Community Unified Forecast System (UFS) Test Plan and Metrics Workshop

The Unified Forecast System (UFS) is a community-based, coupled comprehensive Earth system modeling system. The UFS numerical applications span local to global domains and predictive time scales from subhourly analyses to seasonal predictions. It is designed to support the Weather Enterprise and to be the source system for NOAA's operational numerical weather prediction applications. (UFS SC)

What evidence do we need to operationalize a research result?

Workshop goal

Establish and publish a **test plan** that specifies a testing procedure for UFS development and for selection of candidate configurations for operational implementation.

Test plan and metrics for the Unified Forecast System (UFS)

What does success look like?

- What properties of the system need to be evaluated to demonstrate success?
  - Conservation / balance
  - Forecast error magnitude / growth with forecast length
  - Representation of phenomena (e.g. hurricanes, severe weather, MJO)
  - User-relevant thresholds / user subjective evaluation
  - Computational resources
- What metrics will be used?
  - Standard statistics
  - Categorical/threshold statistics
  - Object-based diagnostics
  - Process-oriented diagnostics
  - Ensemble metrics
- How will these metrics be applied?
  - Model complexity/hierarchical testing
  - From initial development to pre-operational testing
  - Test cases / testing periods / sample size

Outcome:

Test plan that specifies

- What constitutes success or improvement
- The properties that will be evaluated
- How these properties will be evaluated
- Succession of tests from initial sanity checks to a comprehensive coupled system evaluation and the final pre-operational evaluation
- Rigorous and robust evaluation

and at the same it is written

- To allow rapid evaluation and transition of improvements to operations
- Considering computational and personnel resources

Transparency and Applicability

Community Buy-In

- Community wants to know the methods and criteria used to implement new R&D efforts in operations
- Community needs to know the targets are so they can develop accordingly and assess their R&D efforts along the way
- Robust and consensual criteria support community inclusiveness and buy-in

Transparency and Applicability

Who will use and benefit from workshop outcomes

- NCEP Operations
- EMC
- Testbeds
- Developers
- Researchers

Both operational and R&D predictions need to be evaluated!

Toward a Comprehensive Test Plan

What is in a Test Plan

- Motivation, goals, configuration (including use of hierarchical testing), initialization, test periods, evaluation metrics etc.
- No one-size fits all!
- Test plans for R&D differ from pre-implementation evaluation
- Test plans differ among applications (CAM vs S2S etc.)

Outcomes

At the end of the workshop we will have

- Four sample test plans: CAM, Physics for FV3GFSv2, Ensemble/S2S, & Marine
- Recommendations on metrics for assessing prediction by various applications and phenomenon
- A way forward for testing other aspects of the UFS