S2S Ensemble Test Plan

Title: Coupled Atmosphere-Land-Ocean-Sea_Ice-Waves for sub-seasonal and seasonal time frames

Contacts: Jim Kinter, Laurie Trenary

Contributors: Malaquias Peña, Melissa Ou

Date of plan: **Developed during DTC Community UFS Test Plan and Metrics Workshop July 30, 2018. This draft should be considered a work in progress.**

Introduction

The UFS community is developing a coupled model for sub seasonal (UFS-Subseasonal) to seasonal (UFS-Seasonal) time scales. This coupled model is based on the NEMS/NUOPC framework and couples the FV3-GFS global atmospheric model (including the NOAH land surface model) with the MOM6 ocean model, the CICE5 sea ice model and the WW3 wave model. EMC has focused initial development of this model at the sub seasonal time scale. The same framework will also be extended to the seasonal time scale. While the initial focus is on A-L-O-SI-W interactions, aerosols should be added to the framework when available.

Goals and Scope

Scientific questions: How far in the forecast length are the initial conditions relevant? When the model components (ocean, ice, land) become crucial to accurately predict in the S2S range? How relevant are the model resolution and the physics of the model? Do we know the sources of predictability? For certain seasons and locations: soil moisture;

Scope: The S2S test plan could be designed differently depending on whether the goal is to evaluate the prediction system pre- or after-implementation. In the pre-implementation this involves benchmarking; that is; the use of prior runs of a similar system (e.g., operational CFSv2) or a prior stable version of the UFS. The pre-implementation (a.k.a. development phase) is the more difficult particularly because the sample size is generally small. Some centers, like the UKMet have streamlined these types of testing with rapid prototyping, resulting in new versions annually. EMC has developed a pre-implementation test plan with a quick set of key variables and parameters.
Assuming that the UFS has passed the pre-implementation evaluation stage, the post-implementation test plan can be discussed next.

Experimental design

- Ensemble 35-day (180-day for UFS-Seasonal) re-forecasts that sample initial conditions over a minimum of 20 years (30 years for UFS-Seasonal), every calendar month.
- Need to aggregate data in space and time and format data so that matches what is common for users and/or best practice. (e.g. weeks 3-4 should be aggregated as a 2 week period, resolution should be relatively low like 2.5 Deg, format as tercile categorical)
- What ensemble perturbation method? Lagged or batch method? Ensemble initial perturbations in the atmosphere could be based on the NCEP Ensemble DA

system from 2012 onwards. For years prior to that, possibly the Ensemble Transform (ET) will be used. For the other components, the use of other methods and ocean reanalyses (e.g., Penny et al. and CPC) can be used.

- Is once per month enough to sample onset of tropical signals (e.g., MJO)? Should fall-winter periods be sampled more frequently? -CINDY/DYNAMO field campaigns in 2011-2012; ENSO Rapid Response 2016-2017 campaign (Dole et al., 2017)
- Determine ways to be able to evaluate S2S that needs longer training data in reforecast but in a more timely manner. Challenge is that it takes a long time to get reforecasts produced for model upgrade. Ensembles may help/replace having significant data representing events over longer time.
    - Because of the intermittent skill at this time period, verification of large anomalies (extreme events) should become standard. Perhaps "testing" should do exactly this; run the hindcasts for the most extreme cases and stronger ENSO events. If the model is worse in those cases, it may not have hope for other cases. That way testing could be done before a full reforecast is generated and the model is frozen. (Heidke and reliability may still be most useful.)
- EMC will have the benchmarks. The model code is at repository. Everybody will check out. The initial states branches. How much of that is given for the academic input.
- Different test plans: weather, S2S and Seasonal?
- Need to potentially freeze a branch of the model to obtain a reforecast. Too many upgrades precludes getting a consistent reforecast long enough to be used for centers and users that go week-2 and beyond. Can keep a separate branch of the model with more agile updates for < week-2 forecasts.

Source codes (list code repositories and name of branches; when available, add tag names)

- gerrit:fv3, https://github.com/NOAA-GFDL/FMS, https://github.com/NOAA-GFDL/MOM6 (development: https://github.com/EMC-MOM6/MOM6), gerrit:EMC_cice, gerrit:EMC_ww3, gerrit:EMC_FV3-MOM6-CICE5

Initial Conditions (initialization procedures, data assimilation, cycling)

- Atmosphere : Interpolate from archived reanalysis (CFS-v2 or ERA-Interim) states to FV3-GFS grid
- Land : As for atmosphere initialization
- Ocean : Interpolate from archived reanalysis (CFS-v2 or ECMWF ORA-S4)

- Sea Ice : As for ocean initialization
- Ocean Surface Waves : Use archived operational ½ deg initial conditions
- Ensemble generation approach should be taken into account.
- Incorporate lagged ensembles? May be only helpful for certain lead times, e.g. more weeks 3-4 than week 1-2.

Forecast periods/lengths and high-impact case studies (sample sizes)

- Period:
  - 1996-2015 (1986-2015 for UFS-Seasonal)
  - 1985-2010 (Week-2 to seasonal requirements to match reforecast/climate period, especially extremes). This period is used in reforecast to train CPC guidance products. Use of cross-validation/leave one out for verification.
  - At least 10 years? For extremes - percentiles and values. EMC has used 8-9 years of hindcasts for evaluation with success to capture at least some MJO events.
  - Ensemble perturbations from EnKF starts in May 2012
  - Insufficient sample size can especially produce unstable skill verification statistics for extremes. Results can be an incorrect or poor reflection then of the actual skill of the model.
  - Assess by aggregated time regimes, such as warm and cool seasons separately.
- Samples:
  - Cases from every calendar month (12)
- Members:
  - How many members? For the atmospheric problem, the minimum has been suggested to be about 5 to 10 members. The 2012 ESRL Reforecast Project used 11 (1 control + 10 perturbation-) members.
- Model Interval:
  - A maximum interval between re-forecasts of N days with a simultaneous ensemble of M members. Note that N and M are experimental parameters that should be evaluated for statistical validity. For example, values of N=5 and M=10 would mean sampling every 5 days (pentad) of initial states with a perturbed ensemble of 10 members. These values are arbitrary and usually chosen based on computational resource constraints, but a more rigorous evaluation of the values necessary to provide validation and verification data sets is needed.
    - Even though in this situation week-2 is outside of sub-seasonal, reforecast sensitivity tests were done by CPC for week-2 (not for longer time scales) regarding reforecast length. To not observe significant loss in skill, we need 18 years of reforecast, 6 ensemble members (5 members, 1 control), and model run once a

week. (Ou, et. al.) For weeks 3-4 and seasonal probably need more, e.g. 25 yrs. With CFS reforecast only being 1999-2009, it missed “seeing” a major El Nino event and 25 yrs would include 97/98 and 15/16.

Variable Post-processing

Produces derived variables

- For waves: use multi_1 post processing and additionally use ObsOpWaves for observation-model comparisons
- For atmosphere : Use FV3-GFS post for post processing
- For land surface: As for atmosphere
- For ocean: use ocean post package for post processing
- For sea ice: the output files are in NetCDF format

Statistical Post-processing

Calibration and bias correction

- For week-2 and beyond, the Climate Prediction Center (CPC) relies mainly on statistically post-processed model guidance. These tools would need to be evaluated.
- User-oriented methods for downstream models

Graphics

- For atmosphere : Consider using subset of CESM Atmospheric Model Working Group diagnostics (http://www.cesm.ucar.edu/working_groups/Atmosphere/amwg-diagnostics-package/)
  - Plus precipitation rate and 2m air temperature for CONUS
  - S2S Graphics
- For land surface : Consider using subset of CESM Land Model Working Group diagnostics (http://www.cgd.ucar.edu/tss/clm/diagnostics/index.html)
- For waves: Standard production wave graphics plus roughness length, Charnock coefficient, drag coefficient and/or any other variable coupled or strongly influenced by coupling.
- For ocean: mixed layer depth (to see effects of wave-coupling), SST, SSS, SSU/SSV, and SSH; SST and Precipitation rate in Nino3; also consider using CESM Ocean Model Working Group diagnostics (http://www.cesm.ucar.edu/working_groups/Ocean/metrics.html)

Data

- Need to determine some base observational datasets to use for evaluation.

Data archival

- 6 hourly output
- GFS pgrb2 (1 degree), GFS sfluxgrb files (includes land surface variables)
- All ocean and ice variables in NetCDF
- [possible to save tendency terms?]

Computational resources

Evaluation

- Objective assessment
  - Objective verification overview, software to use etc.
    - For waves: WW3 verification packages (both buoys and satellite)
  - Metrics (with table)
    - Process-oriented diagnostics
    - Phenomena
      - Teleconnections, Monsoons, MJO, Extremes, ENSO,
    - Variables sorted by scale to the ones that matter the most
- Case studies
- User-oriented. Forecaster and community participation
- Intercomparisons: NOAA S2S, WMO S2S, SubX projects models
- What ground truth observation datasets to use for what variables?
  - Assumption: Not real-time, although centers may do their own evaluation in real-time in the future.
  - Dataset differences between centers depending on unique time scale challenges.
  - For week-3/4 verification at CPC, use CPC's gauge-only precipitation and tmean temperature grids (land is interpolated station based, over water is GFS "d0" or 24-hours worth of initialization + initial runs of model). Resolution is 2-Deg for verification. Also have 1-Deg but considered potentially too high resolution verif for this time scale.
- Validation (system technical and scientific)
- Verification (e.g. objective/subjective skill)
  - Subjective evaluation:
    - This may be more appropriate for shorter term forecasts, but not necessarily longer timescales.
  - Objective evaluation:
    - Evaluate skill for specific regions that have have had known issues with previous model versions to see if there was improvement.
- What variables? What parts of the ensemble distribution (e.g. terciles, extremes)?
  - Terciles

- How do external developers outside of EMC perform evaluation?
- Partition testing between seasonal and sub-seasonal?

Deliverables and timeline

To make it relevant for model developers EMC's baseline must be taking into consideration.
Need coordination with operational groups to learn the timeline.
Sanity check measures, high priority S2S signals, specific events case studies.

Risks and mitigation

- Lack of coordination between unified efforts and individual centers doing evaluation for center-specific model guidance tools. Not understanding what pieces will be done by whom.

Questions and comments

- Who will be doing each of these tasks? Individual centers, DTC, EMC, etc.? What is the scope of these centers?
    - E.g. EMC and CPC directors coordinating on their own test plan for CFS. Need to make sure that the tests are objectively picked and not cherry picked.
    - Can provide benchmarks from NCEP centers outward to other users.
    - Met UK does freeze then produce benchmarking assessments. ECMWF has a fork approach that creates a fork for approaching seasonal forecast that is branched off compared to shorter time scales. Seasonal forecast is an independent version with slower upgrades.
    - What is the role of DTC within the centers performing evaluation?
- How important is a test plan for S2S/Ensembles?
    - Essential to have a plan to collect data
    - Needed for organized benchmarking with previous model output to compare upgrades.
    - Validation (EMC will be working on this) vs. verification
    - Validation should be handled, there are different definitions though of what validation means. Validation can be interpreted as systems testing. Technical versus scientific validation. Need to transition from validation to verification.
- What topics should be prioritized as part of the test plan?
- Should we work backward, think of challenging timescales or aspects of predictability we can leverage to build into the test plan.

References

- Production wave validation website: http://polar.ncep.noaa.gov/waves/validation/prod/ see info tab for definitions and details.
- Zhang, C., K. Yoneyama, 2017: CINDY/DYNAMO field campaign: Advancing our understanding of MJO initiation.
- UPPs Post: https://dtcenter.org/upp/users/docs/user_guide/V3/upp_users_guide.pdf.v3.0

- Ou, M., M. Charles, and D. Collins, 2016: Sensitivity of Calibrated Week-2 Probabilistic Forecast Skill to Reforecast Sampling of the NCEP Global Ensemble Forecast System
-

Appendix: Definitions and acronyms

Derived variable post-processing - Post processing unified post processor (UPP) software produces derived variables. This type of post-processing involves creating these calculated variables from the raw base variables from the model.

Statistical post-processing - Model post processing using statistical methods to calibrate and bias correct the model.

Verification --Integrity of the model components

Validation --Comparison with the observations.

Scientific Validation: ability to address classes of geophysical problems (applications) for which it was designed.

Unified model - Unified framework not unitary model. There will be different pieces. Framework is built to have different options, e.g. for physics packages, unlike previously it would be considered different models. Unified code base with various pieces.

S2S - Sub-seasonal is week 3-4, to seasonal.