understanding in a model hierarchy).

   Suggested Citation: "3 Strategies for Developing Climate Models:Model Hierarchy, Resolution, and Complexity." National Research Council. 2012. A National Strategy for Advancing Climate Modeling. Washington, DC: The National Academies Press. doi: 10.17226/13430."

Finding 3.4: A hierarchy of models will be a requirement for scientific progress in order to maintain the ability to (1) systematically evaluate model components one at a time and (2) use simpler models to understand complex model behaviors and underlying mechanisms.

Recommendation 3.2: To support a national linked hierarchy of models, the United States should nurture a common modeling infrastructure and a shared model development process, allowing modeling groups to efficiently share advances while preserving scientific freedom and creativity by fostering model diversity where needed.

We have a question for others: are there any operational centers who have embraced and employed hierarchical testing ???

- Also consider a hierarchical testing such as the CESM one, that allows turning on/off individual components (add figure).

Q2/Q3: Is hierarchical testing necessary for the UFS evaluation? How does hierarchical testing help improve model development and the transition from research to operations?
YES
Across the scales of hierarchical testing, it is especially helpful on the R&D spectrum, and potentially useful through the R2O-O2R continuum, but will require additional resources for testing in the full system by the community.

Q4: Which tiers should be considered? Do the tiers depend on which application is being evaluated? If so, how?
Individual sub-components, component testing, column testing, limited-area, regional, global, fully-coupled system (end-to-end; UFS soup to nuts).

Q5: 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?

Process studies; exploring parameter space; sanity checks; idealized testing

Q6: 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?
Testing scale aware-ability of physics; Sanity checks; coupled and uncoupled testing

Q7: Is there value in testing NWP models (GFS, CAM etc.) in cold-start mode (without data assimilation)? If so, under what circumstances?
Yes this is a common practice at NCEP (i.e. low res forecast only), then hi-res f'cast only; then full DA experiments

Q8: 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?
Tradeoff is self evident. Mandatory? - Probably not.
Although there should be benchmarks to meet to advance towards adoption in the operational system

● What tests are essential for advancing prediction capability? What tests are scientifically interesting?
See test plans

● 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?
Yes - include other components.

● How can we separate evaluation of DA from model evaluation? How does one evaluate DA schemes and quantify improvements in the DA system.
E.g. if you need to evaluate a new forward model, error covariance

HIERARCHY, COUPLED MODELING

Q1: Do we need a model hierarchy? *YES.*

Definition from: NRC (2012): A model hierarchy allows key mechanisms of climate [or NWP] response in complex models to be related to underlying physical principles through simpler models, hence giving complex models credibility. [ETC]

Finding 3.4: A hierarchy of models will be a requirement for scientific progress in order to maintain the ability to (1) systematically evaluate model components one at a time and (2) use simpler models to understand complex model behaviors and underlying mechanisms.

Recommendation 3.2: To support a national linked hierarchy of models, the United States should nurture a common modeling infrastructure and a shared model development process…

Q2/Q3: Is hierarchical testing necessary for the UFS evaluation? …improve model development & transition from R2O. *YES, across the scales of hierarchical testing, it is especially helpful on the R&D spectrum.*

Q4: Which hierarchical tiers? *From individual sub-components …to… fully-coupled* ***system*** *(end-to-end; UFS soup to nuts), including other components (ocean, sea-ice, etc).*

Q5: What can we learn from the simplified tiers, SCMs. *Process studies; exploring parameter space; sanity checks; idealized testing.*

Q6: Running models in lower resolution. *Testing scale aware-ability of physics; Sanity checks; coupled and uncoupled testing.*

Q7: Cold-start mode (without DA)? *Common at NCEP (low-res), then normal res, then full testing with DA.*

Q8: Tradeoff in resources? (Worth it?) *Benchmarks to meet to advance towards adoption in the operational system.*

We have a question for others: Are there any operational centers who have embraced and employed hierarchical testing?