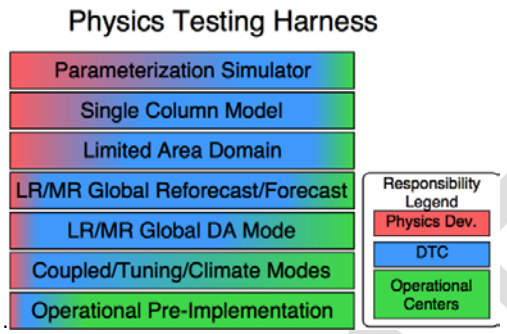
Coupled Systems II Hierarchical Testing Framework Recommendations

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.

Break-out Session 3: Hierarchical Model Testing - Discussion Topics/Assignments



- Consider a hierarchical testing framework such as the one described in the Figure above, 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.

Initial statement to get discussion started: Hierarchical testing may only be useful in a limited sense for coupled systems. We did not circle around to conclusively confirm or deny this at the end of the session.

.

- Also consider a hierarchical testing such as the CESM one, that allows turning on/off individual components.
  - See notes below about how Space Weather community does this

- Is hierarchical testing necessary for the UFS evaluation? How does hierarchical testing help improve model development and the transition from research to operations?
    - Yes. Hierarchical testing is needed during the developmental stage, which is happening at universities, NASA, laboratories. These groups need to know how to test their R&D in a way that is relevant to inform the NWS for potential future R2O. These recommendations can also help graduate students in assessing their results. Recognize that we cannot impose on non-NOAA collaborators how evaluation will be done. These guidelines can help diverse communities in standardizing their evaluation, for example, there are many groups working on sea ice and they can make use of common guidelines for evaluation.

- Which tiers should be considered? Do the tiers depend on which application is being evaluated? if so, how.
    - Single-Column Model
    - Every component (atmos, ocean) etc. should have data forcing and be tested individually.
    - Coupled testing should happen early on (start with low res)
    - Atmos + Land; then add ocean, then add sea ice, then add aerosols. Or not, Malaquias: because land takes many years to spin up, so could add ice before land. Bob: Atmos + land, then waves, then ice+ocean together, then etc.
    - Simplified ocean models (mixed layer) can be used to test atmos etc.
    - Simplified sea ice (ice does not move)
    - Simplified land model (dirty bucket)
    - Small planet testing (full resolution; fewer gridpoints)
    - Is there value in running cold starts? Allows for running longer periods bcs it is cheaper and can be run in parallel,
    - Some development can be done in cold start mode. How quickly we jump into cycled DA mode? For NWP applications (if we have a coupled NWP model), must start cycling soon because initials conditions highly affect the short-term forecast. For S2S applications, much development can be done without cycled DA.
    - 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?
        - Under certain ranges of resolutions, a model can behave very similarly, therefore one can test the model in lower res as long as one stays within this range. These ranges are different between atmosphere and ocean. See Kinter's paper.
        - Danger: Certain experiments are very dependent on forcing, e.g., SCM forcing datasets. Consider having ensemble of forcing or somehow estimate uncertainty
        - Danger: Waiting too long to run coupled.
    - 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?
        - A statistical expert should be involved in experimental design to ensure that the hypothesis being tested can really be verified/falsified and to determine how much computation is needed for a given test
- What tests are essential for advancing prediction capability? What tests are scientifically interesting?

Commented [1]: At ECMWF, each component is examined independently. Then a plan is composed on how to proceed in putting the multiple developments together.

Commented [2]: Is aquaplanet important? Or too far up the funnel?
Are idealized configurations (idealized hills, baroclinic wave etc.) important

- 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?
    - 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
    - Single-component test may be useful for systems with short temporal memories such as waves and aerosols but may not be effective for longer temporal scale memories like oceans and sea ice
- How can we separate evaluation of DA from model evaluation? How does one evaluate DA schemes and quantify improvements in the DA system.
    - Cycled testing with DA may test the DA system ability to remain stable
    - Non-cycled testing may allow for better testing of physics and individual components

Participants:
Jim Kinter
Kevin Burris
Dave Turner
Ligia Bernardet
Lutz Rastaetter
Tara Jensen
Perry Shafran
Tressa Fowler
Martin Janosek
Bob Grumbine
Malaquias Pena

- 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

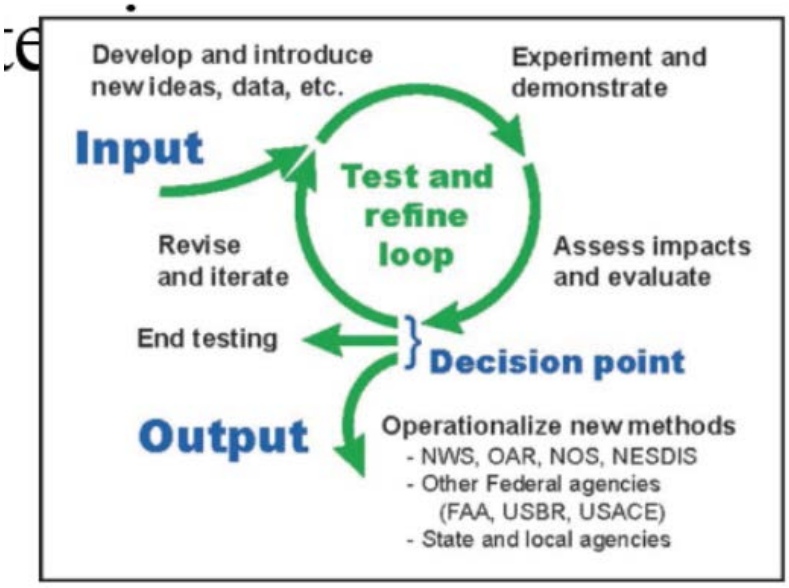
---

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?

~~~~~~~~~~~~

Transcription (or the best Tara could do - please clean up if she mis-represented something)

Jim – intro, including talking about the hierarchical testing
Dave: Some of what is included in a slide Ligia presented is in a normal development loop and don't need external input
~~~~~~~~~~~~



Dave: Some of what you just said is in a normal development loop and don't need external input

Ligia – we need both types of testing – both pre-implementation and hierarchical and look at the development process and look at the error- can we test separate parts of the system

Dave - I agree as a developer – but at what point to we manage

Ligia – testbed needs to sit a little bit higher in the funnel and helping determine

Jim – think about communities not in this building – Steve Pawson's model development shop has similar but different foci. Say they develop a better parameterization – how do they share that with EMC or other operational centers to allow EMC to bring this into their model development… At COLA – looking at marine strato-cu and may have a better entrainment parameterization… There needs to be flexibility in how you entrain innovations

Bob – we have dynamical core, sea ice from DOE, oceans and waves models from other , Space Weather mediator, each of these communities have their own processes and there is no way we can tell other shops how to do this work. There's a path into EMC that's about a 10 page check list from the software perspectives

Kevin – how do you look at something with limited input and tell if it is good

Bob –we don't need all of the computation complexity to do debugging

Jim - in the CESM world, in the past everyone was off in their corners and 6 months before a release they would couple and it would be bad so they learned to do coupled tests often. Maybe one of the dangers is if you don't do coupled testing routinely and you spend too much time in a loop and it will never be released

Ligia: After you test each component – is there a natural next step

Dave – we should probably take a step back and ask how do you get data sets for single column model

Jim – ocean modelers use the CORE core dataset (http://data1.gfdl.noaa.gov/nomads/forms/core/COREv2.html)

Ligia – in the GMTB there's ensembles of forcings –
This is a danger if sub-model being tested is overly sensitive to forcing

Dave – I agree – what is the right order of ingredients

Bob – likes to put everything in there first and if it goes nuts you go back and start figuring out what is forcing the craziness

Bob - Taking atmosphere from operational model – and sea ice model needs to be geared toward what's available to operational models

Jim – one way to answer Ligia's question – is through time-scales of feedback; there's a very rapid feedback between surface flux over land and the water cycle – LSM needs to be very tightly coupled with the land
Atmosphere – Land
Atmosphere – oceans, waves
Atmosphere – sea ice

Malaquias – suggest is the opposite that you start with coupling the longer memory and move toward shorter time-scales… Land has a spin up of 7 years… Ice is even slower

Bob suggested atmosphere – land; +waves; simplified ocean and sea ice and so on

Tara: {I was taking notes so didn't get a chance to suggest atmosphere+land+waves+aerosols+simplified ocean sea ice and so on}

Jim asked Bob if he was saying HTF – simplified atmos, simplified ocean, simplified ice (ice thermos only),

There was a general discussion that picking the right approach is difficult

{I think it was Jim} we try to get a model in a balanced state first then couple

Martin - For each model cycle – there is careful planning including a plan of how to merge the contributions together and the tests they want to run – do “small earth” tests as well

Jim synthesized – So for EC, Seasonal models cycles updates 3-4 years. Global NWP cycles updates 2x per year so that means there are 6-8 NWP upgrades for each seasonal

Lutz - In space weather – many coupled models that are developed individually and then coupled together – then there are several frameworks with the components and run it in many permutations. Testing with a framework with different configurations of components turned on to see what happens. An ensemble of number and type of components turned on. He mentioned they were not constrained by lots of obs

Tressa: Statisticians need to be involved in the development of tests and code to help do that.

Jim: what if there’s a statistician available to steer the testing – what would be the benefit or what would you advise higher up the funnel?

Tressa: The nice part of would be to do a good experimental design. You might be able to throw out some tests (e.g. ensemble design – where do you really get your improvement,
Dave: How often does it bite us that our data is not independent

Tressa: using independent obs it’s usually okay; if you’re using the analysis from the model your evaluating, it’s bad… Statistics are just numbers but comparing it to your standard is what will tell you if it’s meaningful… A skill score of 0.1 for sub-seasonal forecasts can be meaningful

Circling back around from DA

Jim: Do you gain anything with testing with cycled or non-cycled DA – in the coupled system it’s not clear – does anyone have a perspective (i.e. should it be included or not – is there any value – does it inform DA or coupled model development)

Ligia – is there any point running not cycled – cold star?t

Ivanka – non-cycled can be cleaner than if you’re throwing in flux data that it’s imperfect that can cause problems

Dave – there’s a need for both i.e. are initialization of the model so it doesn’t run away and

Bob – we need both – DA is at different levels of maturity for components – atmospheric DA – advanced; sea ice – barely existent; need to have uncoupled runs for consistency to test component

Martin - Small advantage without DA – can cover longer range of time [also parallel processing instead of serial DA]; that's something EC does

Jim – made a reforecast model for 60 years; 2x per year; ensemble initial states; don't cycle with DA; instead get the best state estimates available; need to pick data sources that got back 60 years for 60 years of forecast out to 1 year lead times and there's 3x the el nino's etc and found that there was only a slightly detectable signal with satellite being assimilated; turned out that only Stat Sig - EC state estimator for oceans so for longer time scales – not needed; for short time frame it's needed.

Dave For the 1 week fcst – do we need coupled DA? 4 week? This is a subject for research

Jim: It may be that rather than a single system with DA coupled from day 1, it may be that we have several systems with different DAs

Dave - One question we didn't get to? What can we learn by running models at lower resolution

Ligia – that's a good question because GMTB had to run at 35km rather than 13km

Martin expanded on the use of a small planet. What – make it smaller and tune all the parameters so the forces are the same but you still have a global model at 1km; aqua-planet

Jim 128km; 64km; 32 – uncoupled; 64; 32;16 – oceans is 1 deg so big mis-match;

- There's big sensitivity to forcings between 50-100km
- The results at 16km vs 32km is pretty much the same
- At the mesoscale are non-hydro – need to run at 3km; if you're running at 32km and get a
- If you take a 25km atmos – running over a parameterized ocean (ocean becomes a slave to atmos) but if resolved ocean (ocean becomes master to atmos)
- There are published recommendations;

Bob – maybe the take away is there are resolutions with break-points in skill