2018 DTC Community Unified Forecast System Test Plan and Metrics Workshop

Workshop July 30 - August 1, 2018 (Report issued September 30, 2018)

Table of Contents

Executive Summary

The workshop was very productive in identifying how to construct a test plan, choosing metrics for verification and validation (V&V), and creating recommendations for charting the way forward for the development of a hierarchical test harness. Additionally, development of four sample test plans were initiated, which while not binding, give the Unified Forecast System (UFS) community a good starting point for model assessment.

While the importance of using best practices in the development of UFS software should not be overlooked, this aspect of V&V was not a topic of this workshop.

Recommendations

The following recommendations are put forth to the UFS Steering Committee, the UFS Working Groups (WGs), and NOAA leadership.

1. The generic test plan outline, the metrics spreadsheet, and the recommendations for hierarchical testing created in this workshop should continue to be worked on as living documents. This will enable further development and prioritization of the test plans and metrics in the spreadsheet involving a larger and more representative group.
2. A clear mapping should be developed linking UFS applications, WGs, and stakeholders. Involving the latter can be done in coordination with the NOAA Science Advisory Board's Environmental Information System WG, which engages with stakeholders across the weather enterprise. This will help the WGs to continue evolving the test plans and metrics for the various aspects of the UFS.
3. Pre-implementation test plans for each application should be developed with input from the WG(s) related to that application. In addition, an external and independent review of each test plan is needed. For example, a review by scientists from multiple US and international organizations with experience in developing test plans (e.g. European Center for Medium-Range Weather Forecasting (ECMWF) Technical Evaluation report), could be very valuable.
4. Reports of test results should be made available to the community so that the interpretation of results can be made in a transparent way.
5. Additional workshops will be needed in the future to cover aspects not discussed in depth in this workshop, such as observational aspects (availability, quality control, etc.), data assimilation (DA) use for V&V, hierarchical testing for the DA application, and V&V of ensemble and probabilistic applications.
6. The results of this workshop should be presented at the January 2019 AMS annual meeting and other community venues.
7. A clear mapping is needed between this guidance to funding for groups engaged in these projects, so that sufficient human and computational resources are allocated. For example, there is a clear need to fund the creation of a hierarchical testing framework,

implementation of metrics, and development of test plans.

Overview

The 2018 DTC Community Unified Forecast System Test Plan and Metrics Workshop was held at NOAA's National Center for Weather and Climate Prediction on July 30 - August 1, 2018. The major goal of this workshop was to work towards a community test plan with common validation and verification metrics for the emerging UFS. The plan will serve as a guide for the weather and earth system prediction community for testing and evaluating new developments for the UFS models and components. Through standardized hierarchical testing, comparison of both historical and real-time forecasts with observations and analyses will be conducted.

The workshop had a mix of presentations, discussion periods, and working sessions in which participants contributed to the three topic-based breakout sessions: test plans, metrics, and hierarchical testing. The last activity in the workshop was a summary of the working sessions' discussions by their leads, which was presented to workshop participants and members of the UFS Strategic Implementation Plan (SIP) meeting. The agenda, along with links to the presentations, can be found here.

Workshop organizing committee and participants

The workshop organization was led by the Developmental Testbed Center (DTC), and the organizing committee was representative of various aspects of the V&V enterprise, including voices from those working on research, development, transitions, and operations. The membership of the organizing committee was:

- Curtis Alexander (NOAA/ESRL/GSD)
- Ligia Bernardet (CU/CIRES at NOAA/GSD and DTC)
- Tara Jensen (NCAR and DTC)
- Jim Kinter (GMU/COLA)
- Sherrie Morris (NOAA OSTI)
- Jason Levit (NOAA/NCEP/EMC)
- Ryan Torn (University at Albany, SUNY)
- Ivanka Stajner (NOAA OSTI on detail to NOAA/NCEP/EMC)

The workshop was attended by approximately 100 participants, cross-cutting through various sectors of the V&V community, including international (Taiwan Central Weather Bureau and ECMWF), universities, National Aeronautics and Space Administration, NOAA (National Environmental Satellite, Data, and Information Service, Office of the Federal Coordinator for Meteorological Services and Supporting Research, National Weather Service, and various research laboratories), testbeds, Navy, US Air Force, and the private sector.

Topics Discussed

The workshop talks were structured to cover a broad range of topics. At the beginning of the workshop, the important topic of hierarchical testing was introduced, making it clear that different tests are needed during the research, development, transition to operations, and operational phases of a given innovation. The importance of modular testing, focusing on individual parts of the modeling system, as well as tests of simplified configurations of the modelling system, were stressed. Test plan considerations, such as sample size and independence of the elements in the sample, were discussed. The strategy for stating hypotheses, as well as the process for choosing metrics to answer the questions posed, were covered. Finally, the status of the Model Evaluation Tools (MET), which is extensively used by the Numerical Weather Prediction (NWP) community and has been identified by NOAA's Next-Generation Global Prediction System (NGGPS) V&V group as the software for evaluation of the UFS, was presented.

The subsequent presentations covered topics in weather, medium-range, and subseasonal-to-seasonal (S2S) applications, with special highlights given to ensembles and process-oriented diagnostics. Finally, a special session on verifying coupled applications was held, identifying verification needs and strategies for the marine (coastal, ocean and sea ice), land, hydrology, atmospheric composition and chemistry components of the modeling system.

Three breakout sessions took place, in which participants further discussed how to prepare a test plan, how to choose metrics, and the merits and usability of hierarchical testing.

Test Plans

During this breakout session, workshop participants were divided into five groups, which discussed four specific instances of UFS-relevant test plans. The participants were provided with a draft of the plans and asked to discuss, refine, and edit them. The following plans were produced by the participants: a) choice of physics suite for the operational FV3GFS v2 (which at the time was scheduled for Q2FY20), b) upgrade to the sea ice drift model and the Global Real-Time Ocean Forecast System (RTOFS-G), c) implementation of the operational ensemble/S2S system, and d) operational implementation of the Convective Allowing Model (CAM) configuration of the UFS (two groups focused on the latter). Note that the test plan for atmospheric subgrid-scale physical parameterizations, referred to as the model "physics", is "higher in the funnel", that is, it is not a test of a final model prior to transition to operations, but instead is designed to determine a single aspect of the model, i.e., the physics suite. Additionally, note that the test plans should be considered as a non-binding draft, to be further refined by the UFS WGs.

The leaders of the test plan groups later synthesized the most important aspects discussed in their groups. Table 1 has links to the test plans themselves and the syntheses. Additionally, the outline of a generic test plan was made available for future tests.

Outline for a generic test plan	Test Plan	
Physics for FV3GFS v2	Test Plan	Synthesis
Upgrade to the sea ice drift model and RTOFS-G	Test Plan	Synthesis
Ensemble and S2S	Test Plan	Synthesis
CAM	Test Plan	Synthesis

Test Plans Summary

- There was overall agreement that robust and transparent test plans can lead to an improvement of the UFS. There was agreement on the topics that need to be present in all test plans, and an outline was generated for future use (link to general test plan outline).
- Any test plan must have an end goal, and it needs to be reproducible, transparent, and objective. The community should have the opportunity to provide input on test plans for upgrades to UFS applications. The NGGPS/SIP WGs are forums in which test plans can be discussed.
- 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.
- Stakeholders need to be involved in planning of the testing of UFS applications and validation of their performance (NWS National Centers, stakeholders in aviation, air quality, severe weather, etc.).
- Different stakeholders may have different ways of approaching a test plan and different metrics to measure success, and, thus, a unified modeling system that is trying to address the needs of multiple stakeholders may have multiple conflicting test plans and metrics. How this would be arbitrated was not discussed in detail, but one breakout group suggest an adjudication board.
- Test plans need to consider the balance between improving the science versus improving the predictions, as frequently improving the science decreases the accuracy of the predictions on the short term.
- The modeling system should be evaluated before and after model calibration since differences may be significant.
- Packages developed for components (e.g. the North American Land Data Assimilation System - NLDAS) can provide numerous benefits for the evaluation of individual component and should be considered for inclusion in the UFS testing.
- The four test plans created in the workshop are an excellent, yet non binding, starting point for testing the applications.

Metrics

During this breakout session, workshop participants were divided into five groups, which discussed the metrics for five aspects of the UFS: CAM, Global Weather, S2S, Atmospheric Chemistry and Composition, and Process-Oriented Diagnostics for Multiple Scales. The participants contributed to a draft spreadsheet (link) containing *tabs* for various aspects of the metrics. It should be noted that the initial draft was prepared by the workshop organizers by collecting input from the various NGGPS/SIP WGs. Some WGs, notably the CAM one, spent considerable time discussing metrics before the workshop, therefore, some tabs are more complete than others. Note that not all tabs of the spreadsheet were worked on during the workshop, since there were only five breakout groups.

The leaders of groups later synthesized the most important aspects discussed in their groups. The table below has links to the syntheses.

CAM	Synthesis
Global Weather	Synthesis
S2S	Synthesis
Atmospheric Chemistry and Composition	Synthesis
Process-Oriented Diagnostics	Synthesis

Metrics Summary

- Process-oriented verification is needed to understand the model's deficiencies from a physical perspective. It emcompasses all domains (from land to boundary layer to clouds), as well as radiative feedbacks and teleconnections.
- The group identified an unexplored potential to better leverage DA to examine quality control, bias correction, analysis increments, and verification of forecasts using global NWP analysis.
- Some novel metrics are needed for global weather, such as the analysis of the model's diurnal cycle (of radiation, precipitation etc.), precipitation/clouds, tropics (Madden-Julian Oscillation as a source of predictability) teleconnections, polar weather, aerosol/chemistry, stratosphere (in particular, Quasi-Biennial Oscillation), land-ocean coupling, and feature-based verification of phenomena (e.g. cyclones and jet streams).
- The distribution of errors should be examined to show variability around the mean, not just mean itself, to facilitate verification of extremes.
- Participants suggested looking at how current recommended metrics align with the perspective of the Weather Research and Forecasting Innovation Act of 2017, which

includes: drought, fire, tornado, hurricane, flood, heat wave, coastal inundation, winter storms, high-impact weather, snowpack, and sea ice.

Hierarchical Testing

During the workshop, it was discussed that different tests may be needed during the research, development, transition to operations, and operational phases of a given innovation. The importance of modular testing, focusing on individual parts of the modeling system, as well as tests of simplified configurations of the modelling system, were stressed and defined as hierarchical testing. During this breakout session, workshop participants were divided into four groups, which discussed three specific instances of hierarchical testing focusing on atmospheric physics, tropical cyclones, and coupled systems (two groups focused on the latter). Participants relied on a set of questions provided by the workshop organizers to prepare recommendations.

The table below lists the recommendations and synthesis provided by the groups.

Atmospheric Physics	Recommendations	Synthesis
Tropical Cyclones	Recommendations	Synthesis
Coupled Systems	Recommendations	Synthesis

Hierarchical Testing Summary

- Hierarchical testing is needed for research and development (R&D) at universities and partner agencies so that testing of innovations is relevant for potential future transition of research to operations.
- Different applications (CAM, S2S, etc.) likely have different requirements for hierarchical testing with some overlap among their needs.
- Hierarchical testing for coupled models: every component of the coupled model (atmosphere, ocean, aerosol/chemistry. etc.) should be tested individually with “data” components for all other parts of the coupled model. Fully coupled system testing should also be done as early in the development cycle as possible.
- Hierarchical testing for atmospheric physics: tiers identified were Single-Column Model (SCM), Large-Eddy Simulation (LES) model, case studies, limited area configurations, cold starts, cycled data assimilation, and multi-year integrations. Also, physics needs to be tested at all scales, from CAM to S2S, so we can make progress toward physics unification.
- Hierarchical testing workflows must be user friendly and available to the community. They should be standardized so results can be interpreted consistently throughout the community. Workflows must support R&D and transition of research to operations.

Appendix: Acronyms

CAM - Convective Allowing Model
CIRES - Cooperative Institute for Research in Environmental Sciences
COLA - Center for Ocean, Land, and Atmosphere Studies
CU - University of Colorado
DA - Data Assimilation
DTC - Developmental Testbed Center
ECMWF - European Center for Medium-Range Weather Forecasting
EMC - Environmental Modeling Center
ESRL - Earth System Research Laboratory
FV3 - Finite-Volume Cubed-Sphere dynamical core
FV3GFS - Global Forecast System employing FV3
GMU - George Mason University
GSD - Global Systems Division
LES - Large Eddy Simulation
MET - Model Evaluation Tools
NCAR - National Center for Atmospheric Research
NCEP - National Centers for Environmental Prediction
NGGPS - Next-Generation Global Prediction System
NLDAS - North American Land Data Assimilation System
NOAA - National Oceanic and Atmospheric Administration
NWP - Numerical Weather Prediction
NWS - National Weather Service
OSTI - Office for Science and Technology Integration
R&D - Research and Development
RTOFS-G - Global Real-Time Ocean Forecast System
S2S - Subseasonal to Seasonal
SCM - Single-Column Model
SIP - Strategic Implementation Plan (for the UFS)
SUNY - State University of New York
UFS - Unified Forecast System
V&V - Verification and Validation
WG - Working Group