R2O Initiative

Next Generation Global Prediction System (NGGPS)
Dynamic Core Testing Plan

01/21/2016 v2.1

REVISION HISTORY

Rev.	Date	Author	Description of Change

DTG Signature Acknowledgement Statement:
This Test Plan is effective on the date of approval by the NGGPS Project Manager. Review of this dynamic document will be conducted as deemed necessary by the DTG and/or NGGPS Project Team as tests, results, HPC availability, and evaluations warrant. The latest date of amendment constitutes the new effective date unless some later date is specified.

_____Signed _____ Date ____ February 4, 2016_____
NGGPS Program Manager: Fred Toepfer

Contents

Executive Summary

As part of a Research to Operations (R2O) Initiative, the National Oceanic and Atmospheric Administration (NOAA) National Weather Service (NWS) plans to produce a state-of-the art prediction model system, which will be readily adaptable to and scalable on evolving high-performance computing (HPC) architectures. The modeling system will be designed to produce useful forecast guidance to 30 days. This Next Generation Global Prediction System (NGGPS) will be the foundation for the operating forecast guidance system for the next several decades. Current research and development efforts both inside and outside NWS, including the Navy, NOAA laboratories, NCAR, the university research community, and other partnership efforts, will contribute to the development of this prediction system.

Selecting a non-hydrostatic atmospheric dynamic core (dycore) is the first step in building the NGGPS. Six dycores currently being developed and modified from a variety of institutions are viewed as potential candidates to be evaluated for the new system. In August 2014, modelers attended a workshop to discuss ideal dycore requirements/attributes for the NGGPS. Since then, many telecons, workshops and meetings have occurred to develop the criteria and tests best suited to evaluate what characteristics are predicted to be the most beneficial in the next dynamic core.

This test plan for the selection of a dycore describes these attributes and associated tests to evaluate the dycores. Tests and evaluations of dycore response will include criteria of the fidelity of simulating a variety of common atmospheric phenomena, dycore performance and scalability, conservation properties, minimal grid imprinting, and the capability to support regional nesting (both static and moving) or static mesh refinement capabilities. The NGGPS project will also leverage dycore evaluation data from ongoing High-Impact Weather Prediction Project (HIWPP) activities. A Dycore Test Group (DTG) will be formed to conduct an overall assessment of these tests and evaluations. Assessment results will be provided to NOAA (NWS) management who will then make an overall business case decision on the selection of the next dycore. This test plan details the process of dycore evaluations that will be used in preparation of the DTG assessment for NWS management.

I. Introduction

This document details test procedures for the evaluation of candidate dycores for use in the NOAA NWS Next Generation Global Prediction System (NGGPS) project. Material has been extracted and adapted from the following related documents:

- "R2O NGGPS – Dynamic Core Evaluation Workshop Summary" (R2O NGGPS Summary 08122014 distributed.docx) transmitted via email 13 August 2014
- "NGGPS Benchmarks Effort" (NGGPS Benchmarks 01.docx) transmitted via email 8 August 2014 or subsequent versions

Approaches, tests and/or procedures described in this document may and will require further refinement as the modeling groups, committees and Dycore Testing Group (DTG) proceed with this effort.

II. Participating Dycores

The six candidate dycores are listed below, with sponsors in parentheses.

MPAS (NCAR) – Model For Prediction Across Scales: Unstructured grid with C-grid discretization

FV3 (GFDL) – Finite Volume on the cubed sphere with 2-way regional-global nesting capability: Can be run in the more efficient hydrostatic mode with a run time switch, supports both height-based coordinate and mass-based coordinate

NIM (ESRL) – Non-hydrostatic Icosahedral Model

NEPTUNE (Navy) – Non-hydrostatic Unified Model of the Atmosphere: Flexible grid with adaptive mesh refinement

NMMB-UJ (EMC) – Non-hydrostatic Multi-scale Model on Uniform Jacobian cubed-sphere

GSM-NH (EMC) – Global Spectral Model, Non-Hydrostatic: Non-hydrostatic extension of Semi-Lagrangian Spectral model (as available)[1]

As of August 2014, GFS physics was not running in all dycores but ongoing efforts to provide a repository with a common version of the GFS physics suite for all models to adopt the code and perform testing succeeded in December 2015, and GFS physics was implemented in the two dycores participating in Phase 2 evaluations. The immediate minimum desire for each dycore is that it has a tracer capability that can be measured for conservative properties. The availability of a common GFS physics interface that enables all the models to run with same physics codebase is a critical dependency and will provide insight on conservative properties, grid imprinting, scalability, and simulation fidelity with idealized cases.

[1] The GSM-NH did not participate in any of the tests.

III. Testing Method

a. Benchmark

Ultimate operational implementation of the selected future dycore will be based on a business case decision by NWS management. The existing NWS operational model capability at the time of the decision will be a significant factor as a benchmark in the selection process. The final selection process by NWS management, however, is beyond the scope of this document.

b. Testing Procedures

Testing will proceed through two separate batteries for Phase 1 and Phase 2 evaluation criteria. Each battery will contain tests of all evaluation criteria at that phase. Testing procedures should be negotiated among all participants and be generally acceptable to the dycore group but definitely acceptable to the DTG and NCEP/EMC. Ongoing HIWPP dycore testing efforts will be leveraged to address evaluation criteria where applicable. Conflicts on testing procedures and any scoring or ranking must be resolved by the DTG described in Section IV.b.

Phase 1 evaluation criteria represent fundamental desired attributes of a model dycore. As such, Phase 1 evaluation criteria will serve as the basis for the first stage, Phase 1, screening of dycore candidates. Phase 1 criteria are provided in Section V. Table 1. Phase 2 evaluation criteria will be used in further refining the assessment of dycore candidates following completion of the Phase 1 evaluation. Phase 2 evaluation criteria are provided in Section VI. Table 2. The DTG will be expected to review and refine Phase 2 evaluation criteria prior to the Phase 2 evaluation.

c. Advanced Processors

The future operational HPC configuration could include advanced processors (AdPs), such as the Intel Massively Integrated Cores (MIC) or Nvidia's Graphical Processing Unit (GPU). Candidate dycores codes will be evaluated on a number of software quality criteria, including maintainability, extensibility, and performance portability to processor architectures (including AdPs) anticipated over the life of the NGGPS.

d. NCEP HPC Implementations

The expected evolution of NCEP's HPC capability from the current IBM "Phase 1" system is as follows. IBM Phase 2 will be operational in calendar 2015 and will be similar architecture to the current Linux-based CPU system. Procurement for a new HPC system was awarded in summer 2015 and on schedule to be operational in summer 2016. It is unlikely that the first phase of this new acquisition will have AdPs so that 2019 is the earliest practical date when they could become operational, most likely as a fraction of a heterogeneous operational system.

If the AdP technology is fully established, and meets operational requirements for system stability, performance and other requirements in the 2019 HPC implementation, an operational system employing AdPs is possible by 2021. On current (IBM Phase 1) hardware, the GFS executes on approximately 1600 processor cores.

e. Project Time Line

Ongoing HIWPP activities are expected to address some aspects of NGGPS dycore testing. The HIWPP efforts will be leveraged and augmented, as necessary, to complete Phase 1 tests by April 2015. Phase 2 testing is scheduled for completion by April 2016. One outcome of Phase 1 testing will be a down-selection of the current dycores to a smaller set of candidates for Phase 2 testing. At the end of Phase 2 testing, an overall assessment of dycore test results will be delivered to NWS management by the DTG. This assessment will be one source of information used to make a final business case decision on the dycore selection for the NGGPS. If the decision is to implement a currently non-operational dycore, then development of the operational dycore will proceed in tandem with that of the new dycore (including incorporation of all operational requirements into the new dycore) and comparisons will continue to be made between these two systems until the new dycore is implemented operationally.

IV. Associated Committees and Groups

a. Advanced Computing Evaluation Committee (AVEC)

The purpose of the AVEC is to conduct and oversee technical aspects of dycore computational performance testing and provide technical evaluation of results.

The AVEC will be chaired by John Michalakes, with Phase 1 technical assistance provided by:

Mark Govett – ESRL
Bill Skamarock - NCAR
Rusty Benson – GFDL
Henry Juang – NCEP/EMC (GSM-NH)
Tom Black – NCEP/EMC (NMMB-UJ)
Alex Reinecke – Navy

Phase 2 technical assistance will be provided by:

Mark Govett – ESRL
Rusty Benson – GFDL
Michael Duda – NCAR
Thomas Henderson – ESRL
Mike Young – EMC

The AVEC will provide reports on procedures and performance to inform a down-selection decision at the end of Phase 1 testing. The AVEC will coordinate with modeling groups on testing evaluation criteria to design a fair and objective benchmark methodology and set of evaluation criteria that address criteria pertaining to computational performance, scalability and suitability of model software for next-generation HPC architectures. Modeling groups will be responsible for providing codes, data, verification criteria, and code-specific technical advice and assistance needed to complete the benchmarks and evaluation in a timely fashion. AVEC's responsibilities will include work with NOAA management to arrange access to computational resources suitable for preparation of the computational performance benchmark cases and for conducting the benchmarks. An initial estimate of the requirement for conducting benchmarks is dedicated access to a system comprising >100K current generation conventional Intel processing cores for two short (6-8 hour) periods over the span of a week near the end of the Phase 1 benchmarking period.

The AVEC role continues in Phase 2 with benchmarking and evaluation of computational performance, and HPC readiness. In spring 2016, AVEC will report back to the NGGPS program on details of the benchmarking methodology, cases, model configurations, computational resource requirements, schedule, and results of the Phase 2 evaluations.

b. Dycore Testing Group (DTG)

The purpose of the DTG is to review the technical aspects of all dycore testing and provide evaluation of results in written reports to NWS management (anticipated on completion of both Phase 1 and Phase 2 testing). The DTG is also available to provide guidance on outstanding issues relayed from the AVEC or NGGPS Project Management Team regarding the preparation for and conduct of dycore performance testing. Each candidate dycore shall have one representative on the DTG. A process for internal DTG decision-making will be developed internally by the DTG and proposed to NWS management for approval.

Chair: Dr. Ming Ji, Director, NWS Office of Science and Technology Integration

Membership:
Consultant: Dr. Robert Gall, University of Miami
Consultant: Dr. Richard Rood, University of Michigan
Consultant: Dr. John Thuburn, Exeter
Superintendent, Naval Research Laboratory Monterey: Dr. Melinda Peng (Acting)
Director, Geophysical Fluid Dynamics Laboratory: Dr. Venkatachala Ramaswamy
Director, Global Systems Division, ESRL: Kevin Kelleher
Director, Environmental Modeling Center, NCEP: Dr. Hendrik Tolman
Director, Mesoscale and Microscale Meteorology Laboratory, NCAR: Dr. Chris Davis
NGGPS Program Manager: Fred Toepfer /Dr. Ivanka Stajner (Alternate)
Ex Officio - Test Manager: Dr. Jeff Whitaker
Ex Officio - AVEC Test Manager: John Michalakes
NGGPS Staff: Steve Warren/Sherrie Morris

c. NGGPS Project Management Team

The NGGPS Project Lead, supported by the NGGPS Project Management Team will assist in coordinating any issues regarding conduct of the NGGPS dycore test plan. The NGGPS Project Management Team will coordinate organizational participation in the testing and will coordinate programmatic and administrative support as necessary.

V. Dycore Evaluation Criteria and Phase 1 Testing

Evaluation criteria for Phase 1, the initial phase of dycore testing, are listed in Table 1. **Evaluation criteria for Phase 1 (and Phase 2) testing are not pass/fail criteria. A "no" answer to a "yes/no" evaluation will be compiled as one factor to be considered along with remaining evaluation data in a final decision on model preference.**

Table 1. Phase 1 Testing Evaluation Criteria

Phase 1 Eval #	Evaluation Criteria
1	Bit reproducibility for restart under identical conditions
2	Solution realism for dry adiabatic flows and simple moist convection
3	High computational performance (8.5 min/day) and scalability to NWS operational CPU processor counts needed to run 13 km and higher resolutions expected by 2020.
4	Extensible, well-documented software that is performance portable.
5	Execution and stability at high horizontal resolution (3 km or less) with realistic physics and orography
6	Lack of excessive grid imprinting

A short description of the test procedures referencing the evaluation criteria in Table 1 is given below.

1. Bit reproducibility for restart under identical conditions.

The candidate dycores should be able to restart execution and produce bit-reproducible results on the same hardware, with the same processor layout (using the same executable with the same model configuration).

2. Solution realism for dry adiabatic flow and moist convection

These are tests that measure the model's ability to simulate important atmospheric dynamical phenomena, such as baroclinic and orographic waves, and simple moist convection. Results will be evaluated from HIWPP idealized tests for this item.

3. High computational performance and scalability

Appendix 1 contains the detailed protocol for performance and scalability testing. Two resolutions are used. The following are additional comments. Computational performance is measured by elapsed wall time for each dycore executed, without I/O and initialization and for a representative period. The smaller workload, representing current to near-future NWS domains, is used to measure performance, here defined as the number of cores needed to reach a forecast rate of 8.5 minute per day. A larger workload sized to representing domains that may be routine in ten years is used to measure scalability, defined as the efficiency with which performance increases with the number of processors (speedup divided by the increase in number of processors). Scalability tests will include processor counts exceeding 100,000, as well as in the 10,000 range, which is anticipated operationally for the global model in the 2018 time frame.

4. Extensible, well-documented software that is performance portable

In addition to benchmark results, the AVEC will compile data sheets for each candidate core that includes basic characteristics of the core (numerical formulation, discretization) and technical implementation details including software design (modularity, extensibility, readability, maintainability) and performance-portability, especially with respect to next-generation NOAA HPC architectures and system configurations (decomposition and parallelization strategy, communication patterns, supported programming models, etc.). The clarity of documentation, including that available in the peer-reviewed literature and in-line code comments, will be part of this evaluation.

5. Execution and stability at high horizontal resolution (3 km or less) with realistic physics and orography

As a first test at NGGPS “Life Span” (LS) capability, where LS is defined as a dycore meeting operational needs in the 2025-2030 timeframe, each dycore will be executed at 3 km/ 60 vertical levels on a real (unscaled) earth. Two different sets of operational GFS initial conditions will be used, and forecasts will be run out to 3 days with realistic moist physics (chosen by each modeling group) and high-resolution orography. These tests will be run as part of the HIWPP project

6. Lack of excessive grid Imprinting

Candidate models will be evaluated for level of grid imprinting for idealized atmospheric flows. Results of HIWPP idealized tests will be evaluated to address this criterion.

Overall Phase 1 testing results will be compiled by the NGGPS Project Management Team (consisting of the project manager, test manager and support staff) for presentation to the DTG for review. NGGPS will leverage results from the ongoing HIWPP dycore evaluation efforts where applicable. The DTG will complete a review of the Phase 1 testing data and provide an overall assessment of the testing data to NWS management to inform a down-selection of dycores prior to Phase 2 testing. It is likely that each modeling group will request time for

revisions to their dycore to correct problems that appear in the Phase 1 testing. How these requests will be managed and what criteria, if any, will be applied is TBD (by the DTG).

VI. Dycore Evaluation Criteria and Phase 2 Testing

Table 2 lists evaluation criteria for Phase 2 dycore testing.

Table 2. Phase 2 Testing Evaluation Criteria

Phase 2 Eval #	Evaluation Criteria
1	Plan for relaxing shallow atmosphere approximation (deep atmosphere dynamics)
2	Accurate conservation of mass, tracers, total energy, and entropy
3	Robust model solutions under a wide range of realistic atmospheric initial conditions using a common (GFS) physics package
4	Computational performance with GFS physics
5	Demonstration of variable resolution and/or nesting capabilities, including physically realistic simulations of convection in the high-resolution region
6	Stable, conservative long integrations with realistic climate statistics
7	Code adaptable to NEMS/ESMF
8	Detailed dycore documentation, including documentation of vertical grid, numerical filters, time-integration scheme and variable resolution and/or nesting capabilities
9	Evaluation of performance in cycle data assimilation
10	Implementation Plan (including costs)

A short description of the test procedures referencing the evaluation criteria in Table 2 is given below. All proposed test procedures will be approved by the DTG before testing begins.

All groups will submit code packages to run each set of Phase 2 tests to the test manager so that the results can be verified independently as deemed necessary. The results of these tests will be analyzed by the test manager and synthesized in a final report to the DTG.

1. Plan for relaxing the shallow atmosphere approximation (deep atmosphere dynamics)

The next-generation global forecast system will be required to support both tropospheric and space-weather requirements. Therefore, NCEP/EMC has requested that the NGGPS dycore have the ability to relax the shallow-atmosphere approximation that is currently used in all NOAA operational weather forecast models. This involves letting the distance to the center of the earth and gravitational acceleration (as well as the horizontal distance between model grid

points) be a function of the model level in the model formulation. Since this will require significant development work, and is only one of several features that will need to be added to the NGGPS dycore for whole-atmosphere modeling (WAM) applications, we at this stage simply require that each modeling group submit a plan for incorporating the deep atmosphere equation set, including a description of the development work that will need to be done and an estimate of the time and effort required.

2. Accurate conservation of mass, tracers total energy, and entropy

A variant of the baroclinic-wave test case used in Phase 1 testing with large-scale condensation (DCMIP case 4.2) and extra tracers run at 15 km resolution will be used to assess the conservation of certain derived quantities that have particular importance for weather and climate application. These quantities are:

- Tracer mass (with and without monotonicity constraint, including the ability to maintain a constant tracer mixing ratio)
- Dry mass (including the effect of condensation/precipitation)
- Entropy (including an evaluation of the magnitude of spurious cross-isentropic transport). Note that exact conservation of moist equivalent potential temperature requires that the parameterized physics be 'reversible', which may require some modification of the DCMIP 4.2 simple physics package. The procedures outlined in http://journals.ametsoc.org/doi/full/10.1175/1520-0442%282000%29013%3C3860%3ANUITSO%3E2.0.CO%3B2 and http://onlinelibrary.wiley.com/doi/10.1256/qj.06.10/pdf will be used to evaluate entropy conservation and the magnitude of spurious cross-isentropic transports.
- Total energy

3. Robust model solutions under a wide range of realistic atmospheric initial conditions using a common (GFS) physics package

Retrospective forecast tests with GFS physics will be performed over a wide range of atmospheric conditions. The NUOPC GFS physics API (under development by EMC, delivered June 2015) will be implemented in each of the models. Forecasts out to 10 days will be run once every 5 days for one calendar year, initialized from GFS analyses and run at the resolution of the current GFS (15 km with 64 levels). The vertical distribution and the model top should be as close as possible to what is currently used in the GFS. The intent of these tests will be to evaluate the robustness of the models over a variety of atmospheric conditions, including but not limited to strong hurricanes, sudden stratospheric warmings, and intense upper-level fronts with associated strong jet-stream wind speeds. The NCEP VSDB verification code will be run on the model forecasts to provide a baseline assessment of un-tuned forecast skill and to identify flow regimes that pose particular challenges to the dycores. The forecast output will also be examined for the signatures of grid imprinting at the 8 vertices of the cubed sphere grid and the 12 pentagons on the icosahedral grid.

4. Computational performance with GFS physics (to be performed by the AVEC)

Using the time step and all other model settings used for #3, above, the models will be run at an *effective* resolution of 15 km and scaled up to the number of cores needed for a simulation rate of at 8.5 minutes per day (contingent upon securing additional commitments for HPC resources.) The effective resolution of each dycore is to be determined using kinetic energy spectra, and may not correspond to a nominal 15 km grid resolution. This will be done to ensure that the effects of numerical filters are accounted for when assessing performance. Three nominal horizontal resolutions around the nominal resolution used for the retrospective runs under Evaluation Criterion #3 will be assessed to evaluate/compare the (relative) effective resolution. For example, if the retrospective runs are at 13km, 15km and 11km will be tested, and performance reported as a function of resolution for each model. Since the NWS envisions using the same dynamical core for kilometer scale, convection permitting short-term forecasts (where non-hydrostatic dynamics are needed), and 10-100 km scale climate forecasts (where non-hydrostatic effects are negligible), dycores that have a hydrostatic run-time option will be allowed to run these tests in hydrostatic and non-hydrostatic mode. The results of these tests will reveal the impact of configurations that include realistic physics on the performance and scalability of the dynamical cores. As necessary and to the extent possible, AVEC will use performance profiling other instrumentation to isolate and factor out the performance and scaling of the GFS physics package itself. Appendix 2 contains the detailed protocol for performance and scalability testing. Three resolutions are used.

5. Demonstration of variable resolution and/or nesting capabilities, including physically realistic simulations of convection in the high-resolution region

Although NCEP/EMC has not yet defined requirements for nesting and/or variable resolution for the next generation global prediction system, it is anticipate that some capability will be required, especially for hurricane prediction. The purpose is to demonstrate a baseline capability to provide enhanced resolution over certain regions.

Approximately a 5:1 variation in horizontal resolution will be tested. Individual groups can configure as they choose, using either a variable-resolution mesh or a static nest. Identical physics packages will be used to isolate the effects of the dynamical core. The performance benefit of using locally enhanced resolution will be evaluated for each model by comparing the time to solution with a corresponding run using uniformly high resolution. For tests in which the high resolution region is not convection permitting, GFS physics (with parameterized convection) can be used in the tests involving real data initial conditions. For real-data tests in which the high resolution region is convection permitting, some development work may be needed if simply turning off the GFS deep convection scheme does not produce realistic results. The details of the tests to be performed will be determined by the test manager in collaboration with the modeling teams.

6. Stable, conservative long integrations with realistic climate statistics

Long-term idealized climate integrations at low resolution (50-100km) with GFS physics and specified sea-surface temperatures (i.e. 'AMIP' integrations) will be run. The period of integration will be chosen to include an El-Nino cycle. Long term climate statistics will be computed and compared to a reference simulation with the operational GFS model and CFS-R reanalysis. The degree to which the computational grid is reflected in climate statistics will be evaluated. Conservation of dry air mass, tracer mass and other important quantities will also be evaluated.

7. Code adaptable to NEMS/ESMF

An evaluation of Earth System Modeling Framework (ESMF) and National Unified Operational Prediction Capability (NUOPC) Layer compliance will be performed. The purpose is to demonstrate the ability to operate as an ESMF gridded component and comply with NUOPC standards. The assessment will include the presence of initialize, run and finalize methods, representation of import and export fields and time quantities as ESMF data structures at the component interface, and availability of a set services method for component registration with the framework. The assessment of NUOPC Layer compatibility will include the ability to successfully run the component within the NUOPC compliance checking tools.

8. Detailed dycore documentation

In order to understand the differences between the candidate dycores, each group should provide detailed documentation, including

a) Identification and documentation of numerical filters and fixers.
b) The methods used to couple the parameterized physics and dynamics.
c) Vertical grid and vertical transport schemes.
d) The time-integration scheme and horizontal transport schemes.
e) Methods used to ensure the accurate representation of pressure-gradient forces around steep orography. An idealized test that measures the degree to which a resting state is maintained in the presence of steep orography will be required in conjunction with this.
f) Strategies used for nesting and/or variable-resolution mesh generation.

9) Evaluation of performance in cycled data assimilation tests

Each dycore will be integrated with the NCEP GSI/EnKF ensemble-variational data assimilation system and cycled data assimilation tests will be run. The integration will be done by ESRL and NCEP, with the modelling groups providing software to interpolate from their model native grids to the latitude/longitude grids required by the GSI. The models will be configured to run with GFS physics as in item #3, but will be run at lower resolution due to computational constraints. The intent of these tests is to uncover unforeseen issues that can arise when models are run in

a cycled data assimilation system that might not be evident when they are ‘cold-started’ from another assimilation system.

10) Implementation plan

A plan for implementing each dycore into NCEP operations will be developed collaboratively by EMC and the modelling groups.

Phase 2 testing will be a more extensive evaluation of the down-selected dycore(s) from Phase 1 testing. The DTG will meet to refine Phase 2 evaluation criteria and will provide additional guidance to the dycore candidate teams on the criteria and methods for evaluation. Upon completion of the Phase 2 evaluation, results will be compiled by the NGGPS Project Management Team for presentation to the DTG. The DTG will conduct a review of the results and will generate a final assessment for NWS management. Results from the HIWPP testing will be used in Phase 2 evaluations where applicable. Phase 2 evaluation criterion #3 requires the use of a common GFS physics package being developed at EMC (completion is necessary by June 2015 to maintain the planned schedule for Phase 2 testing).

If the dycore design precludes, or requires excessive revisions to meet, any of the listed evaluation criteria characteristics, it could be removed from contention. Approximate resources (person-months) necessary to meet the evaluation criteria with the current/in operation candidate dycore must be estimated by the sponsor of the dynamic core candidate in consult with the NGGPS Project Team; these costs should be added to the total project effort for implementation.

VII. Evaluation

The computational performance testing will be performed by the AVEC. The NGGPS Project Management Team will coordinate the completion of the remaining test items, and synthesize the results in a report. Evaluation of all test results will be performed by the DTG.

VIII. Further Testing

Dycores will require further testing after Phase 2 is completed. This testing may include accuracy with operational components (e.g. any future upgrade to GFS physics, data assimilation), opportunities for accuracy tuning and further evaluation of computational performance. Emphasis will be on testing under the conditions in which the chosen dycore will eventually operate. Details are TBD.

Appendix 1 NGGPS Phase 1 Test Plan for Computational Performance

Advanced Computing Evaluation Committee

Chair: John Michalakes, NOAA (IMSG)
Co-chair: Mark Govett, NOAA/ESRL
Rusty Benson, NOAA/GFDL
Tom Black, NOAA/EMC
Alex Reinecke, NRL
Bill Skamarock, NCAR

I. Background and Purpose

The Advanced Computing Evaluation Committee (AVEC) was formed in August, 2014 to provide Phase 1 technical evaluation of HPC suitability and readiness of NGGPS candidate models to meet global operational forecast needs at NWS through 2025-30. This document describes the Phase 1 test plan for benchmarking and evaluation of computational performance, scalability, and HPC software design and reporting back to the NGGPS program in spring 2015. This test plan provides details of the benchmarking methodology, cases, model configurations, computational resource requirements, schedule, and results to be reported. The AVEC will leverage related computational performance testing efforts from ongoing HIWPP activities where applicable.

II. Benchmark Cases and Model Configurations

Two sets of benchmarks will be run: performance and scalability. The performance benchmark will measure speed of each candidate model running a near-future workload representing the cost of non-hydrostatic dynamics, including advection, running operationally beginning in 2015.

The scalability benchmark will measure how efficiently each candidate model is able to employ additional processors to run significantly more challenging workloads representing the cost of high-resolution non-hydrostatic dynamics and advection expected to be routine within 10 years.

The benchmarks will be conducted using the idealized baroclinic wave case with monotonically constrained scalar tracer advection, similar to the HIWPP configurations but with the following additional features:

1. The case will include ten extra 3D tracer fields initialized to a checkerboard pattern on the sphere to ensure that the cost of the monotonic constraint is represented in the benchmark workload. The detailed algorithm for initializing the tracers will be the subject of further discussion and agreement by the modeling groups.

2. Two horizontal resolutions (nominally 13 km and 3 km) on the full sphere will be benchmarked using 128 vertical levels. The resolution shall be as close as possible to target resolution.
3. Each group should choose a time step that is their best estimate of what they would use for a real-data forecasting case at each resolution. Rescaling of timing results may be done after the fact if the time step used when actual 3 km real data cases have been run deviates from best-guess time step used during for the Phase 1 benchmarks.
4. For verification, each group will provide a reference solution at each of the resolutions. The benchmark solutions will be evaluated for correctness by calculating differences with these reference solutions.
5. Duration of integrations (subject to computational resource availability): 30 minutes for the high-resolution case and 2 hours for the low resolution case.

Each candidate model's configurations – resolution, number of points, number of levels, and time step – for the two benchmarks has been reviewed and agreed upon by the other modeling groups. The configurations are listed in Table A2-1.

III. Benchmark Readiness

Each team will provide files, data, and scripts sufficient for benchmarkers to compile, run, and verify their model's test cases in rapid fashion during the benchmark period. AVEC will provide instructions to model teams on how to prepare their codes and data sets for benchmarking and evaluation and will work with the teams to conduct pre-benchmarking tests on smaller numbers of processors to ensure the full benchmark testing goes smoothly and within the allotted machine access times.

Model teams will generally use their own HPC resources for development and testing with smaller workloads, but non-dedicated access to larger partitions and time allocations on large the benchmark systems is also planned (under discussion with HPC centers).

IV. Final Benchmark Methodology

Benchmarks will be conducted in at least two sessions of dedicated access to a large system at one of the centers listed under Section VII Computational Resources below.

Performance: For the 13 km resolution performance benchmarks, each model will be run starting on about 1000 cores and then over successively larger numbers of processors until it achieves an integration rate for dynamics and advection required in the full-physics NGGPS to run at the operationally-required 8.5 minutes per day. The starting, ending and incremental numbers of processors will be determined during benchmark readiness phase of this work plan. These may differ from model to model to accommodate different parallelization and other implementation details.

Table A2-1. Model-specific Benchmark Configurations

	NH-GFS (Baseline) *	FV-3	MPAS	NIM	NMMB-UJ	NEPTUNE
Resolution	13 km (TL1534)	13km (C768)*	12km *	13.4 *	13 km	12.5 km *
Grid Points	3072x1536 (unreduced) 3,126,128 (reduced)	6x768x768 3,538,944	4,096,002 **	3,317,762	6x768x768 3,538,944 *	3,840,000 **
Vertical Layers *	128	127 **	127 ***	128	128	128 ***
Time Step	TBD	600s (slow phys) 150s (vertical, fast phys) 150/11 (horiz. acoustic)	72 s (RK3 dynamics) 12 s (acoustic) 72 s (RK3 scalar transport)	72 s	24 s **	60 s (slow RK3 dyn.) 10 s (fast dyn.) ****
Resolution	3 km (TL6718)	3.25 km (C3072) *	3km	3.3 km **	3 km	3.13 km *
Grid Points	13440x6720 (unred.) 59,609,088 (reduced) **	6x3072x3072 56,623,104	65,536,002	53,084,162	6x3072x3072 56,623,104 *	61,440,000 **
Vertical Layers *	128	127 **	127 ***	128	128	128
Time Step	TBD	150 s (slow phys) 37.5 s (vertical, fast phys) 37.5/11 s (horiz. acoustic)	18 s (RK3 dynamics) 3 s (acoustic) 18 s (RK3 scalar transport)	18 s	6 s **	15 s (slow RK3 dyn.) 2.5 s (fast dyn.) ***
Notes * Unless noted, layers refers to the number of layers, not the number of interfaces between layers + top + bottom	* Baseline configuration is tentative, pending test evaluation. ** Rough estimate for reduced Gaussian grid based on reduction factor (0.66) of 13 km grid. This will likely be revised after further testing of accuracy of spectral transform at TL6718.	* True resolution is average over equator and/or from south to north pole. For 13km, max cell size (edge of finite volume): 14.44 km, min: 10.21 km, global avg: 12.05 km. For 3.25 km, divide by 4. ** Favorable OpenMP Performance	* Resolution refers to mean cell-center spacing on the mesh ** Subdivision of 60 km mesh by factor of 5. *** Following the FV3 configuration, we will use 127 levels where density, theta and horizontal momentum are defined (on our Lorenz-grid vertical discretization) and 128 levels for w (that includes both the lower boundary and the model top "lid").	* Generated by 6 bisections followed by 2 trisections. Distances between neighbors: 13.367 average, 12.245 min., 14.397 max.. Maximum ratio of neighboring grid point distances: 1.17577 ** Generated by 8 bisections followed by 2 trisections. Distances between neighbors: 3.3417 average, 3.060 min., 3.601 max.. Maximum ratio of neighboring grid point distances: 1.1765.	* B-grid mass points ** For fast modes and advection of basic model variables. Time step for tracers is longer by 2x.	* Average nodal spacing per element. For 4th-order polynomials: ~12.5 km horizontal resolution will use 200 elements per edge of the cube sphere (grid can use 240,000 cores); ~3.13 km horizontal resolution will use 800 elements per edge (grid that use up to 3,840,000 cores). ** Horizontal grid points is six faces of cube times number of elements per face times polynomial order squared. *** Estimates are for split-explicit. May also use 3d- or 1d-imex method, with ab3/ai2 time integrator for expl./impl. step.

V. Scalability

For the 3km resolution scalability benchmarks, each model will be run starting on a minimum number of processors and then over successively larger numbers of processors until either performance has stopped increasing or the maximum number of processors has been reached. Both raw integration rate and scaling efficiency will be reported. Scaling efficiency is defined as:

$$E = (T_{np\_base} / T_{np\_tested}) \ / \ (np\_tested / np\_base)$$

where *T* is elapsed time (compute-only), *np_base* is the baseline (starting) number of processors and *np_tested* is the number of processors used in a given run. Ideally, E will be one.

As above, the incremental numbers of processors will be determined during the benchmark readiness phase, and may differ from model to model. The starting number of processors will be the maximum over all models of the minimum number of nodes the model fits in memory running the 3 km workload.

In addition to computational scaling, the memory scaling of the models will also be measured by instrumenting the models with the UNIX getrusage() library routine or similar.

For both sets of benchmark, there will be three replications of each benchmark.

VI. Reporting

The final Phase 1 Benchmarking report will provide data along with performance and scalability analysis that supports ranking of candidate model results and subsequent decision making by the NGGPS program, the DTG, and NWS management.

For both lower-resolution performance benchmarks and the higher-resolution scalability benchmarks, the raw timings (wall clock seconds average time step) and simulation speed (wall clock seconds per simulation interval) from each benchmark run will be provided in tabular form and plotted graphically. Simulation speed will be based on the time step used by the candidate models in the performance benchmark runs, but simulation speeds may be scaled upwards or downwards to allow for adjustment of the time step based on subsequent real data tests.

In addition to benchmark results, the AVEC will compile data sheets for each candidate core that includes basic characteristics of the core (numerical formulation, discretization) and technical implementation details including software design (modularity, extensibility, readability, maintainability) and performance-portability, especially with respect to next-generation NOAA HPC architectures and system configurations (decomposition and parallelization strategy, communication patterns, supported programming models, etc.).

VII. Computational Resources

Benchmarks will be conducted on a large homogeneous partition of a supercomputing system provisioned with on the order of 100-thousand conventional Intel Xeon processor cores (Sandy Bridge, Ivy Bridge, or Haswell, but not mixed). Any compiler, library, or other requirements shall be specified well enough in advance to ensure their availability on the benchmark system. Discussions are underway for use of one or more of the following supercomputing systems.

- NSF: Stampede. Texas Advanced Computing Center (TACC) at U. Texas at Austin
    - 102,400 cores over 6,400 dual Xeon E5-2680 (Sandy Bridge) nodes (16 cores per node), each with 32 MB
    - FDR InfiniBand 2-level fat tree interconnect
    - https://www.tacc.utexas.edu/user-services/user-guides/stampede-user-guide

- DOE: Edison. National Energy Research Scientific Computing Center (NERSC) at Berkeley National Laboratory.
    - 133,824 cores over 5,576 dual Xeon Ivy Bridge nodes (24 cores per node)
    - Cray Aries with Dragonfly topology
    - https://www.nersc.gov/users/computational-systems/edison/configuration
- NASA: Pleiades. NASA/Ames Research Center
    - 108,000 cores over 5,400 dual Xeon Ivy Bridge nodes (20 cores per node)
    - Possibility of ~100,000 cores of Xeon Haswell by benchmarking time
    - Dual plane 10D hypercube with InfiniBand interconnect
    - "Dedicated access" to Pleiades will mean to an uncontended section of the hypercube but not exclusive access to whole machine
    - http://www.nas.nasa.gov/hecc/resources/pleiades.html

VIII. Schedule

- October 8, 2014
    - Computational centers contacted and initial approvals for resource availability
- November 8, 2014
    - AVEC completes instructions for benchmark codes and data and provides to Model Teams
- December 12, 2014
    - Model groups provide initial codes and data sets
    - Computational resources finalized and available for benchmark readiness activity
    - Model groups and AVEC test and prepare benchmark codes and datasets
- February 15, 2015
    - Final suite of benchmark codes ready
- March-April, 2015
    - Two benchmarking sessions conducted on dedicated HPC resources
    - Benchmarks completed
- April 30, 2015
    - Final report

Acknowledgements

Nicholas Wright, NERSC. Bill Barth and Tommy Minyard, TACC. William Thigpen, Cathy Schulbach, and Piyush Mehrotra at NASA/AMES.

Appendix 2 NGGPS Phase 2 Benchmarking and Software Evaluation Test Plan

Advanced Computing Evaluation Committee

Chair: John Michalakes, NOAA (IMSG)
Rusty Benson, NOAA/GFDL
Michael Duda, NCAR
Thomas Henderson, NOAA/ESRL
Mark Govett, NOAA/ESRL

Created: Nov. 16, 2015 John Michalakes
Rev. Dec. 15, 2015 draft distributed to AVEC-II
Rev. Jan. 22, 2016 draft agreed to by modeling group reps. to AVEC-II

Introduction

The Advanced Computing Evaluation Committee (AVEC) was formed in August, 2014 to provide Phase 1 and Phase 2 technical evaluation of HPC suitability and readiness of NGGPS candidate models to meet global operational forecast requirements at NWS through 2025-30. This document describes the Phase-2 test plan for benchmarking and evaluation of computational performance, and HPC readiness and reporting back to the NGGPS program in spring 2016 to inform a decision on the modeling system to proceed to Phase 3 testing. This test plan describes the benchmarking methodology, cases, model configurations, computational resource requirements, detailed test instructions, schedule, and results to be reported.

Summary of AVEC Phase 2 Evaluations

The dycore testing criteria to be evaluated by AVEC are items 4, 5, and 7 (in bold) from the following table:[2]

Phase 2 Eval #	Evaluation Criteria
1	Plan for relaxing shallow atmosphere approximation (deep atmosphere dynamics)
2	Accurate conservation of mass, tracers, entropy and energy
3	Robust model solutions under a wide range of realistic atmospheric initial conditions using a common (GFS) physics package

[2] From minutes of the NGGPS Dycore Testing Group (DTG) meeting in 11 December, 2015

Phase 2 Eval #	Evaluation Criteria
4	**Computational performance with GFS physics**
5	**Demonstration of variable resolution and/or nesting capabilities, including physically realistic simulations of convection in the high-resolution region**
6	Stable, conservative long integrations with realistic climate statistics
7	**Code adaptable to NEMS/ESMF**
8	Detailed dycore documentation, including documentation of vertical grid, numerical filters, time-integration scheme and variable resolution and/or nesting capabilities
9	Evaluation of performance in cycle data assimilation
10	Implementation Plan (including costs)

The remainder of this document describes the three areas of AVEC testing in detail.

Eval. Criterion #4: Computational Performance with GFS Physics

The tests will be conducted in the form of two series of benchmarks, similar to what was done during Phase 1 testing. The purpose of the first computational benchmarks will be to measure model performance with representative physics to determine the computational resources required to meet an operational speed requirement of 8.5 minutes per forecast day. The second series will measure the effect on performance of varying the number of tracers being advected. The first series is higher priority than the second. Details of the Criteria #4 testing are as follows:

- The models will be benchmarked using the same physics and physics configurations as the retrospective runs for NGGPS Evaluation Criterion #3 (see NGGPS Test Plan) but will be instrumented to provide separate timing data for the dycore, the physics, and the physics interface from each parallel process.
- *Benchmarking series one (first priority):* The series of performance benchmarks for each code will involve up to **45 runs** of each code with advection of only prognostic fields, including fields required by GFS physics, with varying resolution and core counts:
    - **Three** nominal horizontal resolutions will be benchmarked around the nominal resolution used for the retrospective runs under Evaluation Criterion #3. For example, if the retrospective runs are at 13km, we will test 15km and 11km and report performance as a function of resolution for each model.
    - The models will be run on up to **five** different processor core counts that straddle ±20 percent of the 8.5 minute per day speed threshold and we will report performance as a function of number of processor cores for each model.

- Each configuration/core-count will undergo **three** replications to assess and minimize the effect of run-to-run timing variation on the test system, access to which will not be dedicated. That is, other users will be running on the NERSC system while the NGGPS benchmarks are being conducted.[3]
- As a matter of prioritization, the progression of testing will be to perform the series of 15 runs for each model at the nominal resolution. Results will be evaluated and then, time and resources permitting, the series of runs the lower resolution will be conducted; then the series of runs at the higher resolution.
- Only performance, not scalability will be measured. As in Phase-I testing, the benchmarks are compute-only; initialization and I/O time will not be reported.

- *Benchmarking series two (second priority):* each model will be run **6 times** to measure computational cost as a function of increasing numbers of tracers in the code. AVEC will report the slope and shape of this function for each model.
  - **One** horizontal resolution, same as retrospective runs, will be benchmarked.
  - **One** processor core count will be benchmarked, the one that gave performance closest to 8.5 minute per day in benchmarking series one, above, with an additional 15 artificial tracers initialized to a checkerboard pattern
  - The models will be run on the same processor core count with 30 artificial tracers (time permitting) to determine the curvature of the cost per number of tracers function
  - **Three** replications of each run will be conducted to assess and account for run-to-run variation.
- As with the Phase 1 benchmarks, AVEC will rely on DTG and the individual modeling groups to provide the codes, configurations, input data sets, reference output data sets and verification methodology for each of the test cases described above, including:
  - Forecast start and end times large enough to include at least a one day interval to be timed after an appropriate initial spin-up interval
  - One or more nominal horizontal resolutions to be tested. The purpose of testing multiple model resolutions will be to provide a range of tests for a posteriori consideration of effective resolution by the DTG.
  - Number and spacing of vertical levels.
  - Physics settings, including calling intervals with respect to simulation time.

[3] The performance benchmarks conducted during Phase-1 testing were on a dedicated system; however, multiple runs from the AVEC suite were conducted concurrently on different sets of nodes of the benchmark system so that there was a possibility of variation from contention. In actual, measurements, however, this was generally negligible – only a very small number of timings from the many runs were discarded as outliers.

- Number of tracer fields and required advection settings (i.e., limiters).
- These will be submitted to the other modeling group and the DTG at large for approval.
- Reference output will be submitted to the other team and to the Dycore Test Manager (Jeff Whitaker) for meteorological verification and validation. The reference output is a dycore's output of record for subsequent evaluation of the model's meteorological performance on the test case. Any change to a models code or configuration must be approved in advance by the Dycore Test Manager, who may require resubmission and V&V of the reference output.

- The benchmarks will be conducted on HPC systems with current-generation conventional multicore processors that provide threading, hyperthreading, and fine-grained parallelism in the form of vector instructions.
  - Setup, testing and benchmarking will be conducted on the NERSC Cori Phase-1 system[4], a Cray XC-40 with dual 16-core 2.3 GHz Intel Xeon (Haswell) processors (32 cores per node). Cori is similar to NOAA's new Cray systems Luna and Surge.
  - Dycore test groups are allowed and expected to perform as much optimization of their codes on this architecture as possible prior to submission of the reference output for the benchmarks. As noted above, any subsequent changes will require resubmission and validation of reference output.
- Further, with regard to optimization of the models:
  - The GFS physics codes, settings and configurations must be identical between the two models, and no changes may be made to any code below the NUOPC Physics interface. Different ways the physics interface might be called from a model (e.g. chunking or calling over different threads) are allowed as long as there is no change to the NUOPC interface code or underlying physics itself.
  - Optimizations that induce bit-differences in model output between successive identical runs of the model are not allowed (AVEC discussion and agreement 01-21-2016)
  - The list of compiler settings used by one group will be made available to the other group. Within a given model code, application of different compiler optimizations to different source files (including GFS physics) is allowed as long as these settings are made available to the other group.

Evaluation criteria for assessing computational performance:

[4] https://www.nersc.gov/users/computational-systems/cori/cori-phase-i/

- Number of cores required to meet an operational speed requirement of 8.5 minutes per forecast day, at each given resolution specified in the configurations provided by modeling groups and DTG.
- Efficiency of physics interface, including any copying, transposition, or other reorganization.
- Load imbalances or other computational inefficiencies that result from mapping of physics work to processors because of layout, decomposition, or other properties inherent to a particular dycore. This includes both effects on the performance of the physics component of the model and inefficiencies manifested within the dycore component as a result of being coupled to physics.
- Thread efficiency (including hyperthreading), if applicable.
- Utilization of fine-grained parallel resources on the test hardware: vector efficiency.

Schedule and Coordination of Criterion #4 Testing and Evaluation

Schedule and set of milestones within the initial preparation and benchmark readiness phases:

By January 22, 2016

AVEC provides to Model Groups:

1. Instructions and schedule for preparing and submitting benchmark codes (this document)
2. Accounts and allocations on development system (NERSC Cori) for use in preparing codes for handoff to the AVEC test coordinator
3. Meeting schedule for discussing progress (TBD)

By February 4, 2016

Model groups provide to the AVEC test coordinator:

1. All source code, makefiles and instructions necessary for the test coordinator to independently build, run and verify the code on the test system using the benchmark data set,
   a. README file with build instructions and other information (see below).
   b. Input data, configuration files, and reference output generated on model group's HPC resources for benchmark case (web links, anonymous FTP, tar files)
2. Source code for verification program(s) for benchmark reference output, instructions for running the program, and criteria for verification.

By Feburary 19, 2016

1. All codes have been tested and verified on benchmark HPC system

By March 4, 2016 (may slip, depending on obtaining final codes and data)

1. Final validated versions of code provided by modeling groups; benchmarking commences
2. First set of performance results collected, compiled and prepared into a draft report for review and approval by AVEC
3. Plan for and conduct possible second round of testing

By March 31, 2016

4. Final set of performance results collected, compiled, and prepared into a report for review and approval by AVC
5. Delivery of final report to NGGPS program

README file contents

The following information should be provided in a text or Word file.

1. Specifies all compiler and version requirements
2. Enumerates all package/libraries and version requirements
3. Enumerates any additional environment requirements
4. Describes the data and configuration provided (item #2 above). The configuration descriptions should include instructions for controlling run-length, output frequency, time step, and other settings that might be varied during the process of readying the final suite of benchmarks.
5. Lists the sequence of commands, including any which set the runtime environment, required to run the executable for a given number of processors, threads, etc. Any case-specific instructions should be listed as well.
6. Lists, to the extent possible, estimated run times on known processor counts for the cases provided
7. Describes the data files that will be generated as the model runs – what these contain, their formats, file naming conventions, estimated sizes (if large), and which of these need to be saved and returned to the modeling groups for additional verification/validation/analysis
8. Describes timing information that is output as the model runs, its meaning, and to which files. The location of the main time loop in the code around which additional timers to measure the compute-only performance of the model (excluding I/O and initialization) may be inserted
9. Describes the set of commands to verify model output with respect to reference output, and describe the resulting output statistics and the criteria which should be used to assess correctness of the run
10. Any additional information

Eval. Criterion #5: Demonstration of variable resolution/nesting

Selective grid refinement or nesting over a region of interest is regarded as an expediency, necessitated by the prohibitive cost of running with uniform high resolution over a global domain. The AVEC will evaluate the efficiency with which a model's refinement/nesting schemes mitigate the cost of high resolution and also the cost and complexity of setting up refined meshes or nesting configurations with the respective models.

With respect to run-time cost, the AVEC will measure and evaluate the computational cost of generating the best refined/nested solution possible relative to the cost of running uniformly over the global domain at the targeted resolution. This evaluation will be done under the supervision and direction of AVEC but the timing information will be collected during the runs conducted by the groups themselves, with the possibility of subsequent replication by AVEC to verify results if necessary.

With respect to setup cost and complexity, the AVEC will include in its report a description of the steps to set up nesting or refinement in a candidate model, the methods and tools involved in this setup, their cost to run, and the pre- and post-processing of data necessary to initialize and analyze the data from a refined/nested run. The AVEC will review existing documentation and tutorial information available for the respective candidate models and perform a hands-on setup, run, and post for the model.

Eval. Criterion #6: Adaptability to NEMS/ESMF

AVEC will evaluate candidate dycores' adaptability and state of readiness for ESMF/NEMS of the candidate dycores using self reports by the modeling groups verified where necessary using ESMF/NEMS compliance checking tools (?).

Additional information

Needed