



Dynamical Core Evaluation Test Report for NOAA's Next Generation Global Prediction System (NGGPS)

Submitted
Ming Ji
Dynamical core Test Group Chair

Accepted by
Frederick Toepfer
NGGPS Program Manager

September 2016

Prepared by the NGGPS Dynamical core Test Group (DTG):

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

External Consultants:

- Dr. Robert Gall, University of Miami
- Dr. Richard Rood, University of Michigan
- Dr. John Thuburn, University of Exeter

Candidate Dycore Representatives:

- Dr. Melinda Peng, Superintendent (Acting), Naval Research Laboratory (NRL) Monterey
- Dr. Venkatachalam Ramaswamy, Director, NOAA Geophysical Fluid Dynamics Laboratory (GFDL)
- Kevin Kelleher, Director, Global Systems Division, NOAA Earth System Research Laboratory (ESRL)
- Dr. Hendrik Tolman*, Director, NOAA Environmental Modeling Center (EMC)

NGGPS Program Manager: Frederick Toepfer and Timothy Schneider (Acting)/Dr. Ivanka Stajner (Deputy)

Ex Officio Members:

- **Test Manager:** Dr. Jeffrey Whitaker (NOAA, ESRL)
- **Advanced Computing Evaluation Committee (AVEC) Test Manager:** John Michalakes (UCAR)

Technical Representatives:

- Dr. Shian-Jiann Lin (NOAA, GFDL)
- Dr. Vijay Tallapragada (NOAA, EMC)
- Dr. Stan Benjamin (NOAA, ESRL)
- Dr. James Doyle (Navy, NRL)

Technical Observer: Dr. Rohit Mathur, (Environmental Protection Agency (EPA))

NGGPS Staff:

- Steve Warren
- Sherrie Morris

* Now at NOAA NWS Office of Science and Technology Integration

Contents

Executive Summary.....	5
Chapter 1 Introduction	8
1.1 Background.....	8
1.2 The Dynamical core Test Group (DTG): An Evidence-based Decision Making Process.....	8
Chapter 2 Phase 2 Evaluation	14
2.1 Criterion 1: Plan for Relaxing Shallow-Atmosphere Approximation (Deep Atmosphere Dynamics).....	14
2.2 Criterion 2: Accurate Conservation of Mass, Tracers, Entropy and Energy	15
2.2.1 Overview	15
2.2.2 Test Setup	15
2.2.3 Diagnostics	16
2.2.4 Results.....	17
2.2.5 Conclusions	22
2.3 Criterion 3: Robust Model Solutions Under a Wide Range of Realistic Atmospheric Initial Conditions Using a Common (GFS) Physics Package	23
2.3.1 Overview	23
2.3.2 Test Setup	23
2.3.3 Effective Resolution	25
2.3.4 Global Precipitation Forecasts	26
2.3.5 Forecast Skill	27
2.3.6 Robustness	29
2.3.7 Conclusions	30
2.4 Criterion 4: Computational Performance with GFS Physics	31
2.5 Criterion 5: Demonstration of Variable-Resolution and/or Nesting Capabilities, including Supercell Tests and Physically Realistic Simulations of Convection in the High-Resolution Region....	31
2.5.1 Overview	31
2.5.2 Idealized Supercell Test.....	32
2.5.3 Idealized Tropical Cyclone Test.....	33
2.5.4 Variable-Resolution Tests	35
2.5.5 Conclusions	39
2.6 Criterion 6: Stable, Conservative Long Integrations with Realistic Climate Statistics	40
2.6.1 Overview	40
2.6.2 90-day Means.....	41
2.6.3 Conservation of Mass and Energy.....	45
2.6.4 Grid Imprinting.....	46

2.6.5 Conclusions	47
2.7 Criterion 7: Code Adaptable to NEMS/ESMF.....	48
2.7.1 Overview	48
2.7.2 Conclusion.....	50
2.8 Criterion 8: Detailed Dycore Documentation, including Documentation of Vertical Grid, Numerical Filters, Time-Integration Scheme and Variable-Resolution and/or Nesting Capabilities ...	50
2.9 Criterion 9: Evaluation of Performance in Cycled Data Assimilation	51
2.9.1 Overview	51
2.9.2 Fit of First-Guess Forecasts to Observations	52
2.9.3 Model-Space Verifications Relative to ECMWF Analyses	53
2.9.4 Conclusions	55
2.10 Criterion 10: Implementation Plan (Including Costs)	56
2.10.1 Overview	56
2.10.2 Implementation Tasks.....	58
2.10.3 Human and Computational Resource (Implementation Cost) Requirements.....	61
2.10.4 Conclusion.....	62
Chapter 3 Phase 2 Conclusions	62
Chapter 4 Next Steps	63
Appendix A: Director, NWS Approval Memorandum	A-1
Appendix B: Dynamical core Test Group Charter	B-1
Appendix C: SME Variable-Resolution	C-1
Appendix D: AVEC Phase 2 Report.....	D-1
Appendix E: Effective Resolution	E-1
Appendix F: Vertical Coordinate Analysis	F-1
Acronym List	1

Executive Summary

Motivation

In order to improve global forecasts for its numerical guidance products, the National Oceanic and Atmospheric Administration (NOAA) evaluated candidate non-hydrostatic dynamical cores (dycores) with a battery of tests to culminate in the selection of a new dycore for the Next Generation Global Prediction System (NGGPS). The evaluation considered applications ranging from weather and climate prediction, anticipation of simulating moist convection on the global scale, and efficient usability of present and future high performance computers. The current NOAA operational global forecast model is spectral and hydrostatic; therefore, it cannot explicitly simulate moist convection and will not be able to scale up to take advantage of peta- and exa-scale high performance computing (HPC) systems. The first round of tests (Phase 1) for solution accuracy and computational performance resulted in the selection of two dycores (out of five tested cores) to proceed to the final round of testing (Phase 2). This is a high level summary of the results of the Phase 2 tests – more details are included in the main body of the report.

Candidate dycores

The two candidate dycores evaluated in Phase 2 are listed below, with home institutions in parentheses.

- FV3 (Geophysical Fluid Dynamics Laboratory (GFDL)) – Cubed sphere grid, finite-volume discretization (non-hydrostatic version of the hydrostatic core described in (Lin 2004, Putman and Lin 2007)
- MPAS (National Center for Atmospheric Research (NCAR)) – Unstructured grid, finite-volume discretization with C-grid variable staggering (Skamarock et al., 2012)

More complete documentation can be found at

http://www.weather.gov/sti/stimodeling_nggps_dycoredocumentation

Test results

FV3 and MPAS were subjected to a battery of tests, including idealized simulations of specific phenomena (moist convection and tropical cyclones) and real-world forecast performance with NOAA operational physics and data assimilation. To isolate the performance of the dycores from differences in parameterized physics, the operational Global Forecast System (GFS) physics package was implemented in both models. Computational performance was assessed using GFS physics, with both models configured as they were in the forecast performance tests.

Idealized tests with simplified moist physics showed that both dycores conserve dry mass very accurately. Entropy and energy were also well conserved and deviations from exact conservation were consistent with limitations in the simple moist physics and the approximations inherent in calculating the conserved quantities. Both FV3 and MPAS were able to simulate the development of an idealized splitting supercell thunderstorm and the evolution of an idealized tropical cyclone. The size of the simulated FV3 thunderstorm updraft was slightly larger than the MPAS solution and the MPAS tropical cyclone did not display the expected pattern of concentric rising motion. However, the Dynamical core Test Group (DTG) does not regard any differences in these idealized tests as indicative of any fundamental limitation of either dycore with regards to convective-scale or tropical cyclone forecasting applications or the ability to conserve the expected dynamical invariants in long-term climate integrations. The idealized tests revealed that the impact of the computational grid can be seen in the numerical solutions of both dycores, and it is small and not likely to pose a problem in operational forecast applications.

A set of 10-day forecasts with grid resolutions matched to the current operational GFS were performed using operational GFS initial conditions. The skill of the FV3 forecasts was quite close to that of the operational GFS, even without tuning the GFS physics to the FV3 dycore. MPAS forecasts were significantly worse than either of the FV3 or GFS forecasts and at times exhibited non-physical small-scale noise in the vicinity of the extra-tropical jets. An assessment of the scales resolved by each model revealed that both MPAS and FV3 effectively resolve features at spatial scales nearly half that of the current operational GFS. No significant difference between the scales of features resolved by MPAS and FV3 were noted. With regard to computational performance, FV3 forecasts were able to achieve a specified time to solution using approximately three times fewer processor cores than MPAS and approximately 30% more processor cores than the current operational GFS. In cycled data assimilation experiments at reduced resolution, FV3 first guess forecasts fit the observations much better than MPAS and better than a comparably configured GFS. This suggests that FV3 should be able to outperform the operational GFS when run from its own analyses generated from the operational cycling data assimilation system.

Simulations with variable-resolution grids that telescope down from 13 km to 3 km over and near the continental U.S. were performed for a case of severe convection over the Central U.S. and a major hurricane off the East Coast. MPAS used a variable-resolution mesh, while FV3 used a combination of a stretched global grid and a nest. The FV3 nesting approach was significantly more computationally efficient than the MPAS variable-resolution mesh approach. Both models were run with GFS physics with the deep convection scheme disabled. Acknowledging the limitations of GFS physics at these scales, the DTG found that the simulated convection in both cases was generally similar in both models, although the scale of convective features was somewhat larger in the FV3 solution. This was due in part to the fact that the resolution of the FV3 grid in the high-resolution region was somewhat coarser than MPAS. Both dycores were determined to be suitable for variable-resolution global to convective-scale forecast applications. Significant development work would be needed to develop scale-aware physics packages that work well across scales.

90-day integrations were performed at reduced resolution, using GFS physics. Both dycores produced realistic solutions that were comparable to the GFS. The same grid-imprinting signal observed in the idealized tests was observed to occur, but was only discernible when looking at long-time averages of the vertical velocity and precipitation fields. The results of this test suggest that both models would be suitable for longer term climate forecast applications.

Subjective evaluations of the code and documentation suggest that either model could be adapted to work for whole atmosphere, space weather applications – although a significant amount of further development would be required for either dycore. Similarly, the amount of work required to integrate either dycore into the NOAA Environmental Modeling System (NEMS) framework would be similar. A subjective evaluation of the costs required (in terms of both manpower and computational resources) by the National Centers for Environmental Prediction (NCEP)/Environmental Modeling Center (EMC) found that the cost to implement FV3 as a replacement for the operational spectral dycore in the GFS would be significantly less than for MPAS because of the additional development work needed for MPAS to approach the forecast skill and computational performance of GFS.

Recommendation

FV3 performed much better than MPAS in real-world tests with operational GFS physics and performed at significantly less computational cost. MPAS did not exhibit any clear-cut offsetting advantages in

other aspects of the test suite. Therefore, DTG recommends that the National Weather Service adopt the FV3 atmospheric dynamical core in the Next Generation Global Prediction System¹.

References

Lin, S.-J., 2004: A Vertically Lagrangian Finite-Volume Dynamical Core for Global Models. *Mon. Wea. Rev.*, **132**, 2293-2307. doi: [http://dx.doi.org/10.1175/1520-0493\(2004\)132<2293:AVLFDC>2.0.CO;2](http://dx.doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2)

Putman and Lin, 2007: Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, **227(1)**, 55-78.

Skamarock, W, M. Duda, L. Fowler, S.-H Park and T. Ringler, 2012: A Multi-scale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tessellations and C-Grid Staggering. *Mon. Wea. Rev.*, **240**, 3090-3105, doi: <http://dx.doi.org/10.1175/MWR-D-11-00215.1>

¹Director, NWS approved the DTG recommendation on 26 July 2016 (Appendix A).

Chapter 1 Introduction

On July 26, 2016, the National Oceanic and Atmospheric Administration (NOAA) selected a non-hydrostatic dynamical core (dycore) to serve as the basis for its high-resolution Next Generation Global Prediction System (NGGPS): the Finite Volume Version 3 (FV3) dynamical core, developed by the NOAA Geophysical Fluid Dynamics Laboratory (GFDL). This document describes NOAA's evidence-based decision making process, provides a summary of the key findings and analyses, and includes an assessment of these results.

1.1 Background

As part of NOAA's Research to Operations (R2O) Initiative to expand and accelerate critical weather forecasting, the National Weather Service (NWS) is developing a state-of-the-art next generation global prediction system to be the foundation for the operating forecast guidance system for the next several decades. By upgrading the current operational Global Forecast System (GFS) to a unified, global coupled system within the NOAA Environmental Modeling System (NEMS) infrastructure, this new system will

- Extend forecast skill beyond 8 to 10 days
- Improve hurricane track and intensity forecast
- Support development of products for weeks 3 and 4 to extend weather forecasting to 30 days

Through NGGPS, model developers, including NOAA and other federal laboratories, the Navy, the university research community, and other partnership efforts, are accelerating research and development efforts to identify and refine weather prediction model components, and improve data assimilation and post-processing. The model will utilize NOAA's new high performance computing (HPC) capabilities and allow for a continuously evolving and improving system with the flexibility, and capability, to implement improvements in components as they are developed.

Additional and current information about NGGPS can be found at http://www.weather.gov/sti/stimodeling_nggps

1.2 The Dynamical core Test Group (DTG): An Evidence-based Decision Making Process

Within NGGPS it was decided to choose a dycore from the rich U.S. research community. NGGPS requires an atmospheric dycore that is non-hydrostatic, highly scalable and architecturally compatible with existing and projected HPC architecture. Therefore, a process was developed to select a dycore that provides an accurate representation of atmospheric motions, is cost-effective, can adapt to a changing computational environment, and is projected to support scientific and programmatic goals for the next decade. Six dycores from five institutions were viewed as potential candidates to be evaluated for the new system: Navy/Naval Research Laboratory (NRL), National Center for Atmospheric Research (NCAR), NOAA/Earth System Research Laboratory (ESRL), NOAA/GFDL; and NOAA/National Centers for Environmental Prediction (NCEP). The process began in August 2014, when modelers attended a workshop to discuss ideal dycore requirements/attributes for the NGGPS. Since then, many teleconferences, workshops and meetings have occurred to develop a test plan; and to identify, document and execute the criteria and tests best suited to evaluate what characteristics are predicted to be the most beneficial in the next dycore.

To provide impartial oversight and guidance to the process of selecting a new dycore, a Dynamical core Test Group (DTG) was formed, and a DTG charter was drafted in the spring of 2015 (Appendix B). The DTG, comprised of federal, academic, selected subject matter experts, and representatives from each of the candidate dycores (Figure 1.1), provided an objective and unbiased assessment of the tests and evaluation results. In order to maintain a professional, unbiased process, and confirm all data was reviewed and any issues fully addressed before being released, each participant signed a non-disclosure agreement, noting that no communication or release of results would occur until agreed upon by the group. The DTG, working in accord, developed and approved the Dycore Test Plan (http://www.weather.gov/sti/stimodeling_nggps_implementation_atmdynamics), and executed, reviewed and assessed the tests and results throughout Phases 1 and 2. These are summarized below; however the primary focus of this report is on Phase 2.

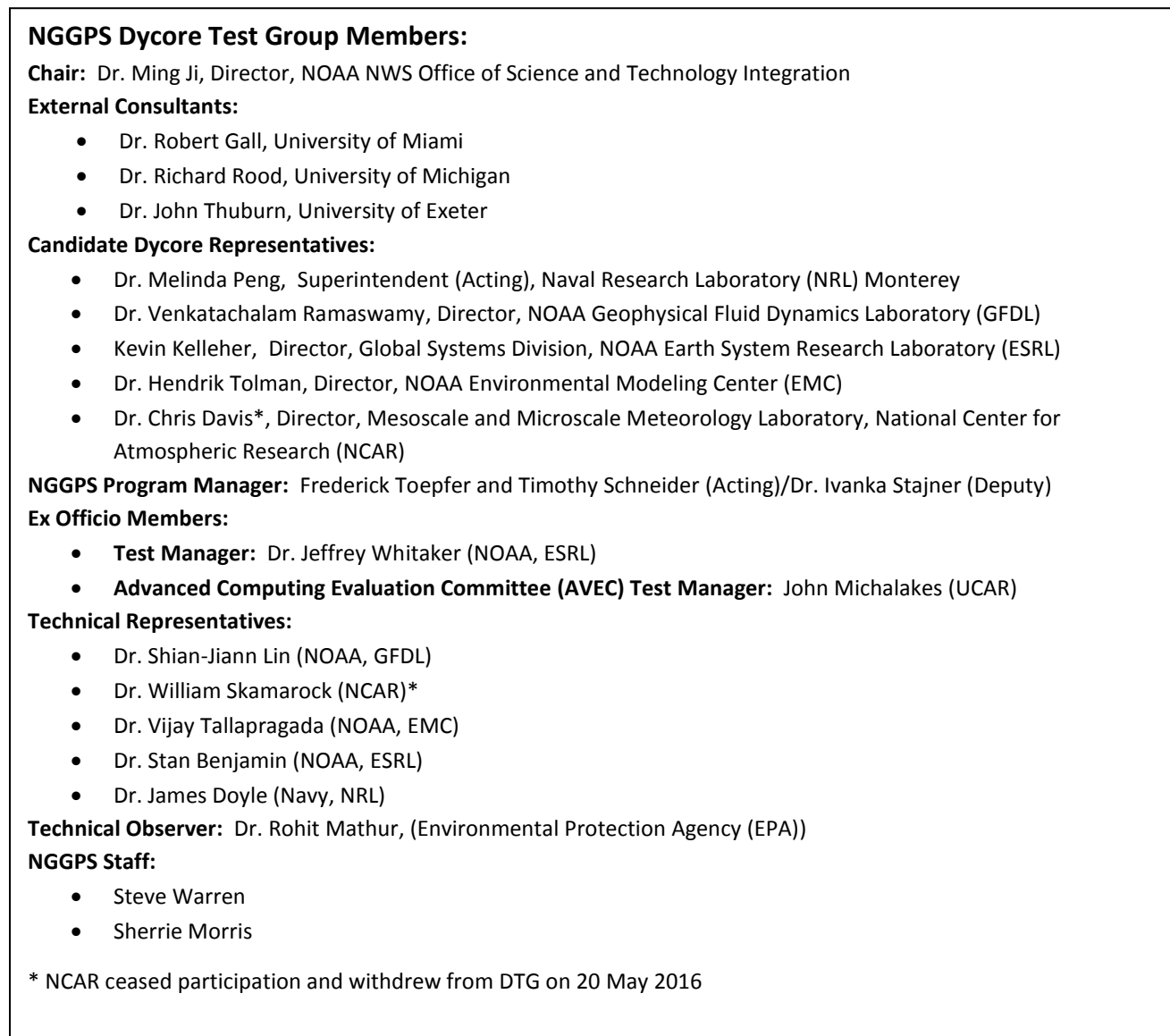


Figure 1.1: The composition of the NGGPS DTG.

A group of subject matter experts, convened as the Advanced Computing Evaluation Committee (AVEC), managed computational performance and scalability, benchmark tests, HPC suitability and readiness during Phase 1 and Phase 2, and provided technical evaluations of the results to the DTG. AVEC membership included:

Phase 1:

- John Michalakes², UCAR (Co-Chair)
- Mark Govett, NOAA, ESRL (Co-Chair)
- Rusty Benson, NOAA, GFDL
- Tom Black, NOAA, EMC
- Alex Reinecke, Navy, NRL
- Bill Skamarock, NCAR
- Henry Juang, NOAA, EMC

Phase 2:

- John Michalakes, UCAR (Chair)
- Mark Govett, NOAA, ESRL
- Rusty Benson, NOAA, GFDL
- Michael Duda, NCAR
- Mike Young, NOAA, NCEP

Michael Duda participated fully in AVEC Phase 2 discussions and activities, but ceased participation in the AVEC after 20 May 2016, when NCAR formally withdrew MPAS from consideration as a dycore for NNGPS and ceased participation in the DTG.

The initial Dycore Test Plan, including the AVEC Test Plan, was developed by the DTG in the summer of 2015 for the Phase 1 evaluation. During the September 2015 DTG face-to-face meeting, the Phase 2 dycore testing criteria were vetted, discussed, and finalized, and incorporated into the Dycore Test Plan, which was revised and approved by the DTG in January 2016.

As described in the Dycore Test Plan, the dycore evaluation process is separated into three phases of tests and assessments. Phase 1 was designed to evaluate the dycores for solution accuracy, and technical performance and scaling; and to determine if any of the dycores had distinctive desirable characteristics, and/or were mature enough to continue to Phase 2. The outcome of Phase 1 reduced the field from six to two cores for the next phase of testing. Phase 2 of the dycore testing was designed to evaluate HPC performance, suitability and readiness for transition into an operational system, and to select the next dycore. The DTG provided assessments of the results to NOAA (NWS) management who then made an overall business case decision on the selection of the next dycore. Phase 3 will address the path forward to integrate and implement the new dycore. See Figure 1.2 below for the dycore evaluation timeline.

² Now at UCAR.

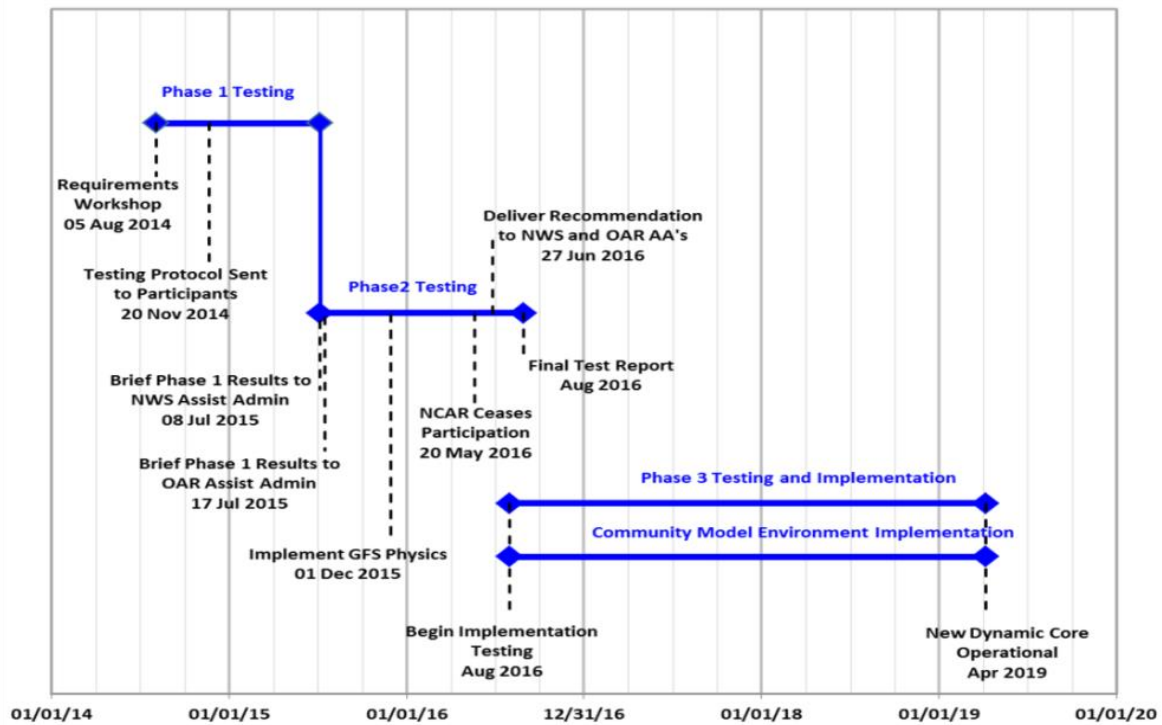


Figure 1.2: Dycore testing and implementation timeline.

During Phase 1, the DTG and the program office worked together to identify evaluation criteria and assessments to be conducted on six candidate models:

- Non-hydrostatic Global Spectral Model (GSM) - EMC³
- Finite Volume Model Version 3 (FV3) - GFDL – Cubed sphere grid, finite-volume discretization (non-hydrostatic version of the hydrostatic core described in Lin (2004) and Putman and Lin (2007))
- Model for Prediction Across Scales (MPAS) - NCAR – Unstructured grid, finite-volume discretization with C-grid variable staggering (Skamarock et al., 2012)
- Navy Environmental Prediction System Using the NUMA CorE (NEPTUNE) - NRL – Cubed sphere or icosahedral grid using a spectral element discretization with the Non-hydrostatic Unified Model of the Atmosphere (NUMA) core (Giraldo et al. 2014)
- Non-hydrostatic Icosahedral Model (NIM) - ESRL – unstaggered finite-volume A-grid implementation
- Global Non-hydrostatic Mesoscale Model (NMM & NMM-UJ) (NCEP/EMC) – Finite-difference cubed-sphere grid version of the B-grid lat/lon mesh core used for operational regional modeling (Janjic and Gall 2012).

For Phase 1, the DTG defined six specific evaluation criteria to assess the dycores, including criteria that required computational testing such as solution accuracy and computational performance and scaling, along with criteria that could be evaluated with a yes/no response. The evaluation criteria were not pass/fail, and a “no” answer to a “yes/no” evaluation was one factor to be considered along with remaining evaluation data in a final decision on model preference.

³ The Non-hydrostatic GSM was an initial candidate in Phase 1, but did not participate in tests.

Phase 1 Criteria #	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

Table 1.1: Phase 1 Dycore Evaluation Criteria.

The AVEC designed fair and objective benchmark methodologies and evaluation criteria pertaining to computational performance, scalability and suitability of model software for next-generation HPC architectures. Modeling groups were responsible for providing codes, data, verification criteria, and code-specific technical advice and assistance needed to complete the benchmarks and evaluation to the AVEC. The AVEC Phase 1 results can be found in an Addendum to the Phase 1 Report and at the NGGPS website http://www.weather.gov/sti/stimodeling_nggps_implementation_atmdynamics.

Results from idealized test cases and 3 km, three-day full physics simulations, including leveraging ongoing High Impact Weather Prediction Project (HIWPP) activities, were evaluated and attributes considered in determining which two dycores would proceed to Phase 2. FV3 and MPAS produced solutions that were generally of higher quality than the other models. An assessment of the Phase 1 test results also determined that going forward to Phase 2 with two dycores was a low technical risk and that no unique dycore quality would be lost. Data and reports from Phase 1 are available on the NGGPS website http://www.weather.gov/sti/stimodeling_nggps_implementation_atmdynamics.

Therefore at the conclusion of Phase 1 testing, the NGGPS Program Manager recommendation to the NWS Director was to proceed to Phase 2 testing on schedule with two dycores (FV3 and MPAS).

The Phase 2 selection evaluation criteria were determined and refined during the September 2015 face-to-face meeting, and are listed in Table 1.2 below. The focus for Phase 2 was testing under the conditions in which the chosen dycore will eventually operate.

Phase 2 Criteria #	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
4	Computational performance with GFS physics
5	Demonstration of variable-resolution and/or nesting capabilities, including supercell tests and 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 cycled data assimilation
10	Implementation plan (including costs)

Table 1.2: Phase 2 Testing Evaluation Criteria.

The GFS physics was implemented in both FV3 and MPAS for Phase 2 evaluations. In addition, each dycore was required to have a tracer capability that could be measured for conservative properties. The availability of a common GFS physics interface that enabled the models to run with the same physics codebase was critical to provide insight on conservative properties of the dycores, grid imprinting, scalability, and simulation fidelity with idealized cases.

As testing progressed, data results and assessments associated with each evaluation criterion were addressed during bi-weekly and then weekly telecons and through extensive email communications amongst the DTG.

As the DTG began to discuss and evaluate the results from the variable-resolution tests (Criterion 5), in particular the high-resolution, real-data hindcast of the Moore Tornado, May 19-21, 2013 and the supercell Idealized case described by Klemp et al. (2015), a number of questions arose pertaining to the use of GFS physics in the simulations. The DTG determined it would be beneficial to obtain external guidance from independent convective-scale subject matter experts. The DTG consultants drafted specific questions and requested insight from a group of subject matter experts, agreed upon by the DTG. Specific details describing the questions and responses can be found in Appendix C. The general outcome of this interaction was a determination that neither core had any obvious defects that would preclude its development into a model fully capable of forecasting and simulating moist deep convection.

The DTG met for a face-to-face meeting in Silver Spring, MD, May 4-6, 2016, to review, evaluate and discuss the Phase 2 results. During the meeting, the DTG continued to deliberate on issues related to the GFS physics implementation in MPAS, and NCAR was given 60 days to examine the issues and report the findings to the DTG, and then to resolve and resubmit the test results. On May 20, 2016, NCAR formally withdrew from the process by submitting a letter terminating their participation in the dycore

evaluation. Following the NCAR withdrawal, the DTG continued finalizing any remaining gaps in the testing results with available data and resources.

The AVEC contributed to the evaluation of Phase 2 criteria 4, 5, and 10, and the results were submitted in a final report to the DTG (AVEC Report: NGGPS Phase 2 Benchmarks and Software Evaluation, Appendix D). Both dycore groups submitted code packages to run each set of Phase 2 tests to the test manager so that the results could be verified independently as deemed necessary. The assessment, results, evaluations, and conclusions are described for each criterion within Chapter 2.

Chapter 3 documents the findings of the DTG that GFDL's FV3 is the optimal candidate for NOAA's new operational dycore.

Chapter 4 summarizes Phase 3, the integration and implementation of the new dycore into the GFS.

In summary, the hallmarks of this evidence-based decision making process include the following: an objective/outcome was clearly identified; test plans were developed, executed, and revised as needed; independent test leads conducted the testing and analyses; all of the developers could review each other's model configurations prior to testing and could also weigh-in on the analyses; internationally regarded external subject matter experts provided independent expert guidance, which was invaluable in resolving differences of opinion and interpretation; clear decision authority resided with the program leadership; finally, the DTG operated in accordance at each step of the way and the final recommendation was representative of the DTG⁴.

Chapter 2 Phase 2 Evaluation

2.1 Criterion 1: Plan for Relaxing Shallow-Atmosphere Approximation (Deep Atmosphere Dynamics)

NGGPS will be required to support service requirements spanning the full atmosphere from sensible weather near Earth's surface to space weather. Therefore, NCEP/EMC requested that the NGGPS dycore have the ability to relax the shallow atmosphere approximation currently used in all NOAA operational weather forecast models. In a deep atmosphere model, the distance to the center of the Earth, the gravitational acceleration, and grid-cell areas are all height-dependent. 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, the DTG requested 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. In response, the modeling groups included the requested information as part of their overall documentation package (which was submitted in response to Criterion 8). This information was forwarded to the Space Weather Prediction Center (SWPC) for evaluation. Below is the verbatim email response from Dr. Rodney Viereck from SWPC:

⁴ Director, NWS approved the DTG recommendation on 26 July 2016 (Appendix A).

“The SWPC WAM development team has considered approaches to space weather requirements outlined in the FV3 and MPAS-A dycore descriptions of Jan-Feb 2016. Both teams address most of the requirements presented by SWPC to the DTG and dycore developers in Aug 2015. Similar approaches are proposed to some of the requirements such as conversion to deep atmosphere equations and implementation of elliptic solvers for implicit treatment of fast diffusive processes in the thermosphere. Potential impact of these changes on computational efficiency is presently unknown. They will require substantial development and testing efforts, perhaps after the delivery of a dycore selected for tropospheric weather application.

There are also some differences such as a novel suggestion by the MPAS-A team to implement a grid with a horizontal resolution variable in height to enable larger time steps in the presence of strong horizontal winds higher up. However, the feasibility, development effort needed, and computational efficiency of this approach will have to be further explored.

Some requirements are not fully addressed by either team, such as the approach to thermodynamics in a whole atmosphere. Both dycores presently rely on potential temperature as the thermodynamic variable, which cannot be formally defined in a multi-species gas. This will require further communications between WAM and dycore developers.

In conclusion, presently both the FV3 and MPAS-A dycores are very close in addressing the space weather requirements for the next generation WAM and no preference may be given to either team based on this criterion. However, it is important to note that neither version has been run, tested, or validated with a raised lid to 600 km with updated physics for space weather needs. Significant effort still remains to adapt both dycores to the full atmosphere altitude/pressure domain currently covered by WAM. The DTG expects WAM model developers at SWPC will need to work closely with the chosen dycore developers and EMC to ensure future applications to space weather operational needs are accommodated.”

2.2 Criterion 2: Accurate Conservation of Mass, Tracers, Entropy and Energy

2.2.1 Overview

A variant of the idealized baroclinic wave test case used in Phase 1 testing with large-scale condensation (DCMIP 2012 case 4.2) and extra tracers run at 15 km resolution was used to assess the conservation of certain derived quantities that have particular importance for weather and climate applications. These quantities are dry mass, equivalent potential temperature, entropy and total energy. This test was also used to assess the degree to which the computational grid imprints itself on the numerical solution ('grid imprinting'), making use of the fact that the southern hemisphere is expected to be quiescent so that the grid imprinting signal is not masked by background variability.

2.2.2 Test Setup

The dry baroclinic wave test case described by Jablonowski and Williamson (2006) was used, with and without a simple parameterization of large-scale condensation as described in section 4.2 of the DCMIP 2012 test specification (available at https://www.earthsystemcog.org/site_media/docs/DCMIP-TestCaseDocument_v1.7.pdf). FV3 used a nominal horizontal resolution of 13 km (a 768x768x6=3,538,944 point grid) while MPAS used a 15 km nominal mesh (2,621,442 points). MPAS was run with 64 vertical levels with a dynamics time step of 60 seconds. FV3 ran with 60 vertical levels and a dynamics time step of 150 seconds. Both groups were instructed to run 20 days of simulation.

The FV3 group delivered all 20 days while the MPAS group only delivered results to 15 days for the moist case. The MPAS group pointed out that the initial conditions as specified in the DCMIP 2012 document resulted in a convectively unstable atmosphere in the tropics. As a result, moist overturning resulted in vertical velocities exceeding 20 ms^{-1} in the moist simulation at day 15 of the MPAS solution and the simulation subsequently aborted. The MPAS group elected not to submit a revised simulation running the full 20 days with a smaller time step.

A reference solution with the operational spectral semi-Lagrangian GFS dycore was also run, at a spectral resolution of T1534 (nominally 13 km) with 64 levels.

2.2.3 Diagnostics

2.2.3.1 Dry Mass: For MPAS, dry density (the density of dry air) is a prognostic variable. Therefore, the dry mass was computed from the vertical integral of the dry density in height coordinates. For the GFS and FV3, dry mass was inferred from the total surface pressure by subtracting the surface pressure associated with the total water content of the atmosphere in each column. Since MPAS is a height coordinate model with a rigid lid at a specified height, only the mass between the surface and lid is accounted for. FV3 and GFS use vertical coordinates based on pressure and the surface pressure implicitly takes into account the implied mass above the top of the model. Therefore, the dry mass estimate for MPAS is less than the value inferred from the surface pressure fields from FV3 and GFS.

2.2.3.2 Entropy: The total entropy of an air parcel should be conserved under reversible moist thermodynamic processes. The simple moist condensation scheme used in the test is not quite reversible, since it allows the condensed water to instantaneously fall out as precipitation. The material conservation of entropy is measured using the equivalent potential temperature, which is approximated by

$$(1) \quad \theta_e = T \left(\frac{p_0}{p_d} \right)^{R_d/C_{pd}} e^{rL_v/C_{pd}T}$$

where T is temperature, p_0 is a reference pressure (taken to be 1000 hPa), p_d is the ‘dry’ pressure (not including the effects of moisture), R_d is the specific gas constant for dry air, r is the water vapor mixing ratio, L_v is the latent heat of vaporization and C_{pd} is the specific heat capacity of dry air at constant pressure. This approximation neglects the component of the heat capacity associated with liquid water and a term associated with relative humidity raised to a negative power much smaller than one. The entropy is simply the natural log of equivalent potential temperature multiplied by C_{pd} .

2.2.3.3 Total Energy: The total energy should be conserved globally if there are no fluxes through the boundaries (which is true for this simple test case). However, the simple moist physics used for this test does not convert latent heat released into heat stored in liquid water. It simply removes the relevant amount of water vapor. Therefore, one may expect a slight loss of total energy for the moist case. The total energy density is computed as

$$(2) \quad \rho \left(U + \Phi + \frac{1}{2} |\vec{u}|^2 \right)$$

where ρ is density, U is the internal plus latent energy per unit mass of moist air, Φ is the geopotential and $\frac{1}{2} |\vec{u}|^2$ is the kinetic energy computed using the vector wind.

2.2.4 Results

2.2.4.1 Dry Mass: Figure 2.1 shows the time series of dry mass (converted to units of surface pressure) for FV3, MPAS and GFS for the dry and moist versions of the test case. The continuity equations in FV3 and MPAS take into account the change in total air mass associated with condensation and, as a result, both models accurately conserved dry air mass for both the dry and moist test cases. However, the GFS continuity equation is expressed in terms of total mass and does not take condensation into account. For the moist case, there is an apparent 5 Pa increase in dry surface pressure in the GFS solution. The GFS does not exactly conserve dry mass even in the absence of moist processes, although the increase in dry surface pressure is less than for the moist case. When run in operations, the GFS includes a ‘dry mass fixer’, which artificially forces the total dry air mass to stay equal to the initial value at every time step. Such a fixer will no longer be necessary if either FV3 or MPAS replaces the hydrostatic spectral dycore in the current GFS.

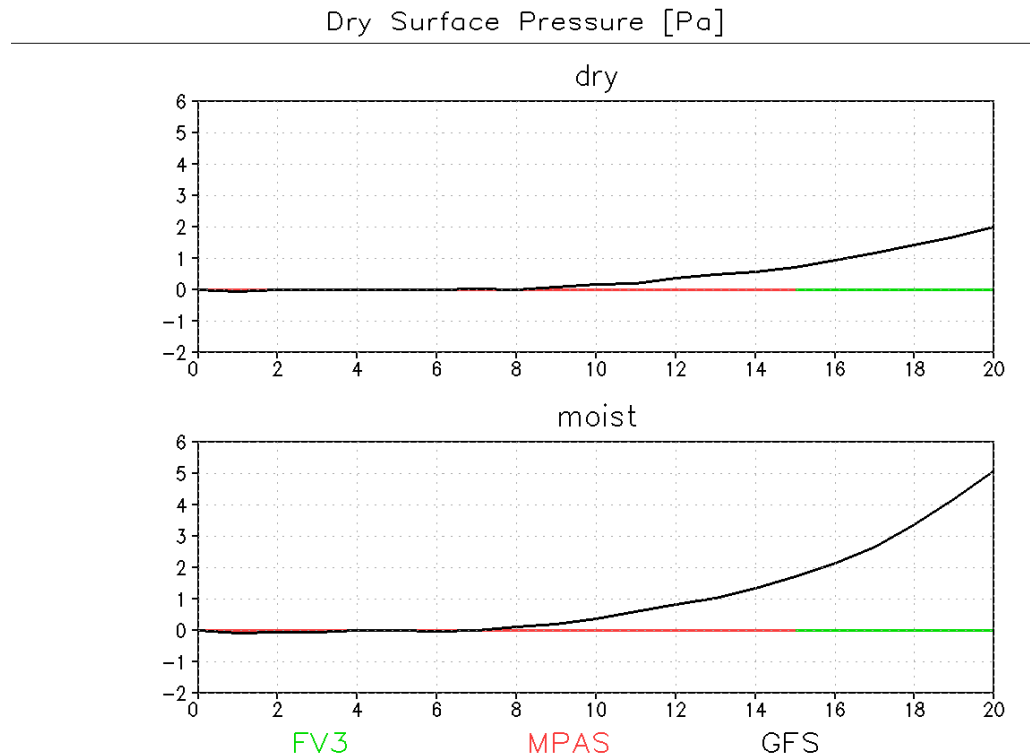


Figure 2.1: Change in global average dry surface pressure by day (x-axis) from the initial value for the dry (top panel) and moist (bottom panel) test cases in Pa.

2.2.4.2 Total Energy: Figure 2.2 shows the time series of the global integral of the total energy for the dry and moist test cases. The total energy change is very small for all three models in the dry test (much less than a hundredth of a percent of the initial value). For the moist test case, GFS loses significantly more energy than either FV3 or MPAS. The energy loss for FV3 and MPAS is on the order of 0.02-0.03 percent (versus 0.07 percent for GFS) over 15 days, which is consistent with the fact that the simple moist physics scheme used for this test does not convert latent heat released into heat stored in liquid water.

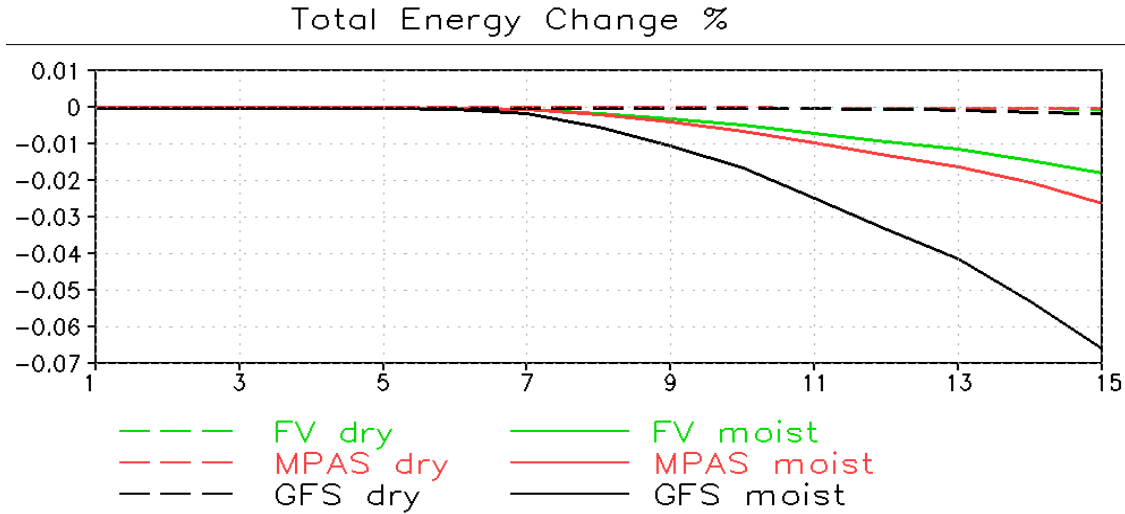


Figure 2.2: Total energy change by day (x-axis) for the dry (dashed) and moist (solid) test cases.

2.2.4.3 Total Entropy: The time series of globally integrated entropy are shown in Figure 2.3. Total entropy is very well conserved in all three models for the dry test. For the moist test, there is about a 0.01 percent decrease in the total entropy over 15 days for all three models. This decrease is consistent with the approximations made in the calculation of total entropy, including the neglect of the heat capacity associated with liquid water.

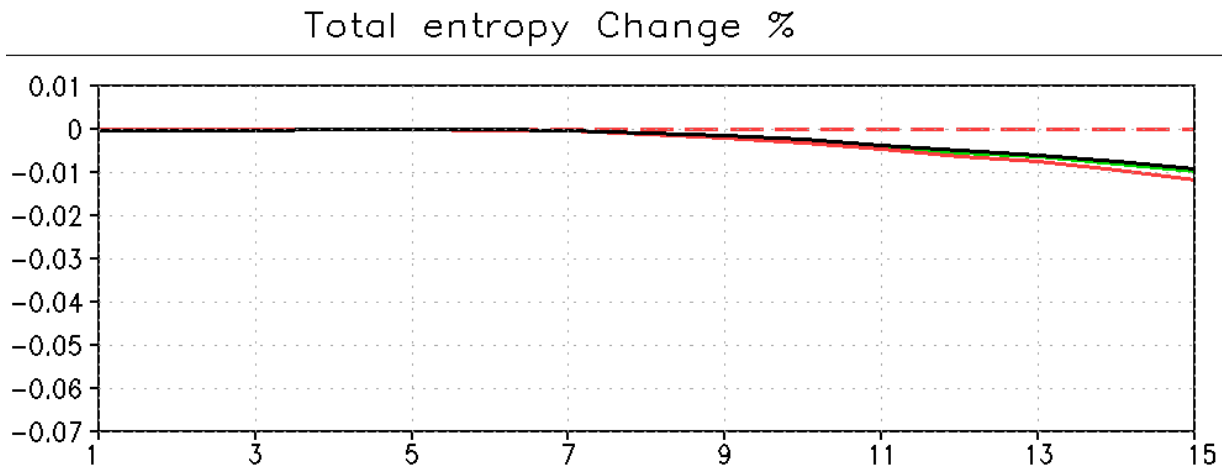


Figure 2.3: Total entropy change by day (x-axis) for the dry (dashed) and moist (solid) test cases. The color scheme is the same as in Figures 2.1 and 2.2.

2.2.4.4 Equivalent Potential Temperature: The material conservation of equivalent potential temperature is evaluated using the diagnostic framework of Johnson et al. (2000). An additional advective tracer is added to all three models and initialized equal to the equivalent potential temperature diagnosed from the temperature, pressure and humidity fields using equation (1) on page 15. The difference between the forecast tracer values and the diagnosed equivalent potential temperature is then used as a measure of the accuracy of the model numerical approximations in

simulating reversible thermodynamic processes. Figure 2.4 shows a scatter plot of the advected tracer versus the diagnosed equivalent potential temperature at day 15 from the moist test case solution. The RMS differences are small for all three models (0.1-0.2K), at least when compared to the values computed for the NCAR CCM3 model presented in Johnson et al. (2000), which were $O(10K)$ (see Figure 1 of that paper). The RMS differences are an order of magnitude smaller for the dry test case (see left most panel in Figure 2.5). Figure 2.5 shows the RMS differences at day 15 for the dry and moist cases as a function of pressure, with insets illustrating the differences very near the surface and top of the model domains. The large RMS values for MPAS and GFS near the top of the model are likely related to the upper boundary condition and the highly diffusive ‘sponge’ layers used to control noise near the upper boundary. FV3 has relatively small RMS values near the model top, likely due to the differences in the upper boundary and sponge layer treatments. The baroclinic wave solution involves the formation of sharp fronts at the surface in the potential temperature and moisture fields. Numerical errors are likely to be greater near such features; this might explain the large RMS values near the surface seen in the moist case for all three models. FV3 was configured to use a longer time step for tracer advection than dynamics for this test, and the variable used in the vertical re-mapping algorithm was temperature, not equivalent potential temperature. When FV3 is re-configured to use the same time step for the tracer advection algorithm as the dynamics, and potential temperature instead of temperature in the vertical remapping scheme, the RMS differences are smaller and comparable to those of MPAS in Fig. 2.5 (not shown⁵).

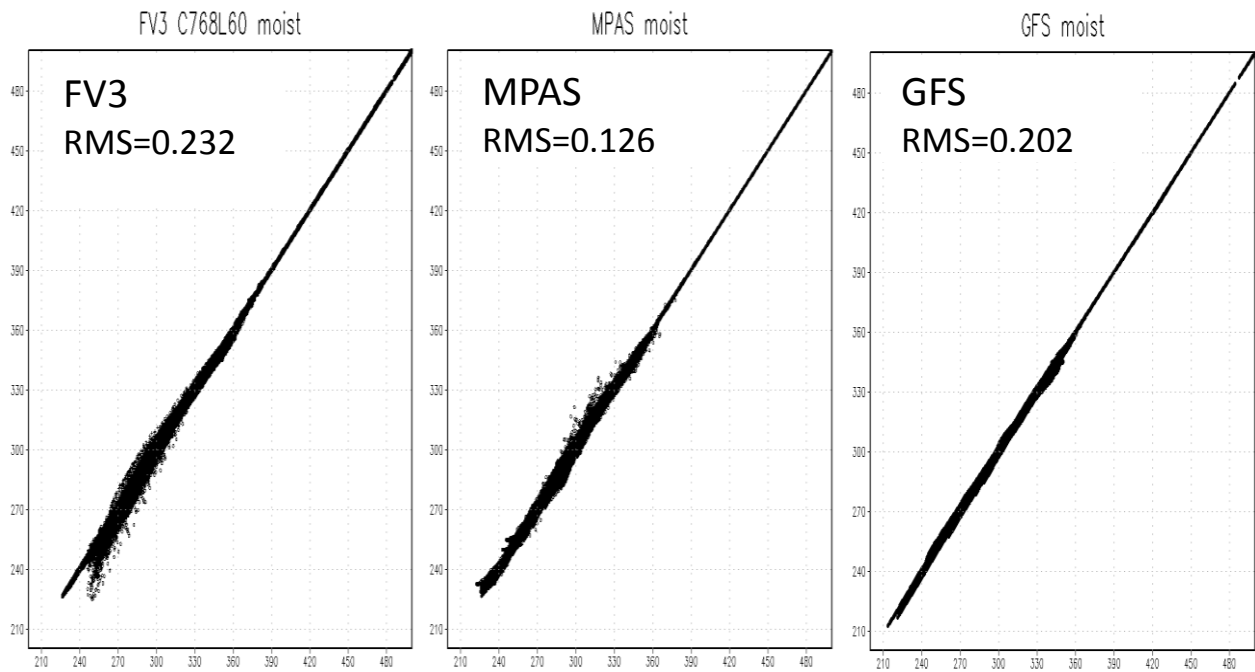


Figure 2.4: Scatterplot of advected tracer (y axis) versus equivalent potential temperature diagnosed from the moist test case solution at day 15 for FV3 (left), MPAS (middle) and GFS (right). These plots and RMS values use only equivalent potential temperatures up to 500K. See Figure 2.5 for the behavior near the model top.

⁵ The FV3 team re-ran the test in this configuration.

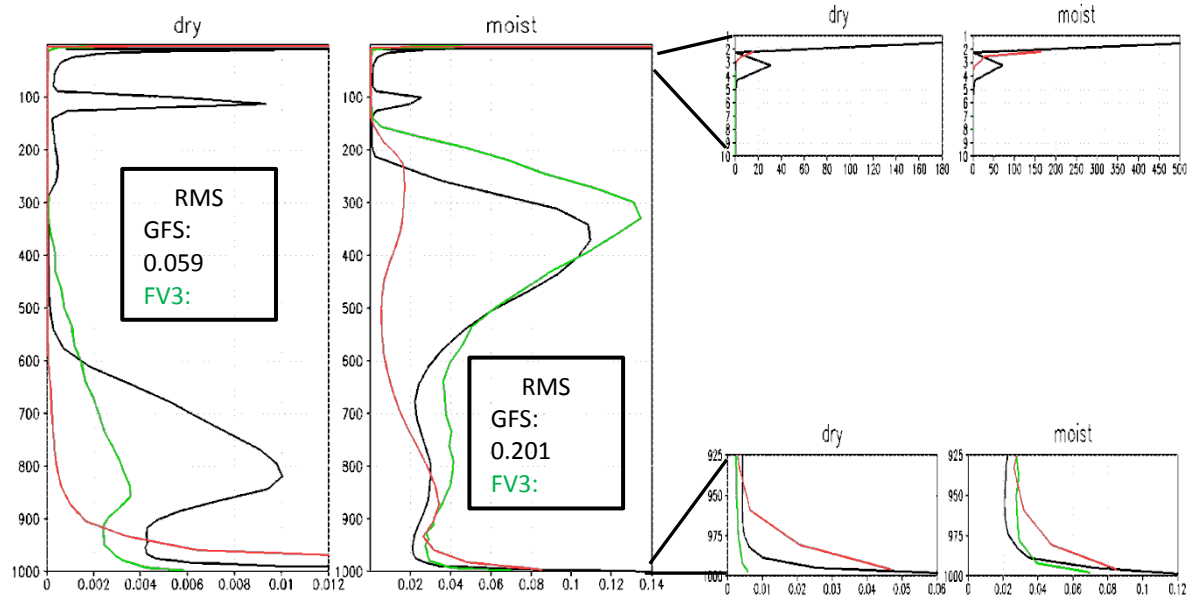


Figure 2.5: RMS differences between the advected tracer and the day 15 equivalent potential temperature as a function of pressure for the dry and moist test cases. Insets to the right illustrate the RMS values near the top of the model and the surface. The inset for the two plots on the left show the vertically integrated RMS values between the surface and 10 hPa. The RMS values for the entire model domain (from the surface to the model top) are 0.227 for FV3, 0.287 for MPAS and 0.7 for GFS. Caution is required in interpreting these values, since they are sensitive to the location of the upper boundary and the treatment of the upper boundary condition and sponge layer.

2.2.4.5 Grid Imprinting: The DCMIP baroclinic wave test case involves the growth of an unstable baroclinic wave packet on a zonal jet in the Northern Hemisphere. The baroclinic wave packet is initiated by a zonally localized perturbation in the mid-latitudes of the Northern Hemisphere. The initial conditions in the Southern Hemisphere are zonally symmetric and although the zonal jet is unstable, it is a steady solution of the non-hydrostatic equations and therefore the solution should remain zonally symmetric as long as no perturbations propagate across the equator from the Northern Hemisphere. Barring energy propagation from the Northern Hemisphere, any zonally asymmetric perturbations in the Southern Hemisphere solution are likely related to numerical errors, including those related to the increases in truncation error near the vertices and edges of the cubed-sphere grid and the pentagons on the icosahedral grid. Figure 2.6 shows the vertical velocity field at the first model level (with the zonal mean removed) at day 1 in the dry test case solution. The domains cover a 60 x 60 degree box centered on one of the cube corners for FV3 and one of the pentagons for MPAS. The vertical velocity field does show the signature of the computation grid. In particular for FV3, the cube edges near the corner are clearly visible. For MPAS, the pentagon is clearly visible in the vertical velocity field but the grid imprinting signal is smaller and more localized. The vertical velocity field is on the order of 10^{-4} meters per second, which is several orders of magnitude smaller than the vertical velocity signature associated with the fully developed baroclinic wave packet in the Northern Hemisphere (not shown). Figure 2.7 is a zoom in showing the detailed structure in a 2 x 2 degree box surrounding the points of interest. The structure of the computation grid is clearly visible in the vertical velocity field, including the polygonal cell in the MPAS mesh. The vertical velocity field in the vicinity of these ‘special points’ is steady in time (not shown) suggesting that it indicates a vertical circulation that is a dynamical response to time invariant spatial variations in truncation errors across these regions.

A small-scale numerical instability trapped near the lower boundary arises in both the Northern and Southern Hemispheres of the MPAS solution for both the dry and moist cases (Figure 2.8). The scale of the instability is 5-10 times the model grid spacing and much smaller than the expected baroclinic wave signal. No such small-scale instability is apparent in the FV3 solution.

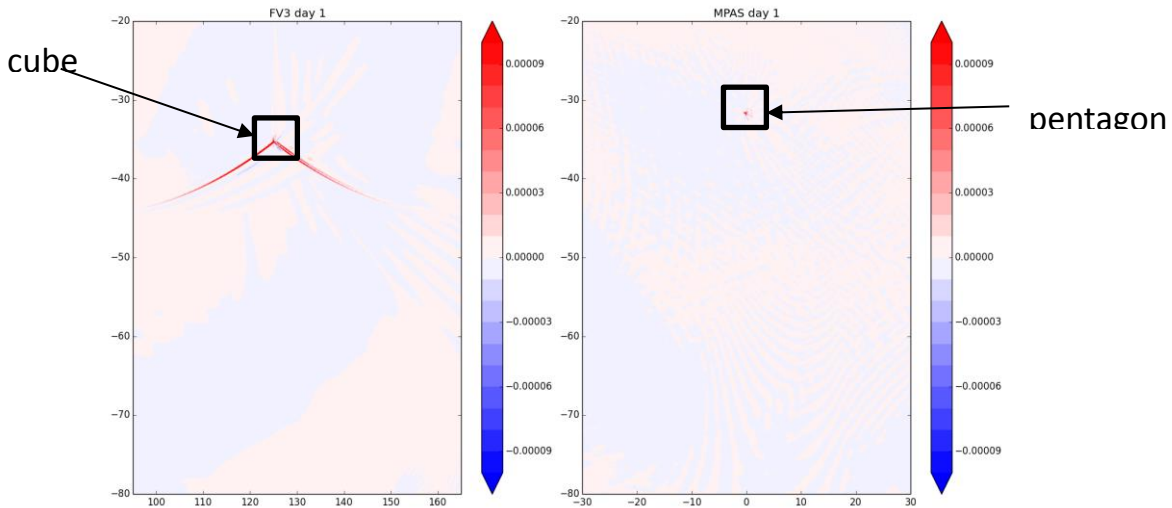


Figure 2.6: The vertical velocity field at the first model level at day 1 for the dry test case (zonal mean removed). Units are meters per second. Note the domain plotted for MPAS and FV3 are different. The FV3 domain is chosen to encompass a cube corner, and the MPAS domain is chosen to encompass one of the pentagons in the icosahedral mesh.

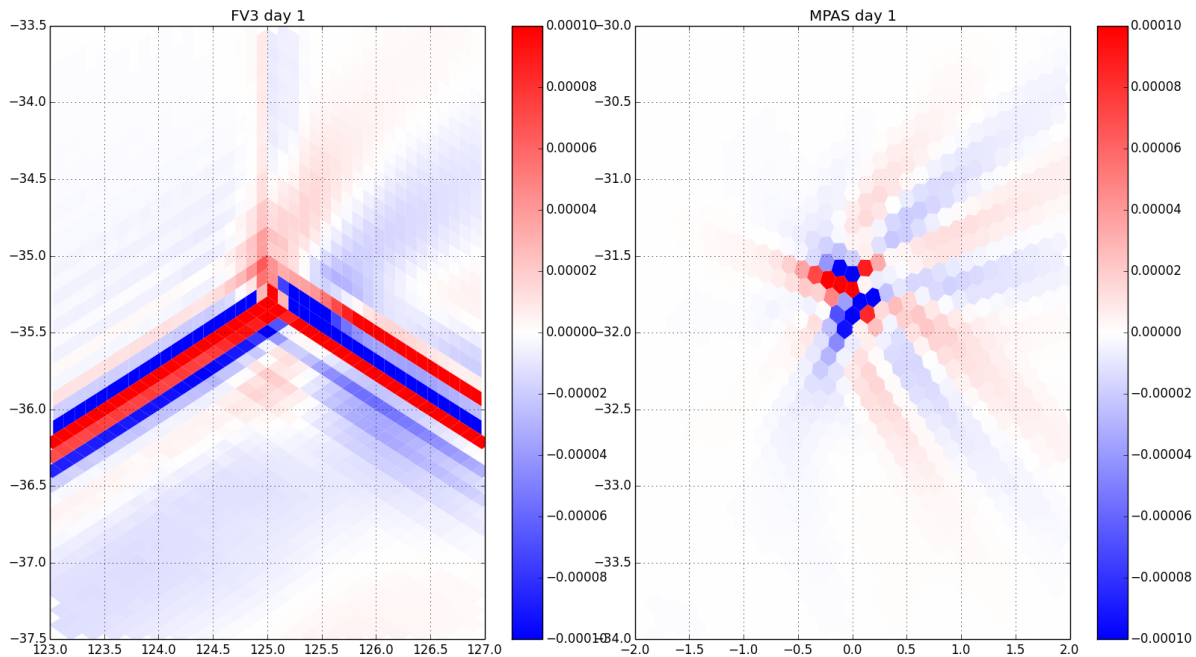


Figure 2.7: As in Figure 2.6, but zoomed in to a 2x2 degree box surrounding the points of interest. The vertical velocity is displayed by filling individual model grid cell polygons.

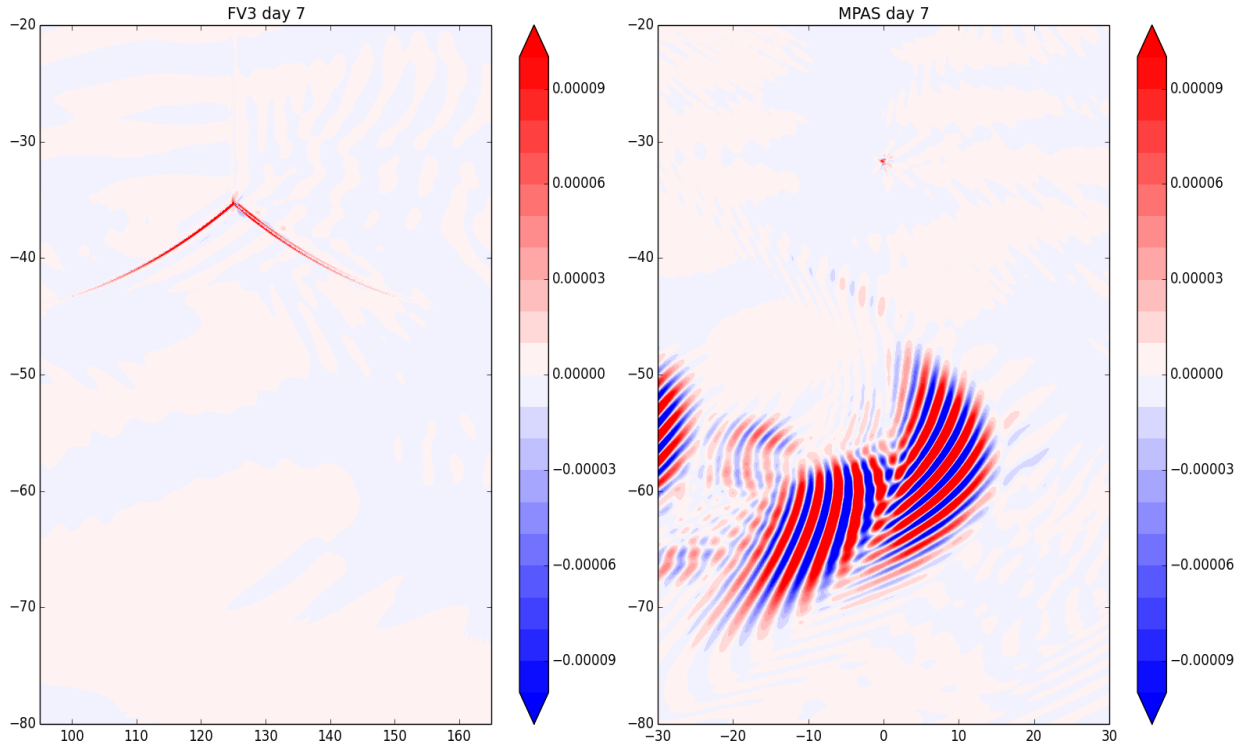


Figure 2.8: As in Figure 2.6, but for day 7 in the dry test case solution.

2.2.5 Conclusions

Accurate conservation of mass, tracers, total energy, and entropy - The conservation tests for energy and entropy (both moist and dry) reported above are impressive for both models (MPAS and FV3) and results are improved over the GFS. The DTG identifies no conservation issues with either model for these quantities.

The conservation of equivalent potential temperature in all three models shown here (FV3, MPAS and GFS) are all good and much better than for the NCAR CCM3 as shown by Johnson et al. (2000). Even the GFS shows great improvement over the CCM3, comparable to MPAS and FV3. The very large errors in MPAS and GFS near the top are likely related to the sponge layer treatments near the top of those models. All three models show significant errors near the surface particularly for the moist case. Note here that the errors shown in the upper troposphere for FV3 for the moist case were greatly reduced when the vertical remapping was changed to potential temperature and the tracer advection time step was made identical to the dynamics time step.

The equivalent potential temperature conservation in both FV3 and MPAS is considered excellent by the DTG.

Grid Imprinting - It was expected that there would be some evidence of anomalies from the grid appearing in the solutions of the global models analyzed in Phase 2. For MPAS, there are twelve pentagons required to complete spherical coverage with mostly hexagons. For FV3, there are the edges and vertices of the cube onto which the grid points are projected from the sphere. For the cubed-sphere grid there are eight vertices and three edges incident each of those vertices. The difference

equations must take on a different form for these special grid points and this can and usually does lead to higher truncation error.

Figures 2.6 and 2.7 clearly show some of the effects of this changed truncation error in the form of very weak stationary vertical circulations at the special points. As noted, this can be expected but the effect is very small, many orders of magnitude smaller than the vertical motions found in the Northern Hemisphere in these simulations with the growing baroclinic waves. It is not expected that the grid imprinting in either MPAS or FV3 will be noticeable in operational forecasts.

The low level instability noted in Figure 2.8 for MPAS is not explained here, but was not evident in either the FV3 or GFS solution.

References

Jablonowski, C. and Williamson, D. L., 2006: A baroclinic instability test case for atmospheric model dynamical cores. *Q.J.R. Meteorol. Soc.*, **132**, 2943–2975. doi: 10.1256/qj.06.12

Johnson, D. R., A. J. Lenzen, T. H. Zapotocny, and T. K. Schaack, 2000: Numerical uncertainties in the simulation of reversible isentropic processes and entropy conservation. *J. Climate*, **13**, 3860–3884, doi: [http://dx.doi.org/10.1175/1520-0442\(2000\)013<3860:NUITSO>2.0.CO;2](http://dx.doi.org/10.1175/1520-0442(2000)013<3860:NUITSO>2.0.CO;2)

2.3 Criterion 3: Robust Model Solutions Under a Wide Range of Realistic Atmospheric Initial Conditions Using a Common (GFS) Physics Package

2.3.1 Overview

This test component was designed to illustrate how the candidate dycores perform when confronted with realistic orography and initial conditions. The concept of “robustness” in this context means that the model solutions are stable and physically realistic. In order to isolate the impact of the dycores, it was necessary to implement a common physics package. The operational GFS physics was the obvious choice, since it allows for direct comparisons with the operational model. Forecast skill comparisons with the operational GFS also allowed EMC to estimate the work required to implement each of the candidate models in operations (which involves replicating or exceeding the forecast skill performance of the operational system).

2.3.2 Test Setup

Both FV3 and MPAS were configured to have close to the same nominal horizontal grid resolution as the operational T1534 GFS. For FV3, this resulted in $768 \times 768 \times 6 = 3,538,944$ grid points. For MPAS, a 3,504,642 point mesh was used. FV3 and MPAS both used a dynamics time step of 112.5 seconds. For MPAS this resulted in 6 forecast failures, and these forecasts had to be re-run with a smaller time step of 75 seconds.

NCEP EMC provided a developmental version of the National Unified Operational Prediction Capability (NUOPC) GFS physics application programming interface (API) (including input datasets) to both modeling groups, along with implementation assistance. The physics parameters were configured as in NCEP operations. Both models used a vertical distribution of model levels very similar to the GFS, with the top level omitted, resulting in 63 model layers with a model top close to the 2nd highest GFS model layer interface (~ 6.4 Pa or 68 km). Both modeling groups examined the other’s implementation of GFS physics, and some inconsistencies were noted and fixed. After some discussion within the DTG, it was

decided that the implementation of the sponge-layer near the model top should be left to the dycore and not be considered part of the GFS physics. The physics time step used for both FV3 and MPAS is 225 seconds, except for radiative processes which were calculated hourly. These are the same settings used in the operational GFS. Unlike the operational GFS, surface properties such as Sea Surface Temperatures (SSTs, vegetation fraction, albedos, etc. were held fixed at their initial condition specification for the duration of each forecast. Applications to interpolate the operational GFS analyses and build the orography, including orographic gravity wave coefficients, were delivered by NCEP to each modeling group. The MPAS team chose to use its own in-house applications. The FV3 team modified the NCEP applications to work with generic latitude-longitude grids, of which the cubed-sphere is but one representation. Figure 3.1 shows a plot of the orography over the Andes, indicating the resolution is similar in all three models. Neither FV3 nor MPAS has ‘spectral ringing’ over oceanic regions seen in the GFS orography.

GFS operational initial conditions were used, every 5 calendar days at 00UTC from January 16, 2015 to January 16, 2016 (a total of 74 cases). Forecasts were run out to 10 days, with output saved every 6 hours. Both models were converted from the native grid model output to a 3072 x 1536 regular latitude-longitude mesh (the same mesh used by the operational GFS) using tools provided by both modeling groups. The re-gridded data was saved in ‘GFS-lookalike’ files that were ingested by the NCEP post-processor. The files were post-processed at EMC and run through the suite of verification tools routinely used to validate pre-implementation versions of the GFS. Over 6000 diagnostics are available via an easy-to-use graphical interface. Only a very small sample is shown here to illustrate the key results. The full set of generated plots is available at <http://www.emc.ncep.noaa.gov/gmb/wx24fy/nggps/web>.

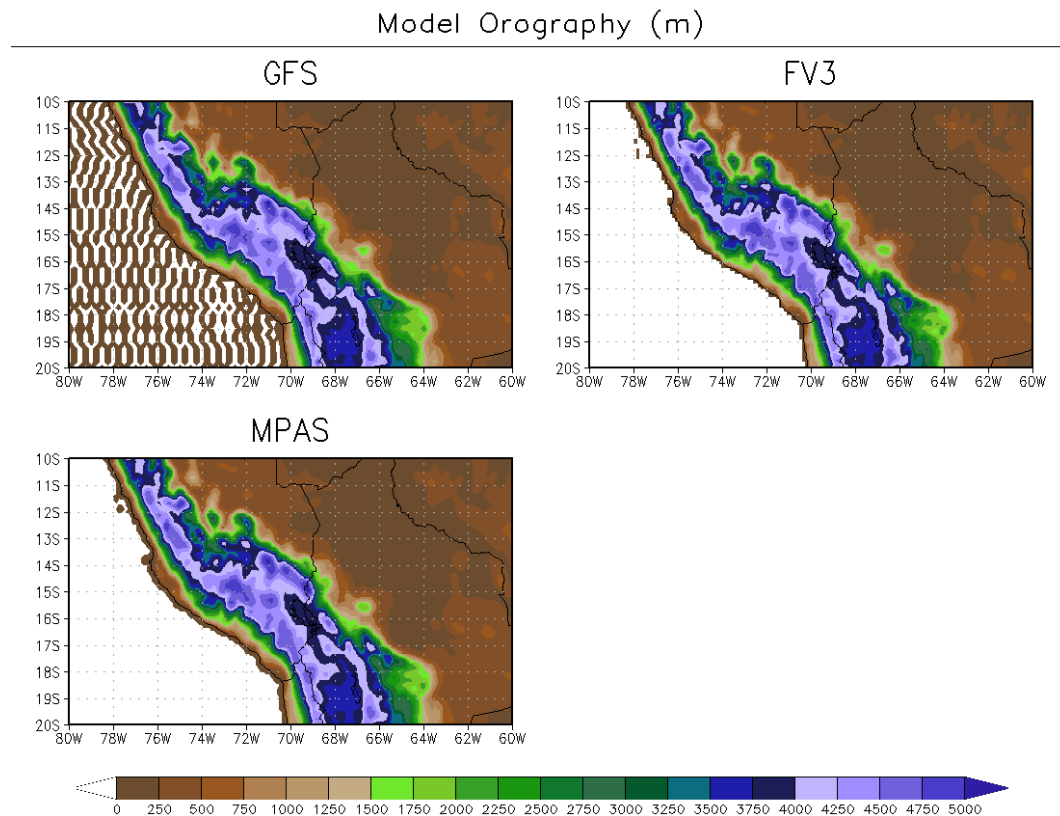


Figure 3.1: Model terrain over the Andes for GFS, FV3 and MPAS. Units are meters.

2.3.3 Effective Resolution

Appendix E describes the issues involved in estimating the ‘effective resolution’ of a forecast model. For the purposes of this discussion, ‘effective resolution’ is defined to be the scale at which numerical diffusion dominates and dynamical features can no longer be accurately simulated. This scale is estimated using kinetic energy spectra of winds near the tropopause (200 hPa). Figure 3.2 shows the kinetic energy spectra computed using all of the forecast output at day 10. For reference, the -3 and -5/3 power-law spectra (characteristic of two-dimensional and three-dimensional turbulence) are shown along with the scales corresponding to 10 and 4 times the nominal grid spacing (10Δ and 4Δ). Both the MPAS and FV3 spectra start to fall off sharply due to diffusion at approximately 4Δ . The GFS, however, falls off at a scale closer to 10Δ , indicating that MPAS and FV3 both have a significantly higher effective resolution than the current operational GFS even though the nominal mesh spacings are very similar. Both FV3 and MPAS show the expected shallowing of the spectrum in the mesoscale indicating a transition from two to three-dimensional turbulence, while the GFS does not. The fact that the effective resolution of MPAS and FV3 is so similar means that interpretation of forecast skill and computational performance comparisons should not be complicated by differences in the scale of features resolved by the models.

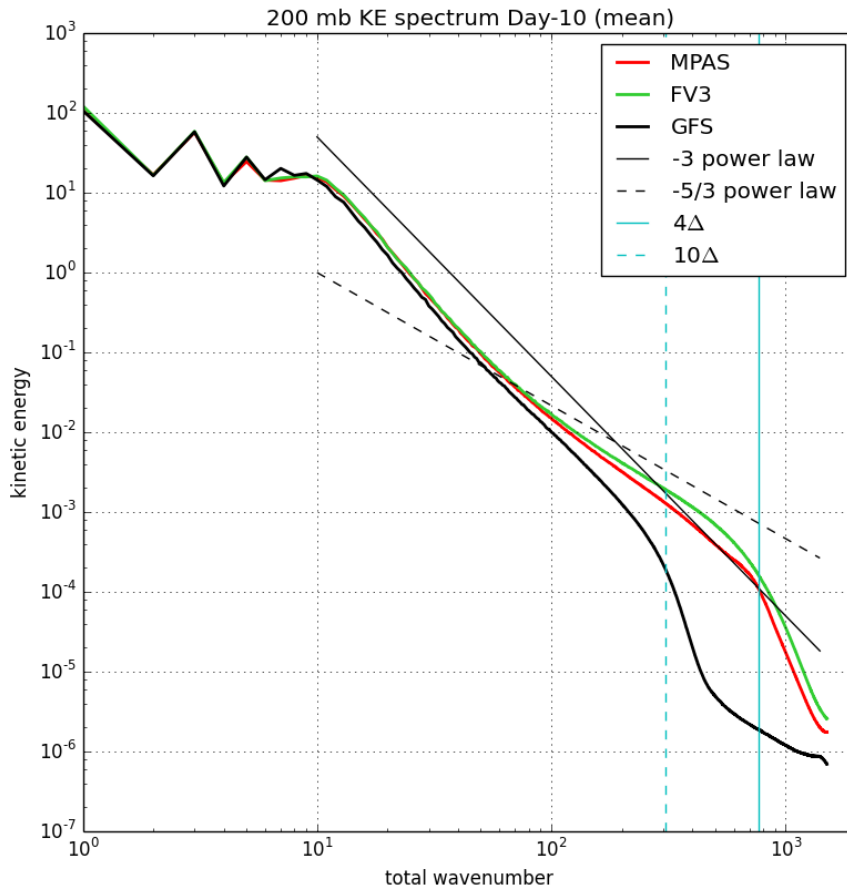


Figure 3.2: 10-day forecast 200 hPa kinetic energy (KE) spectra, averaged over all 74 forecasts. Reference power-law spectra corresponding to powers of -3 and -5/3 are shown for reference, as well as scales corresponding to 4 and 10 times the nominal grid resolution.

2.3.4 Global Precipitation Forecasts

Figure 3.3 shows the global mean precipitation rate for the three models as a function of forecast lead time. MPAS has slightly less precipitation and a somewhat longer spin-up period. The global mean value at day 10 for FV3 (MPAS) is slightly larger (smaller) than the GFS value of 3.1 mm/day. On a global scale, maps of precipitation rate averaged over the 74 cases look very similar (not shown), but zooming in over the Andes shows some differences (Figure 3.4). The GFS precipitation is quite noisy over and to the east of the Andes, with grid-scale maxima as high as 240 mm over the 6-h period. Both MPAS and FV3 appear to have a more realistic representation of precipitation near high terrain.

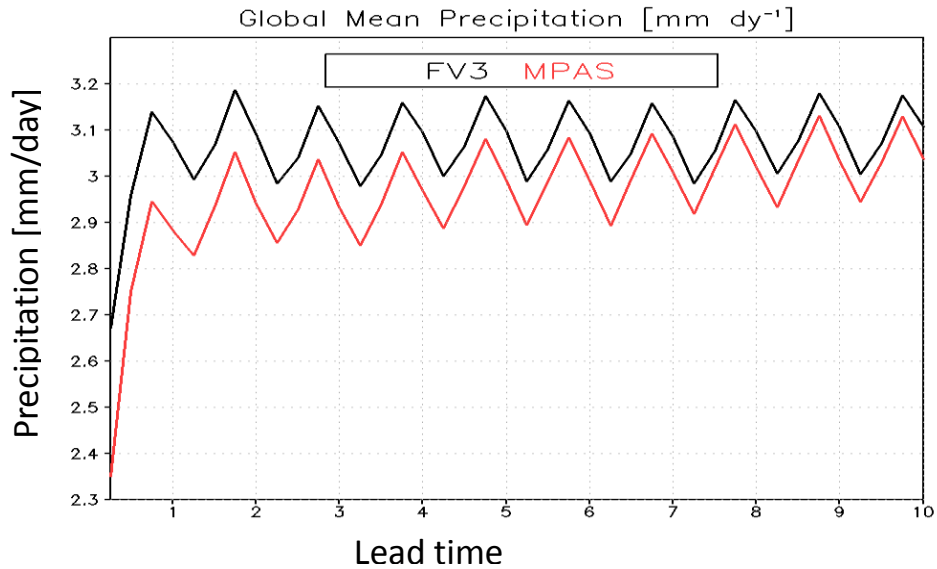


Figure 3.3: Global mean precipitation [mm day⁻¹]. For reference, the GFS global mean at day 10 is 3.1mm.

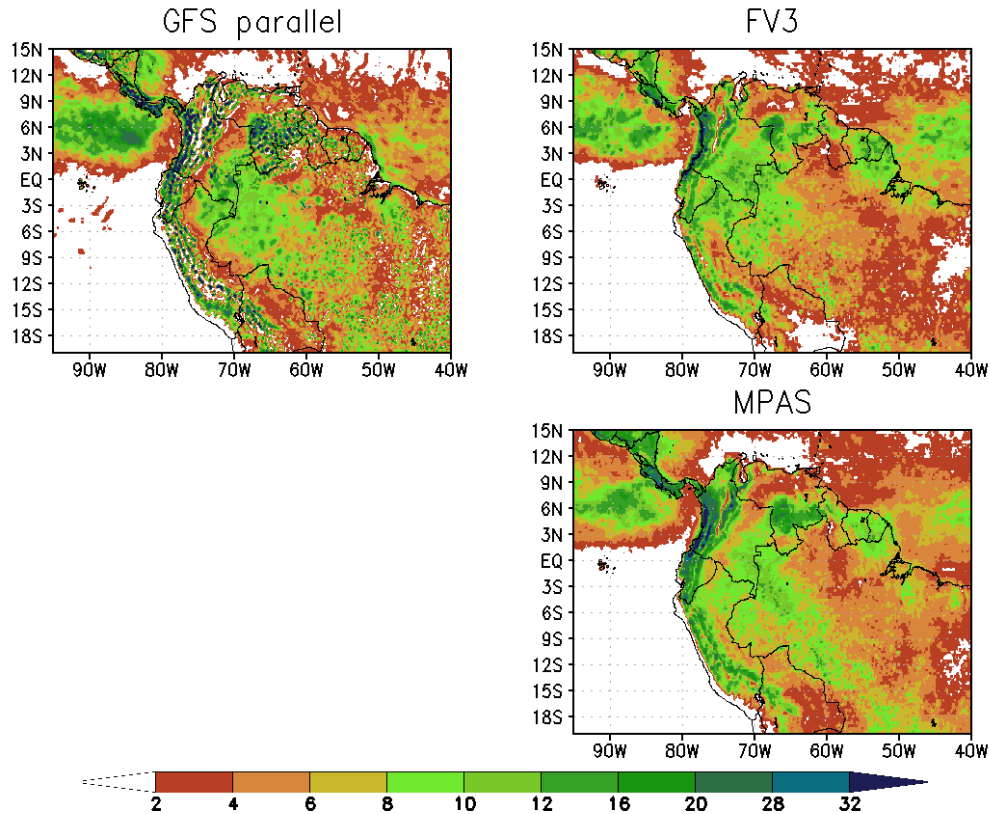


Figure 3.4: Maps of precipitation accumulated [mm] between forecast hour 234 and 240, averaged over all cases.

2.3.5 Forecast Skill

Figure 3.5 shows a time series of 5-day 500 hPa height anomaly correlation for the Northern Hemisphere extra-tropics, taken from the NCEP EMC verification statistics website at <http://www.emc.ncep.noaa.gov/gmb/wx24fy/nggps/web>. FV3 and GFS have very similar skill scores (they are not statistically significantly different at the 95% level), and the MPAS scores are significantly worse. Results are similar in the Southern Hemisphere extra-tropics, but in the tropics the vector wind RMS errors for the GFS are significantly better than either FV3 or MPAS at lead times less than 96-h (although FV3 still out-performs MPAS, see Figure 3.6). This difference in skill between FV3 and MPAS is confirmed by a wide range of diagnostics, such as time series of anomaly correlation, RMSE and bias for different fields, and error maps at different levels for different fields. All forecasts have been verified against the GFS analysis. This may be a factor favoring the GFS forecasts, particularly in the tropics and at short forecast lead times. However, given that the FV3 (and MPAS) dycore has significantly different effective resolution than the GFS, the DTG finds it remarkable that the FV3 skill scores so closely match the GFS (even without tuning of the resolution-dependent physics parameters to the FV3 dycore). Due to the relatively poor performance of MPAS, the NCAR development team was given an extra 60 days to investigate the possible presence of error(s) in their implementation of the GFS physics and rerun the forecasts. NCAR was also offered the opportunity to submit forecasts run with their own physics package. While errors were discovered in the initialization of sea ice and the cloud-water mixing ratio, their impacts were not believed to be significant enough to improve the overall forecast skill. It was early in this 60-day extension when NCAR made the decision to withdraw MPAS from the evaluation.

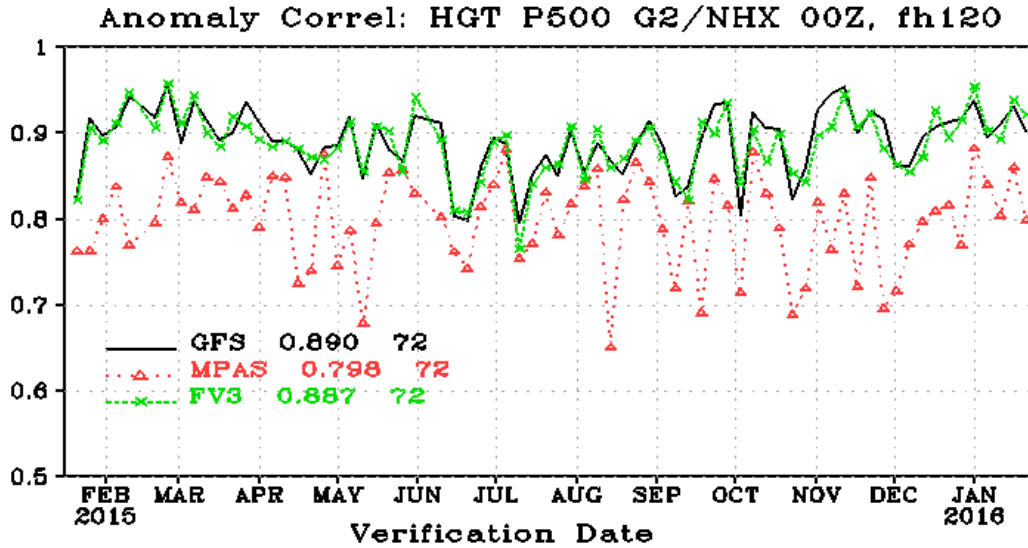


Figure 3.5: 500 hPa 5-day forecast anomaly correlation time series for the Northern Hemisphere poleward of 20 degrees.

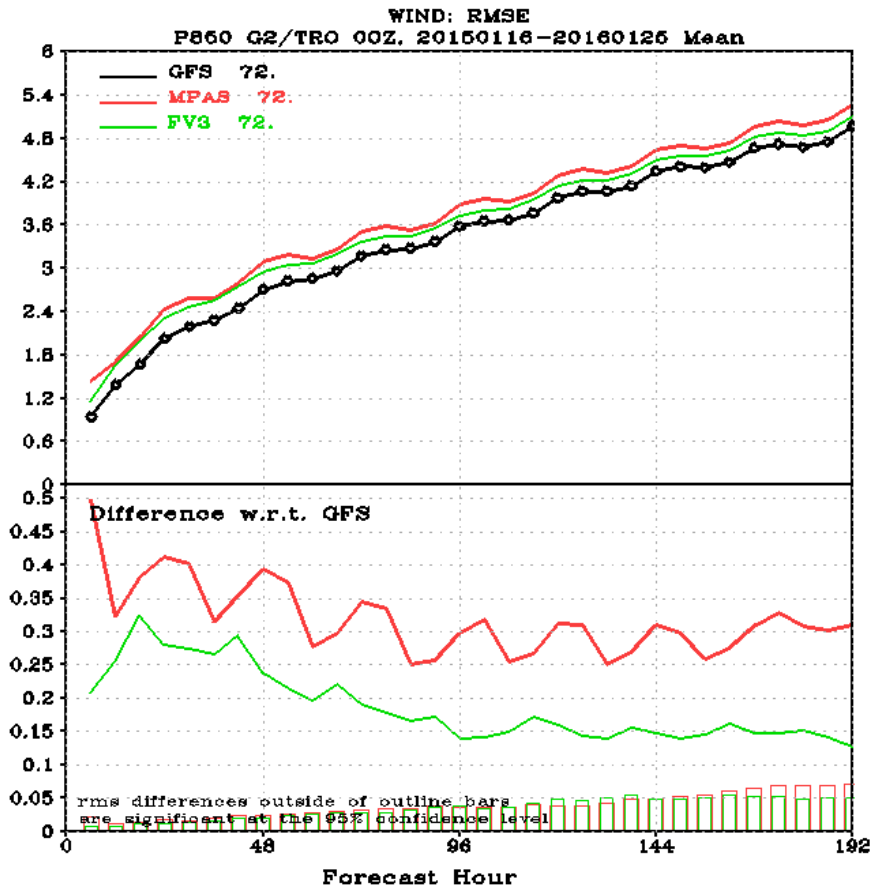


Figure 3.6: RMS vector wind error [m s^{-1}] at 850 hPa for the tropics (between 20 degrees North and South).

2.3.6 Robustness

As mentioned in section 3.2, six MPAS forecasts became computationally unstable and did not complete when using the same 112.5 second time step as FV3. Those six forecasts did run to completion with a 75 second time step. However, in many of the MPAS forecasts there was evidence of noise in the upper-tropospheric wind fields at a scale of roughly 4 times the nominal grid spacing. The signature of this is evident in the kinetic energy spectra (Figure 3.7), if all of the 74 individual spectra are plotted instead of just the mean as in Figure 3.2. There are a number of cases for which MPAS exhibits a spectral peak at around 4Δ . A map of 200 hPa zonal wind is shown for one of these cases (Figure 3.8). Noise near the grid scale is clearly evident in the jet stream over the Pacific Ocean for the MPAS forecast, but is absent in the forecast from FV3.

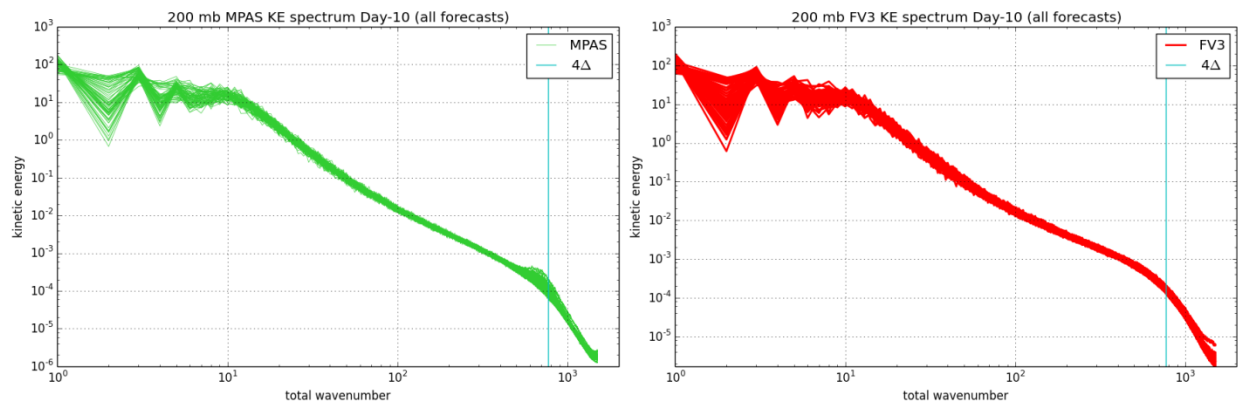


Figure 3.7: 10-day forecast 200 hPa kinetic energy spectra for each of the 74 forecasts for FV3 (left) and MPAS (right). For reference, the scale corresponding to 4 times the nominal grid spacing is shown as a vertical line.

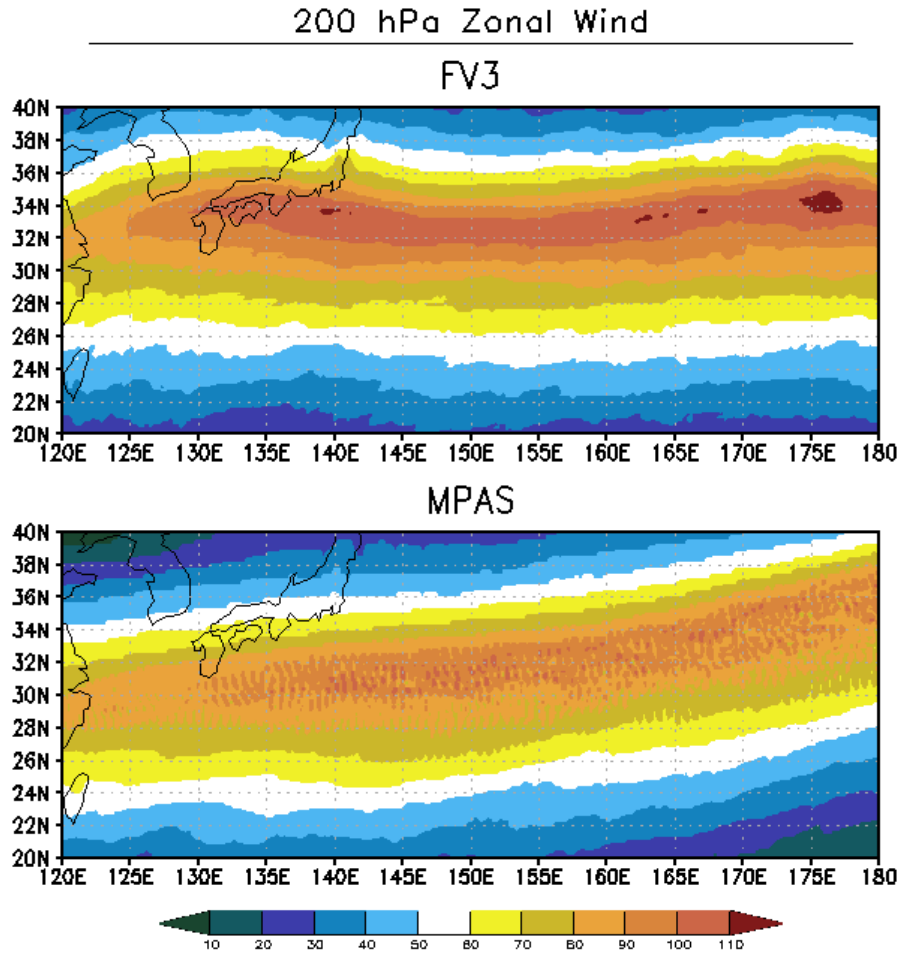


Figure 3.8: Zonal wind [m s^{-1}] at 200 hPa for a 10-day forecast initialized at 00UTC 16 Jan 2015.

2.3.7 Conclusions

Effective resolution - There are a number of ways to infer the effective resolution of a model and after considerable discussion, the DTG elected to use the kinetic energy spectra for that purpose. In particular, the minimum effective resolution for the model was interpreted as occurring where the spectra slope deviated significantly from either a -3 or $-5/3$ power law. This marks the point where the diffusive processes remove energy from the spectra and prevent buildup of spurious noise. Figure 3.2 illustrates a very important result concerning the effective resolution of FV3 and MPAS as compared to the GFS. The point where the FV3 and MPAS spectra start to drop off is around 4Δ or 4 times the nominal grid spacing, compared to around 10Δ for the GFS. Thus for the same resolution in grid spacing, both FV3 and MPAS have over twice the effective resolution of the current operational GFS. Both MPAS and FV3 develop a $-5/3$ power law range in their spectra in contrast with GFS, suggesting that those two dycores are better at representing mesoscale phenomena than GFS.

Both MPAS and FV3 have higher effective resolution for the same nominal resolution as the current operational GFS.

Global precipitation forecasts - The global precipitation forecasts for both FV3 and MPAS were similar to each other and similar to GFS. The main difference was that the GFS precipitation shown in Figure 3.4 was noticeably noisier than either FV3 or MPAS, especially near steep orography.

Forecast Skill - It is in this test where the main difference between the skill performance of FV3 and MPAS is apparent. Figure 3.5 shows that the skill of FV3 is comparable to that of GFS while MPAS lags considerably. Similar differences are apparent in other plots of the EMC verification statistics available at <http://www.emc.ncep.noaa.gov/gmb/wx24fy/nggps/web>. In addition, the similarity in skill to the GFS suggests that the work necessary to bring FV3 to a performance level equal to or better than the operational GFS will be considerably smaller than for the MPAS.

The DTG feels that the differences between the two cores shown in this comparison indicate that FV3 is a better choice for the new GFS dycore than MPAS.

Robustness - There were several examples where the MPAS showed some tendency to become unstable at the time steps used in tests. Examples include those noted above (Figures 3.7 and 3.8), and Figure 2.8 in the previous section. It is likely that all these noted instabilities could be eliminated with a shorter time step but of course that would increase the computational cost of the dycore. None of these issues were noted with FV3.

2.4 Criterion 4: Computational Performance with GFS Physics

The AVEC, a DTG subcommittee, compared computational performance in benchmark tests of MPAS and FV3 conducted during dedicated access to the Cori supercomputer at the U.S. Department of Energy's National Energy Research Scientific Computing Center (NERSC). Test results showed FV3 provided significantly better computational performance, more efficient tracer advection, and more computationally efficient mesh refinement than MPAS. The full Phase 2 AVEC report is appended to this document as Appendix D.

2.5 Criterion 5: Demonstration of Variable-Resolution and/or Nesting Capabilities, including Supercell Tests and Physically Realistic Simulations of Convection in the High-Resolution Region

2.5.1 Overview

Although NCEP EMC has not yet defined requirements for nesting and/or variable-resolution for the NGGPS, it is anticipated that some capability will be required, especially for prediction of hurricanes and severe convection. Indeed, the fact that all of the candidate dycores are non-hydrostatic anticipates the likely eventual unification of global and convective-scale regional prediction systems. The purpose of this test criterion is to demonstrate and evaluate a baseline capability to provide enhanced, 'convection-permitting' resolution over certain regions. Two idealized tests were added in order to isolate the effect of dycore numerical methods on simulations of explicit moist convection and tropical cyclones, independent of the variable-resolution and/or nesting infrastructure. The two idealized tests are the supercell test (also used in Phase 1) and the DCMIP 2012 idealized tropical cyclone test.

2.5.2 Idealized Supercell Test

Klemp et al. (2015) describe an idealized supercell thunderstorm test that can be used to evaluate a dycore's ability to simulate deep, moist convection. The test is run without rotation on a sphere with radius 120 times smaller than Earth, so that global convection permitting resolutions can be achieved with minimal computational expense. A simple Kessler-type warm rain microphysics scheme is the only physical parameterization used, and numerical convergence can be achieved by specifying a constant Laplacian diffusion. In the Klemp et al. (2015) specification, the diffusion is applied to the full fields in the horizontal and to deviations from the initial profile in the vertical. The test was modified to use only horizontal Laplacian diffusion since 2nd-order vertical diffusion is not available in the FV3 dycore. Because of the Lagrangian vertical coordinate, it would be difficult to apply diffusion to the deviations from the initial profile. In Phase 1, each modeling group chose their own diffusion settings for this test, making it difficult to isolate the impact of dycore numerical methods on the simulations. Here FV3 and MPAS use the same constant Laplacian horizontal diffusion of $2000 \text{ m}^2\text{s}^{-2}$, applied to all dynamical and microphysical prognostic variables. The monotonic constraint for advection is disabled in both models, and the tracer advection scheme is integrated with the same time step as the dynamics. Solutions out to two hours are computed using nominal mesh spacings of 0.5, 1, 2 and 4 km. MPAS used a dynamics time step of 3, 6, 12 and 24 seconds (with 6 acoustic steps per dynamics step) and microphysics tendencies were computed every dynamics time step. FV3 used a dynamics time step of 2.5, 5, 10 and 20 seconds (with 5 acoustic time steps per dynamics time step) and microphysics computed every 20 seconds. The vertical resolution was set to 0.5 km for all four horizontal resolutions, with a model top at 20 km.

Figure 5.1 shows the FV3 and MPAS 500 hPa vertical velocity field 60 minutes into the simulation for each of the four resolutions. Convection is initiated by a single warm bubble centered on the equator in an unstable sounding with vertical shear. As expected, both FV3 and MPAS produce two separate updrafts north and south of the equator one hour into the simulation. The scale of the updrafts in both models is roughly 10 km, although the scale of the FV3 updraft appears to be slightly larger. The solutions for both models appear to be numerically converged at 500 m resolution. As the resolution is degraded, the structure of the grid mesh becomes more apparent in the solution. The 2 km solution in both models retains the basic character of the numerically converged solution. At 4 km resolution both models still produce two distinct updrafts, but the updrafts are poorly resolved and highly distorted. At later times (90 minutes and 120 minutes), additional circulations develop and the model solutions become less similar to each other (not shown) due to subtle differences in the interaction between the multiple updrafts and downdrafts.

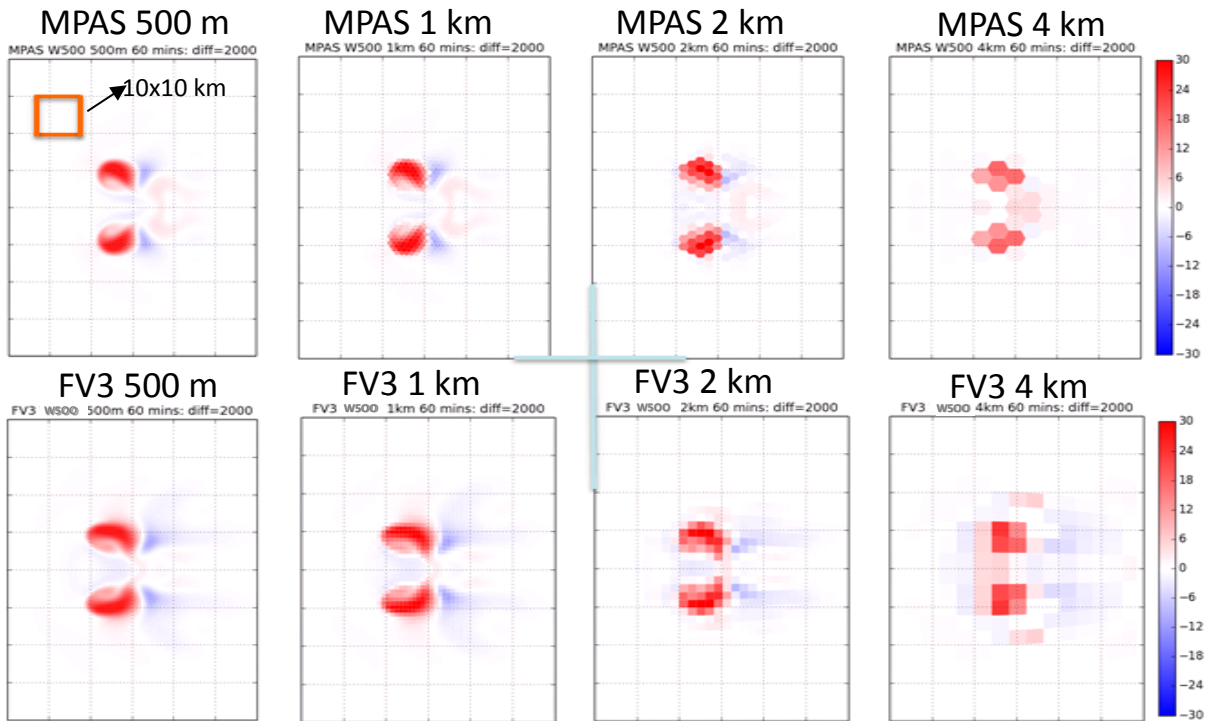


Figure 5.1: 500 hPa vertical velocity in m/s for MPAS (top) and FV3 (bottom) at 60 minutes into the simulation, for horizontal resolutions of 500 m, 1 km, 2 km and 4 km (from left to right). The orange box in the top left panel is 10 x 10 km in scale. Individual model grid cells are color-filled, so the influence of the computational grid on the solutions at coarser resolution can be seen. The background grid lines are drawn every 10 degrees of latitude and longitude on the reduced radius sphere.

MPAS and FV3 use very different grid staggering strategies (MPAS uses the Arakawa C-grid, while FV3 uses the D-grid, with winds computed by a C-grid solver used for flux calculations). With the magnitude of the horizontal diffusion used here, these differences of grid staggering do not appear to have a large impact on simulations of deep moist convection. Both dycores appear to be able to produce simulations of splitting convective supercells that are quite consistent with previous solutions in both spherical and Cartesian geometries. Several examples of solutions with other global dycores are available at the DCMIP 2016 website (test 3 at <https://www.earthsystemcog.org/projects/dcmip-2016/results>).

2.5.3 Idealized Tropical Cyclone Test

This test was part of DCMIP 2012 and it is described in detail in the DCMIP 2012 test specification document (available at https://www.earthsystemcog.org/site_media/docs/DCMIP-TestCaseDocument_v1.7.pdf) and in Reed and Jablonowski (2012). It was designed to elucidate the impact of dycore numerical methods on the simulation of tropical cyclones, in a setting intermediate in complexity between full-physics aqua-planet and idealized axi-symmetric settings. The test was run at 13 km nominal resolution (similar to the operational GFS) on the full sphere with 64 vertical levels, distributed similarly to the GFS. Since this test includes rotation, it is difficult to configure a reduced-radius version of this test that preserves the dynamical behavior of the full sphere. Therefore, this test was not run at convection permitting resolution due to computational constraints. Also, the DTG did not perform a reference run of the GFS for this test.

The simple physics suite used for this test includes a parameterization of large-scale condensation, surface fluxes of momentum, sensible and latent heat, and boundary layer turbulent diffusion. The surface stresses in the boundary layer formulation are sensitive to the height of the first model level – Reed and Jablonowski recommend that the first model level be placed close to 70 meters above the surface. The height of the lowest level in both FV3 and MPAS is close to 50 meters (51.2 for FV3, 46.9 for MPAS). It is not clear whether these differences impact the test results.

Figure 5.2 shows the initial test results for day 6, submitted by the MPAS and FV3 modeling groups. The FV3 tropical cyclone has a much larger circulation than the MPAS cyclone, and propagates farther north over the six-day simulation period (consistent with the enhanced beta-drift associated with the larger vortex). After some experimentation it was discovered that FV3 was configured with a Richardson-number 2Δ vertical filter enabled. This filter effectively enhanced vertical mixing in the boundary layer, increasing the depth of the inflow layer and resulting in a very large circulation. The test was re-run without the vertical filter (Figure 5.3). The size of the FV3 storm without the vertical filter is much more similar to MPAS, as is the storm location.

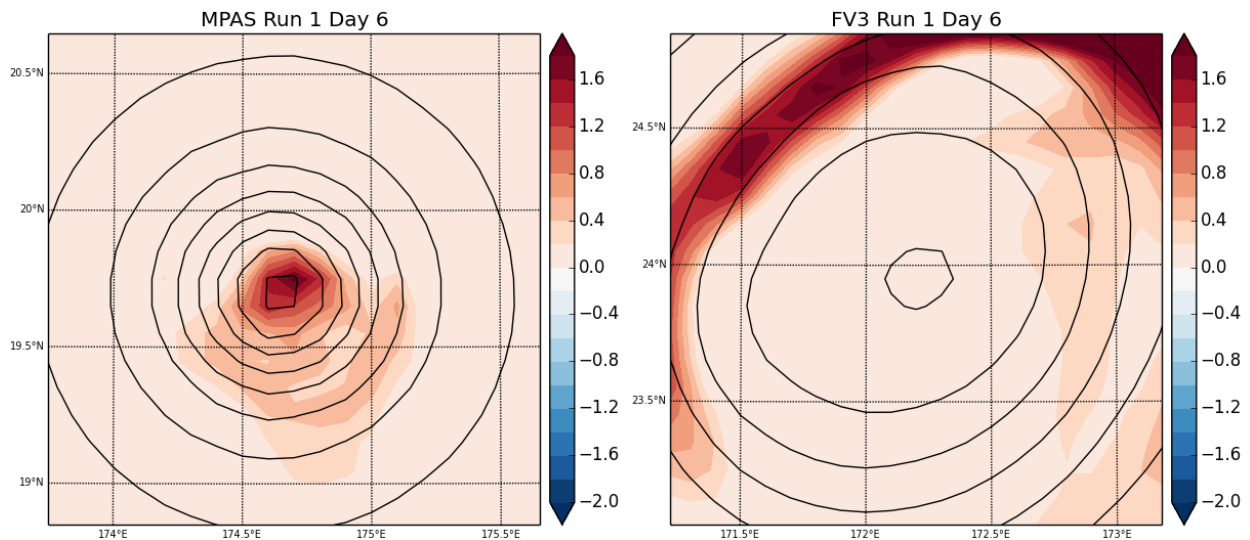


Figure 5.2: Surface pressure (contours, every 4 hPa) and vertical velocity (colors, in m/s) at day 6 for the MPAS (left) and FV3 (right) as for the tropical cyclone test. The domain plotted is storm-centric – the latitude and longitude labels are different for each plot.

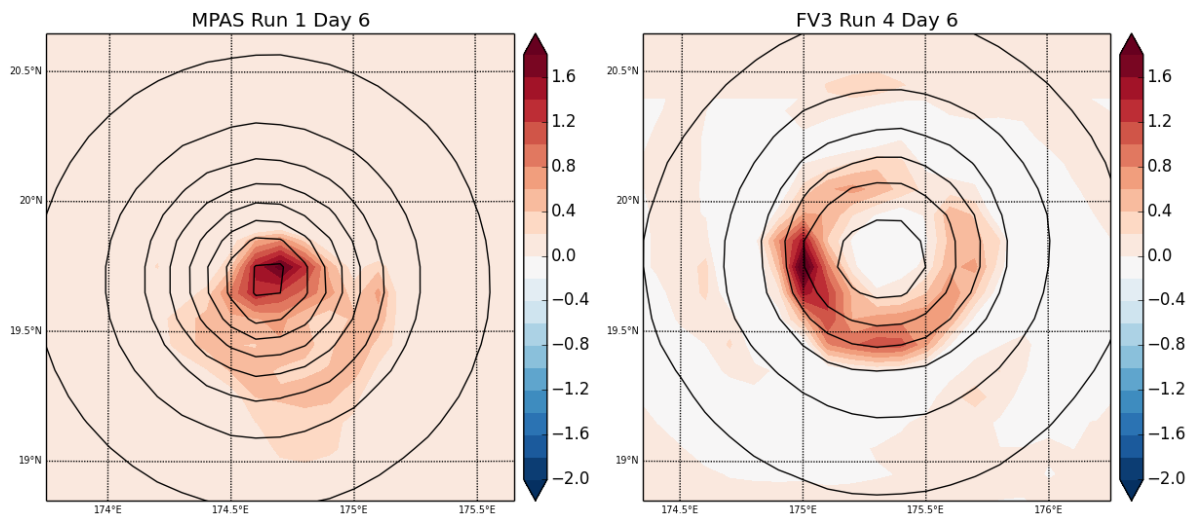


Figure 5.3: As above, except FV3 was re-run excluding the 2Δ vertical filter.

The primary difference between the MPAS and FV3 solutions shown in Figure 5.3 is the structure of the vertical velocity field. The FV3 solution shows a quasi-circular region of rising motion in the eye-wall region with weak subsidence in the center of the storm, consistent with observed tropical cyclones and previously published solutions from this test. The MPAS solution does not show a circular updraft – in fact the updraft appears to have the maximum near the center of the storm. The reason for the lack of subsidence in the center of the MPAS tropical cyclone is not clear. The MPAS tropical cyclone does exhibit a warm-core structure in the middle and upper troposphere (not shown), suggesting the presence of subsidence there at some point in the life-cycle. The fact that the MPAS tropical cyclones in the real-data forecasts do have quasi-circular updrafts (not shown) suggests that this feature of the solution may be specific to this idealized test framework.

Finally, note that the DTG did not run this test with the GFS for comparison, nor did the DTG have available a reference solution for each model that was converged with respect to horizontal and vertical resolution.

2.5.4 Variable-Resolution Tests

In Phase 1, two real-data forecast tests were performed at 3 km global resolution. Initial conditions from 18UTC October 24 2012 and 00UTC May 18 2013 were used. The 3-day forecast period included the early development of Hurricane Sandy and the Moore, Oklahoma severe weather outbreak. Forecasts were run with each group's own physics package. This made it difficult to attribute differences in the simulated moist convection in the forecasts to the differences in the dycores. Here both FV3 and MPAS are configured to use GFS physics, with the deep convective parameterization disabled. The same two test cases were run, but instead of running uniform 3 km global resolution, MPAS used variable-resolution mesh and FV3 used a combination of a stretched grid and a variable-resolution nest. This allowed us to examine the simulation of moist deep convection in a more controlled environment, while at the same time testing the variable-resolution capabilities of both systems. The detailed variable-resolution configurations used in both models are shown in detail in Figures 4 and 5 in the AVEC report (Appendix D). Both models were configured so that the resolution over and near the continental U.S. (CONUS) is close to 3 km, with up to 15 km resolution far away from the CONUS, and use the same

vertical level distribution as in the 10-day retrospective forecasts described for Criterion 3. FV3 was configured to call physics every 18 seconds in the 3 km nest. MPAS was configured to call physics every 18 seconds everywhere in the domain. FV3 used the same grid and nest in both cases, while MPAS centered the refined grid over Moore, Oklahoma and over in the vicinity of Hurricane Sandy in each case. A more detailed description of the GFDL nesting approach is given by Harris and Lin (2014). The variable-resolution MPAS approach is described by Park et al. (2014).

Figure 5.4 shows the forecast 500 hPa vertical velocity for 00UTC May 19, 2013, which is 24 hours into the forecast. The Moore, Oklahoma tornado occurred at 20UTC May 20, nearly three days into the forecast. Focus here on day 1 because both models produce strong updrafts in the same region in western Oklahoma, where severe weather was observed to occur (Figure 5.5). Although both models do produce strong updrafts on day 3 near the time of the Moore tornado, they are more widely separated from each other and the observed severe weather making a direct comparison more difficult. The character of the updrafts simulated by FV3 and MPAS is overall quite similar – the main difference is the larger and stronger FV3 updrafts. The larger scale of the FV3 updrafts is likely partly due to the fact that the resolution of the FV3 nest is slightly coarser than the high-resolution part MPAS variable-resolution grid mesh (Figure 5.6). However, it is also likely related to the general tendency of the FV3 dycore to produce slightly larger deep convective updrafts than MPAS for the same grid resolution and diffusion settings, as shown in the idealized supercell test (Figure 5.1). Comparisons of convective updrafts simulated at 00UTC May 20 and 21 by FV3 and MPAS show qualitatively similar differences – with FV3 producing somewhat larger and slightly stronger updrafts.

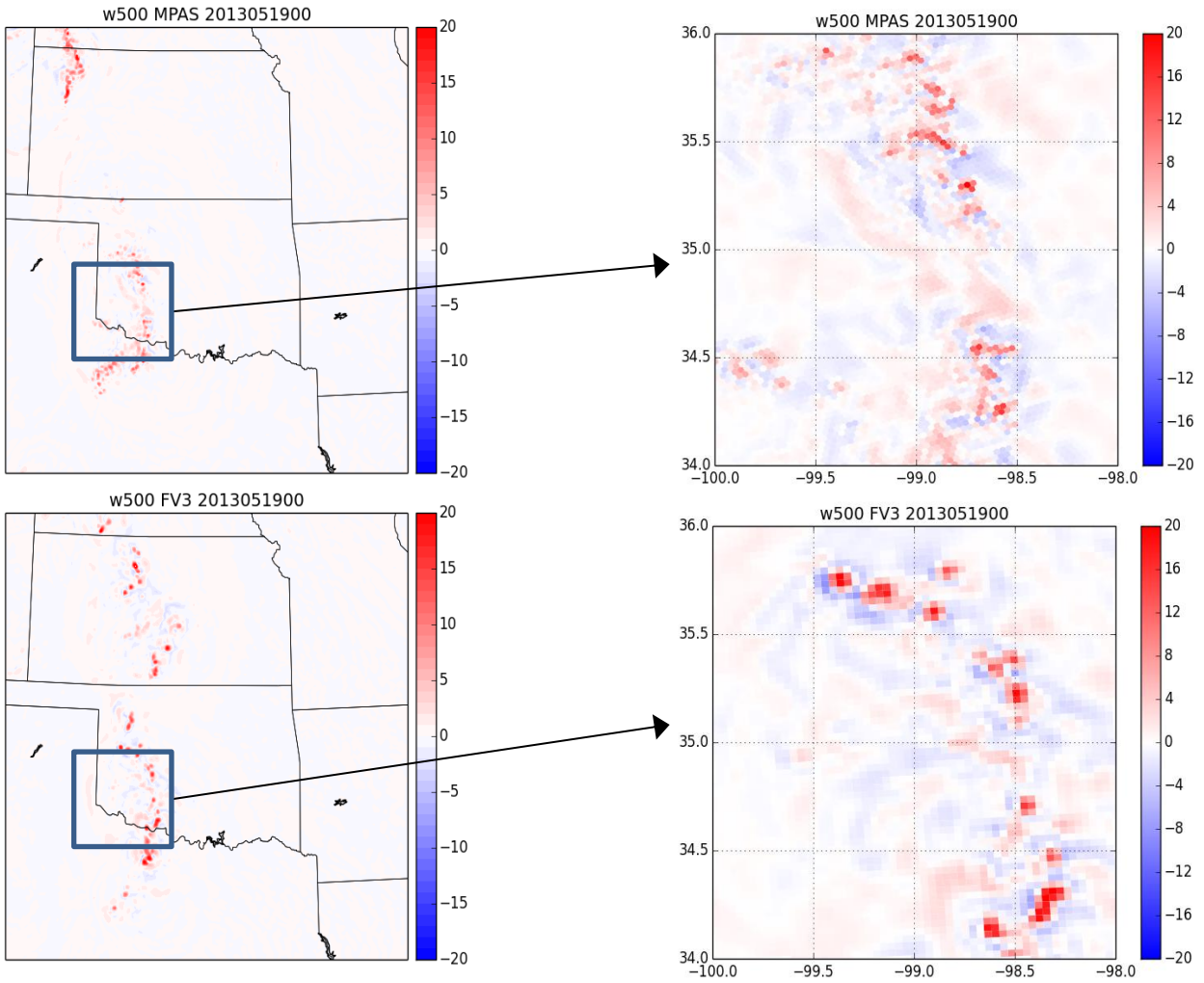


Figure 5.4: 24-h forecast 500 hPa vertical velocity (in meters per second) from the high-resolution region of the MPAS and FV3 forecasts. The plots on the right zoom-in on the inset box centered over western Oklahoma. The individual model grid cells are filled to avoid any smoothing and to illustrate the impact of the model grid on the simulated updrafts.

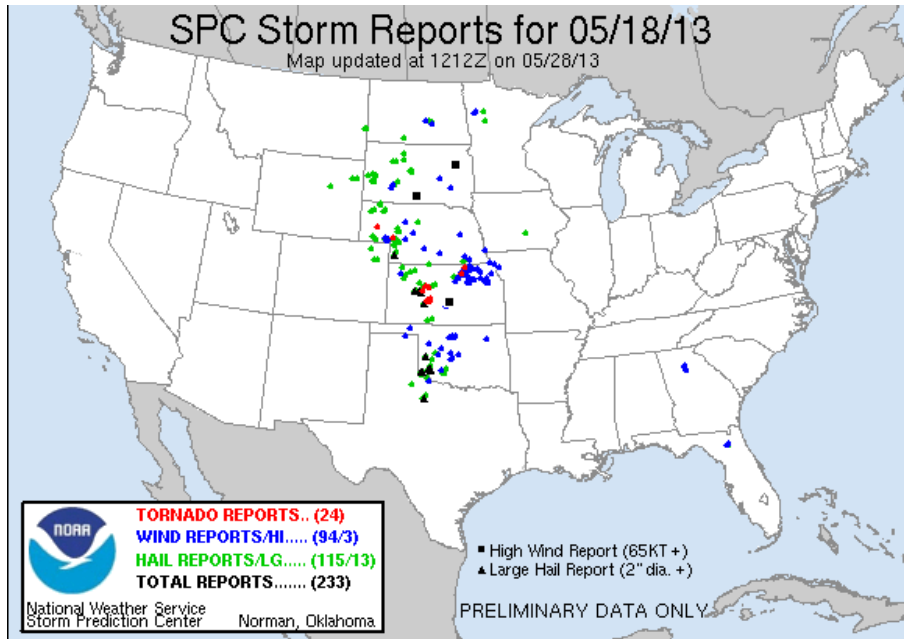


Figure 5.5: Storm Prediction Center storm reports for May 18, 2013. Note the cluster of high wind and large hail reports in western Oklahoma.

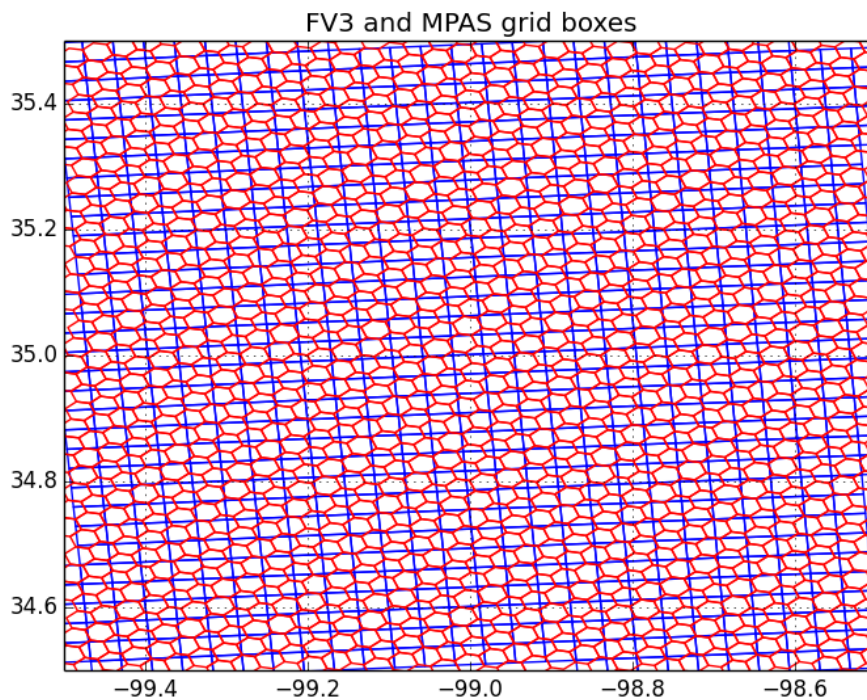


Figure 5.6: The individual grid cells for FV3 (blue) and MPAS (red) near the center of the zoomed-in domain shown in Figure 5.4.

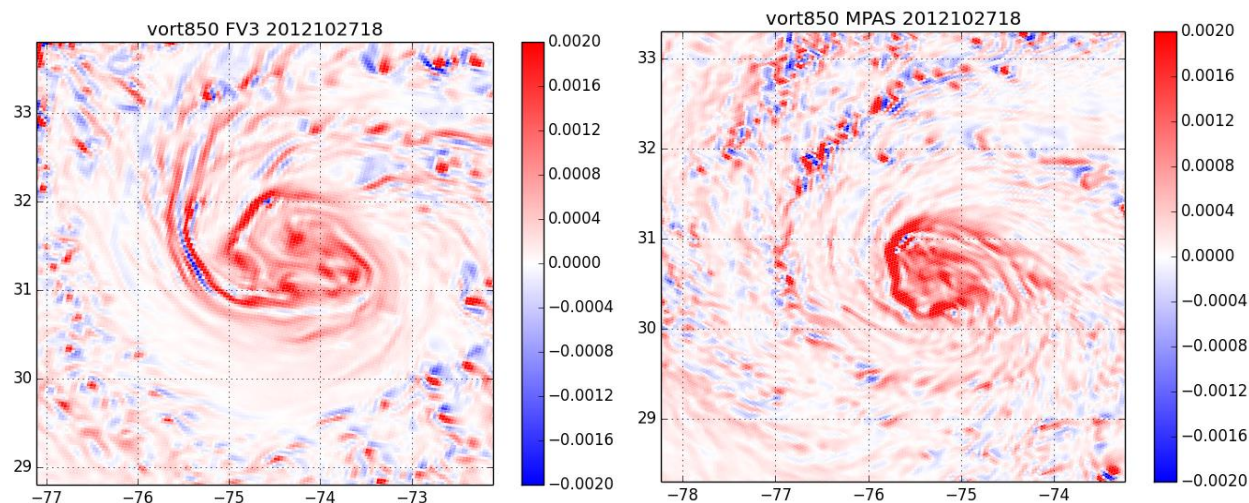


Figure 5.7: 72-h forecast 850 hPa vertical vorticity for the Hurricane Sandy case. The domain plotted is a 5-degree by 5-degree box center on the mean sea-level pressure minimum for each forecast. The MPAS storm is slightly farther south and west than the FV3 storm at this time.

Figure 5.7 shows 850 hPa vertical vorticity for a storm-centric domain at 18UTC 27 October 2012, 72-hours into the Hurricane Sandy forecast. The detailed structure of the simulated rain bands is evident in the vorticity field in both models. The main difference, as in the Moore case, appears to be the tendency for the FV3 convective structures to be slightly larger scale. This difference is apparent at all forecast times and in other forecast fields (such as vertically integrated cloud condensate and 500 hPa vertical velocity). The difference in the grid structures over the region of interest in the Hurricane Sandy case is very similar to that shown for the Moore tornado case (Figure 5.4).

The DTG enlisted the help of several subject matter experts to interpret the results of the variable-resolution test cases and the idealized supercell test. The results of that discussion are summarized in Appendix C. In short, the subject matter experts expressed the concern that the use of GFS microphysics, which was not designed to work well at these scales, limited the ability of either model to accurately simulate important aspects of observed convection, particularly downdrafts and cold-pools. However, even with this limitation, the DTG feels that the tests were useful in illustrating the ability of both dycores to simulate moist convection. The results indicate that the numerical formulation of either dycore does not preclude their future application to operational convective-scale forecasting. Furthermore, either the FV3 nesting capability or the MPAS variable-resolution mesh capability could form the basis for a unified global to convective-scale forecasting system. However, significant research and development work will need to be done regardless of which model is chosen to realize that vision. A key component of this R&D effort will be the development of advanced physics packages that work across scales, from global to convective.

2.5.5 Conclusions

Idealized supercell test - Both models produce very similar solutions for this test though the updrafts in FV3 appear to be a bit larger than for MPAS. Analytical solutions aren't available for this case so precise structure of the solutions isn't known. Both models appear to have converged by a resolution of 500 meters and as the resolution is decreased each model behaves similarly until the resolution becomes too coarse to effectively resolve the convective structure.

The DTG feels that both cores adequately simulate the convection in this test.

Idealized tropical cyclone test - The DTG is unsure how to interpret the results of this test. FV3 gave a solution similar to what was expected and similar to other simulations with other models of this test. MPAS did not. In particular, the MPAS solution had the updraft concentrated in the center of the vortex with no central downdraft as is normally expected for a tropical cyclone. Again as in the idealized supercell test, there is no analytic solution so the exact solution is unknown.

There was considerable debate within the DTG as to the cause of the difference but no conclusions were reached.

Variable-resolution tests - This was simply a test to see whether each of the models, MPAS or FV3, are capable of simulating real convective systems and two cases were chosen; the Moore, Oklahoma Tornado and Hurricane Sandy. Both models produced similar simulations of the convection in both cases. In the Moore Tornado case both models diverged from each other and from observations as the simulations went to longer forecast lead times, as is expected for the scales of interest here, but the DTG noted that the day 1 forecasts of the two models were qualitatively similar to each other and to observations at that forecast lead time. The main difference between the models was that the updrafts were somewhat larger and stronger for FV3 as compared to MPAS, which is similar to the results of the idealized supercell test. As noted, this difference may be related in part to a slightly coarser grid structure in FV3 compared to MPAS.

The DTG feels that this test shows that both models are suitable for forecasting convective permitting scales though a more suitable physics package for these scales would be required.

References

Klemp, J., W. Skamarock, and S.-H. Park, 2015: Idealized global non-hydrostatic atmospheric test cases on a reduced-radius sphere, *J. Adv. Model. Earth Syst.*, **7**, 1155–1177, doi: 10.1002/2015MS000435.

Reed, K. A. and C. Jablonowski, 2012: Idealized tropical cyclone simulations of intermediate complexity: A test case for AGCMs, *J. Adv. Model. Earth Syst.*, **4**, M04001, doi: 10.1029/2011MS000099.

Harris, L. M. and S. J. Lin, 2014: Global-to-Regional Nested Grid Climate Simulations in the GFDL High Resolution Atmospheric Model. *J. Climate* **27**, 4890–4910, doi: 10.1175/JCLI-D-13-00596.1.

Park, S.-H., Klemp, J. B., and W. C. Skamarock, 2014: A Comparison of Mesh Refinement in the Global MPAS-A and WRF Models Using an Idealized Normal-Mode Baroclinic Wave Simulation, *Mon. Wea. Rev.*, **142**, 3614-3634. doi: 10.1175/MWR-D-14-00004.1

2.6 Criterion 6: Stable, Conservative Long Integrations with Realistic Climate Statistics

2.6.1 Overview

Since it is anticipated that the NGGPS dycore will be used not only for shorter-term weather forecasts, but also for seasonal to inter-annual forecasting applications, it is important to make sure that the candidate dycores can be run stably for longer integrations, produce reasonable climate statistics, and conserve the appropriate invariants. To this end, a single 90-day simulation with a lower-resolution

version of the configuration used in the retrospective 10-day forecast tests (Criterion 3) was performed. The resolution was reduced from a nominal resolution of 13 km to a horizontal mesh of $192 \times 192 \times 6 = 221,184$ points for FV3 (nominally 52 km) and 256,002 point mesh for MPAS (nominally 48 km). These are resolutions similar to what is currently used for operational seasonal forecasting in NCEP operations. The number of vertical levels was kept at 64, the same as in the higher resolution 10-day forecasts. The simulation was started at 00UTC September 1 2015, from the GFS operational analysis. Surface conditions, including observed SST, sea ice, and seasonally-varying land surface conditions (such as vegetation fraction), were updated daily during the integrations. A reference run of the GFS was also performed at T382 resolution (nominally 52 km). Only the orographic gravity wave drag parameters in the GFS physics were changed to account for the lower resolution topography, using values suggested by NCEP EMC.

2.6.2 90-day Means

Figure 6.1 shows the 90-day, September-October-November (SON), mean precipitation rate (in mm/day) for the three simulations. The estimate from the Tropical Rainfall Measuring Mission (TRMM) 3B42 satellite product for the same period is shown for reference. The enhanced precipitation over the eastern Pacific associated with the El Niño event (and the associated suppression of precipitation over the western Pacific warm pool) is evident in all three model simulations. Both the suppression and enhancement appear to be overdone relative to the TRMM estimate, especially for the GFS. Overall, the character of the simulated precipitation is similar in all three models, likely because the influence of the GFS physics suite is overwhelming differences due to the dycore formulations.

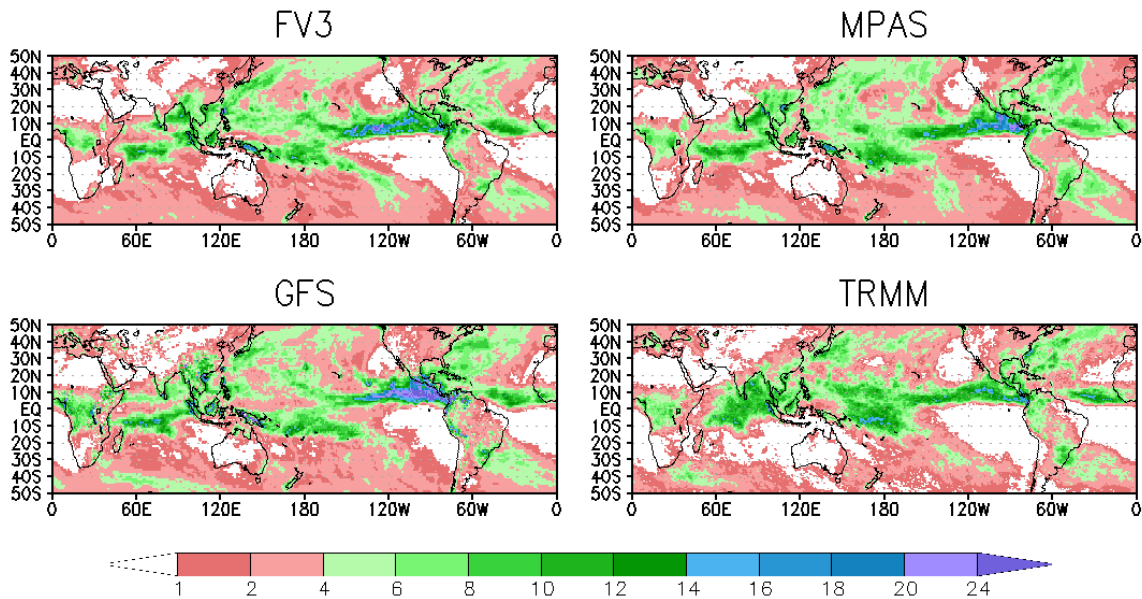


Figure 6.1: Seasonal Mean precipitation rate [mm day^{-1}] averaged for 1 Sept 2015 through 30 Nov 2015.

Figure 6.2 shows the 90-day zonal mean temperature difference with the European Center for Medium Range Weather Forecasting (ECMWF) ERA-Interim reanalysis. The FV3 model appears to have a warm bias in the upper-troposphere, where both the GFS and MPAS have a cold bias for this period, but all three models' biases are comparable to other state of the art climate models (Stevens et al. 2013). In addition, the differences in the temperature biases seen in the three models are of the same order as the differences that can result from simply changing the resolution and model top for a given dycore with a fixed physics package (Stevens et al. 2013, Figure 12). This suggests that there are some parameters in the GFS physics, perhaps those microphysics parameters that affect the distribution of upper-level clouds, sensitive to the dycore formulation and resolution and will need to be tuned differently for each dycore.

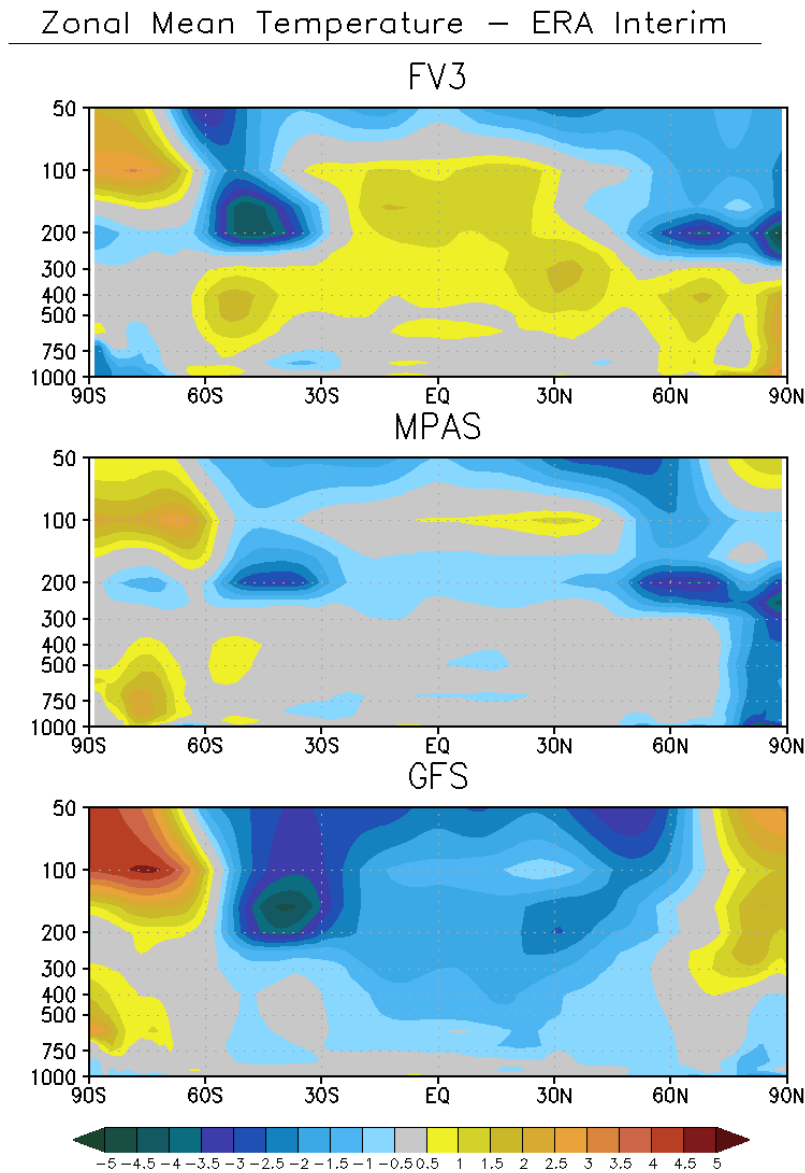


Figure 6.2: SON 2015 zonal mean temperature bias with respect to the ERA-Interim analysis [K].

Figure 6.3 shows the 300 hPa eddy geopotential height, showing the position of the stationary waves simulated by each model, compared to the ERA-Interim reanalysis for this period. All three models produce qualitatively similar stationary wave patterns that correspond well to those in ERA-Interim reanalysis.

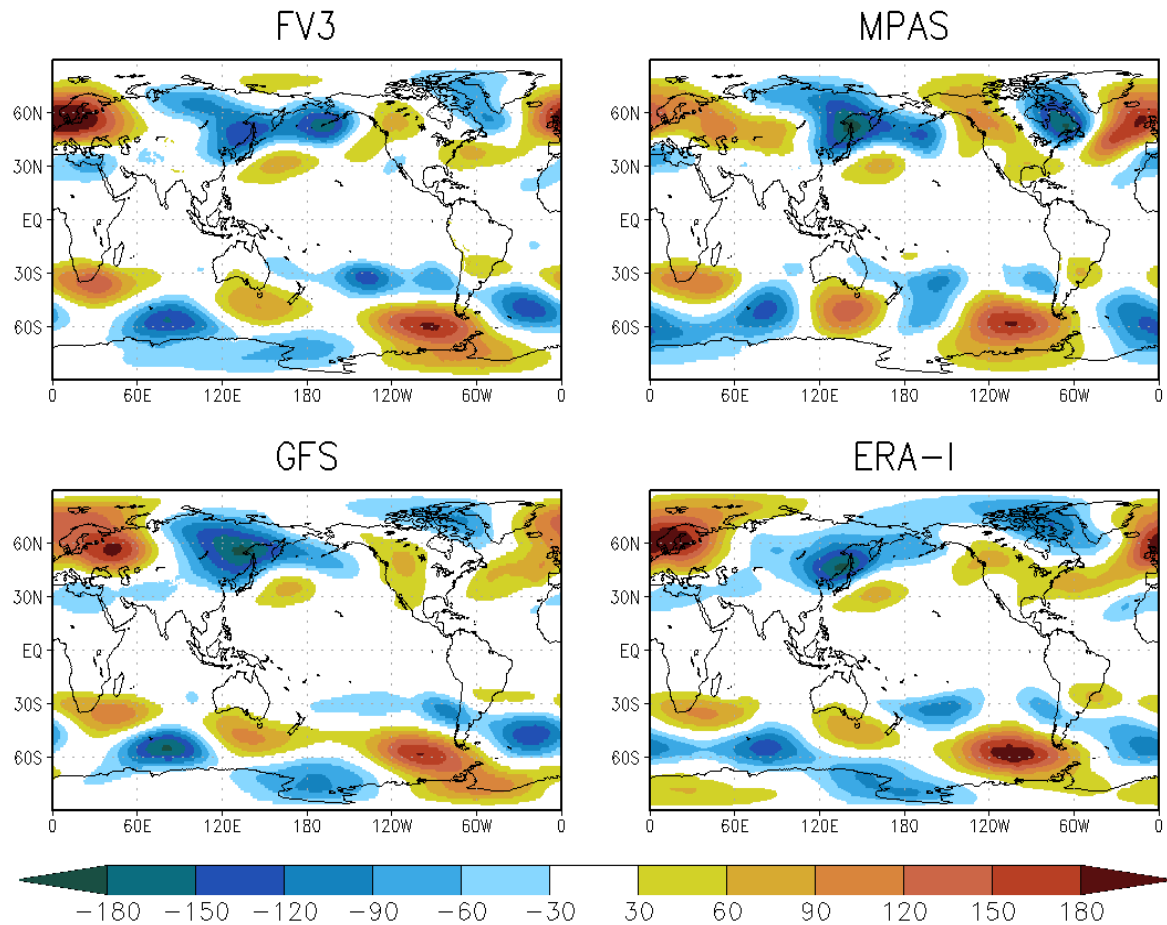


Figure 6.3: Eddy geopotential height at 300 hPa for SON 2015 [m]. Bottom right is ERA-Interim analysis.

Figure 6.4 shows the difference between the 90-day mean, 2m temperature simulated by the three models and the ERA-Interim reanalysis estimate for the same period. MPAS has a strong overall cold bias, particularly over sea ice and over North America. The FV3 and GFS 2 m temperature biases are similar, except over the Arctic, where FV3 has a warm bias and GFS has a cold bias. It is not clear to what extent these differences reflect differences in the dycores, differences in the way GFS physics was implemented, or simply reflect sampling uncertainty. The fact that the character of the near-surface temperature bias in MPAS is so different than FV3 or GFS does suggest that there may be problems in the implementation of the GFS land-surface scheme in MPAS.

2m Mean Temperature – ERA Interim

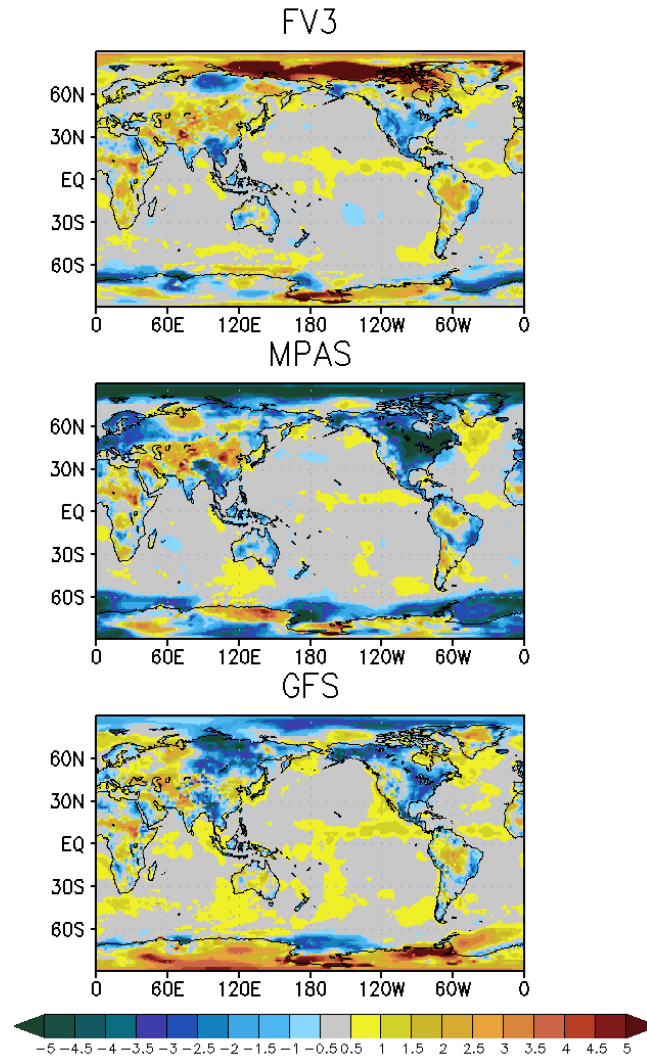


Figure 6.4: 2 m temperature bias with respect to ERA-Interim analysis for SON 2015 [K].

2.6.3 Conservation of Mass and Energy

Figure 6.5 shows the evolution of total energy in each of the three simulations. All three models show an energy decrease of about 0.6 percent over the 90-day period. The DTG does not expect energy to be exactly conserved by an atmospheric model forced by seasonally varying boundary conditions and solar forcing. All that can be said regarding energy conservation in this case is that each of the models is behaving similarly and there are no obvious outliers.

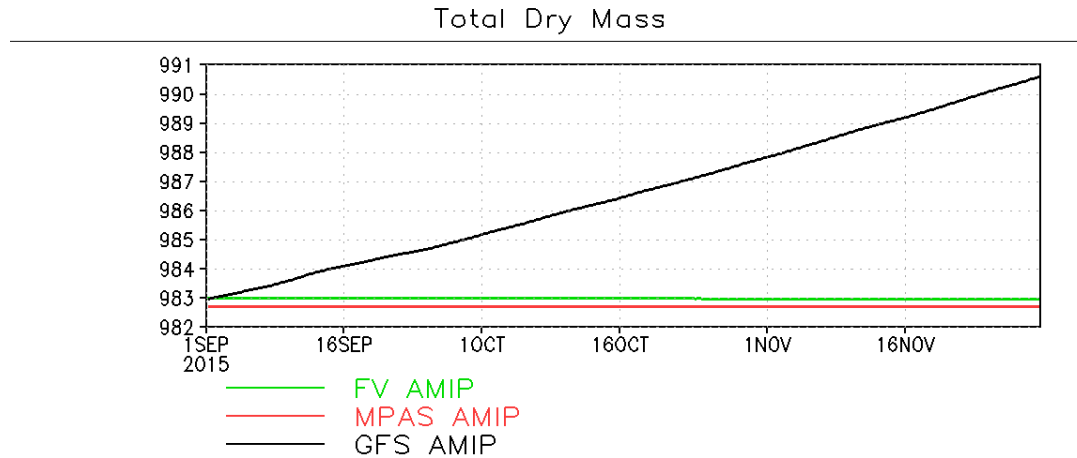


Figure 6.5: Total Dry mass [hPa] averaged globally for each 90-day run. Although it looks like FV3 is conserving, it is actually decreasing slightly. GFDL has acknowledged this bug is related to the interfacing with the GFS physics.

The dry mass of the atmosphere should be conserved exactly. The operational GFS has a ‘mass-fixer’ that sets the total dry mass back to the initial value by modifying the global mean surface pressure at each model time step. The DTG disabled the mass-fixer for this test. Figure 6.6 shows that without the mass fixer, the GFS increases the total dry mass by almost 1% over the 90-day period. FV3 and MPAS do a comparatively much better job at conserving dry mass, although FV3 loses some dry mass over the 90-day period whilst MPAS exactly conserves dry mass. The GFDL modeling team subsequently discovered a bug in the implementation of the GFS physics that accounts for this slight mass loss. Based on the results of the Criterion 2 idealized test, the DTG expects that if the physics is implemented correctly, both dycores should conserve dry mass.

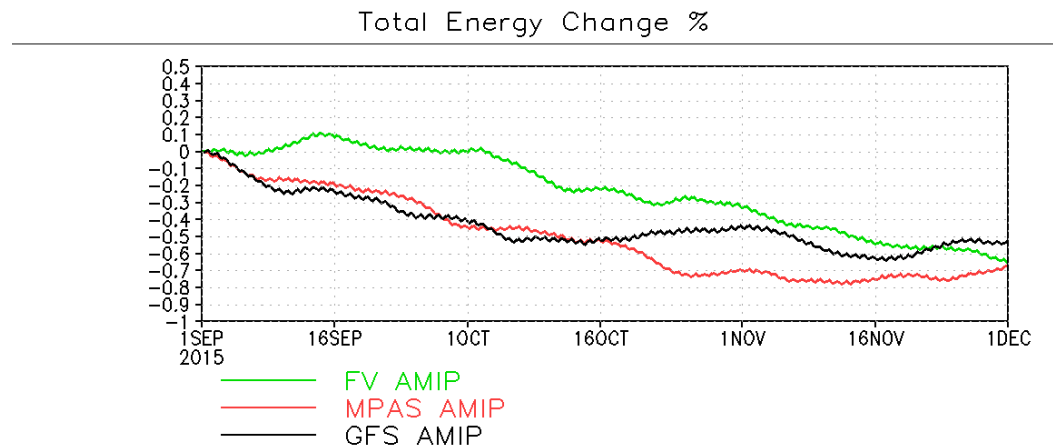
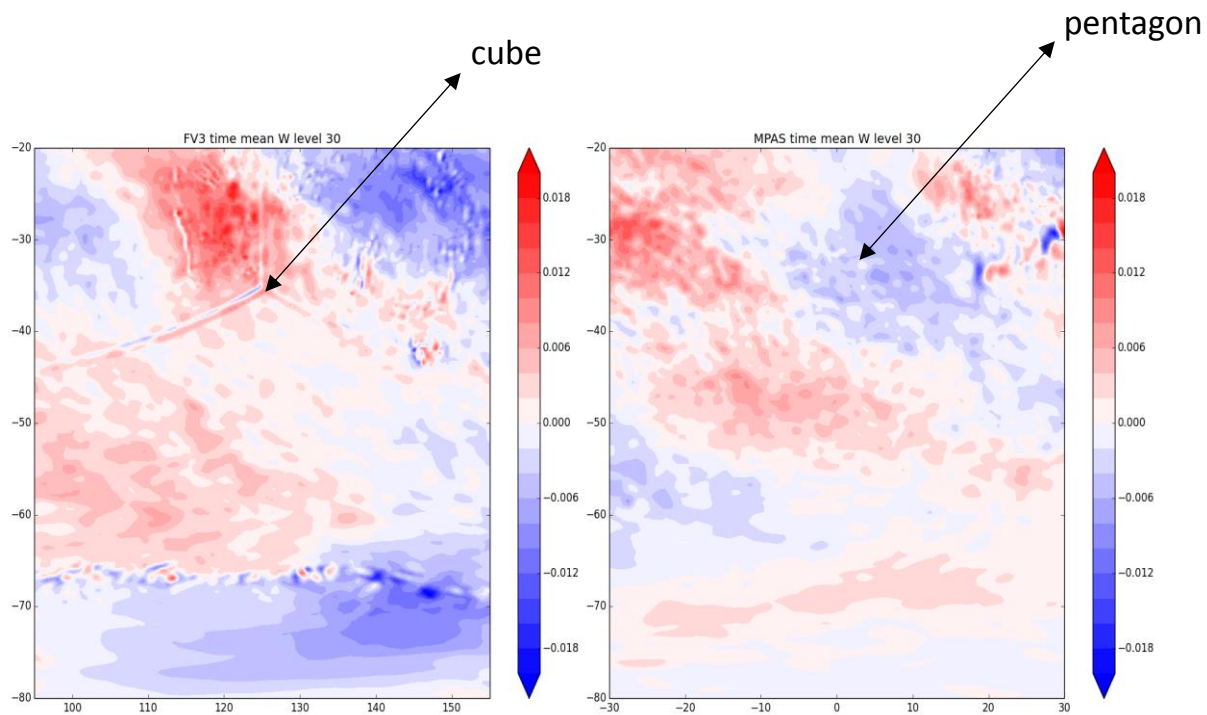


Figure 6.6: Total energy change from the initial conditions over 90 days.

2.6.4 Grid Imprinting

In the Criterion 2 idealized test, some evidence of grid imprinting was seen in the vertical velocity field. Spurious stationary vertical circulations along the corners of the cubes on the FV3 cubed sphere grid and at the pentagonal grid cells on the icosahedral MPAS mesh were evident, although they were very weak. Figures 6.7 and 6.8 show that these same circulations are evident in full-physics simulations, and as in the idealized case, they appear to be stronger and more widespread for FV3 than MPAS. However, these signals are still small $O(10^{-3})$ meters per second) compared to the day-to-day variability of vertical velocity, which is one to two orders of magnitude larger at this resolution. Some evidence of the grid imprinting signal can be seen in maps of time mean precipitation (not shown), but it is even harder to detect. These results suggest that some further work on reducing the variation of truncation error across cube boundaries may be warranted in FV3, particularly for long-term climate integrations, since the grid imprinting signal is stationary and shows up most prominently in longer means. The GFDL development team is aware of this and has been working on improvements to the dycore to reduce the



grid-imprinting signal.

Figure 6.7: 90-day mean vertical velocity for FV3 (left) and MPAS (right) at model level 30 (located near 300 hPa). The domain shown is different for each model.

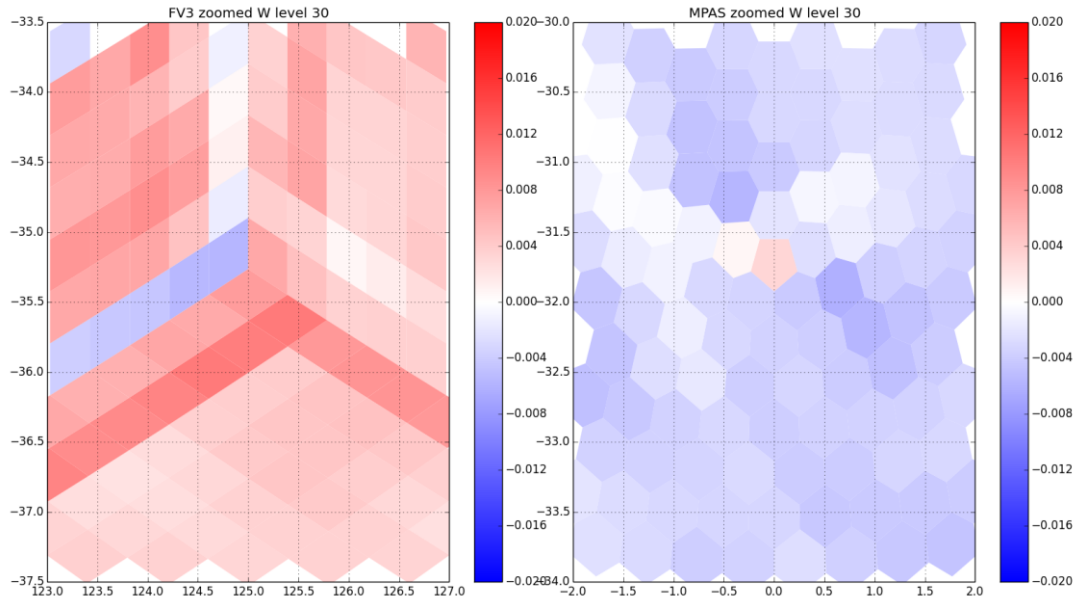


Figure 6.8: A zoom-in of the vertical velocity field shown in Figure 6.7, centered on the cube corner for FV3 and the pentagonal grid cell for MPAS.

2.6.5 Conclusions

90-day means - Both FV3 and MPAS produced similar long term means over the 90-day integration period, but there are differences.

FV3 and MPAS produced similar global precipitation rates and compared well with GFS, but all three models over-predicted the enhanced precipitation in the eastern Pacific and lowered precipitation rate over the western Pacific related to El Niño compared to TRMM estimates. It is likely that these differences are related to the GFS physics and not the dycores.

The zonal-mean temperature anomalies for all three models are again similar, but with notable differences particularly in the stratosphere in the polar regions. These differences are comparable to what is often seen when introducing a new physics package into a model without significant tuning. The DTG does not find these differences concerning for either FV3 or MPAS.

Finally, the 2 m temperature differences for all three models have considerable differences from observations and for each model they are different. In particular there is a warm bias for FV3 in the Arctic and a cold bias there for MPAS. For both models the bias amounts to several degrees, which is significant. It is not clear to what extent these differences reflect differences in the dycores, differences in the way GFS physics was implemented, or simply reflect sampling uncertainty. Either FV3 or MPAS would require work to identify the source of these biases.

While there are some issues noted for FV3 and MPAS from this test, the DTG sees no reason why each of these dycores would not be suitable for long-term integration forecasts.

Conservation of mass and energy - This test complements the conservation tests of Criterion 2. Whereas the Criterion 2 test only went out to 15 days for both models, this test goes out to 90 days. For

dry mass (which should be conserved exactly) both models conserve dry mass, MPAS exactly and FV3 with a slight error due to a known problem in implementing the GFS physics, which is being fixed.

All three models, FV3, MPAS and GFS lose total energy because of seasonally varying boundary conditions and solar forcing and all three lose at the same rate and approximately the same amount.

Even with these longer integrations, there are no serious issues related to conservation of mass and energy.

Grid imprinting - The grid imprinting described in Criterion 2 which was just for day 1 forecast lead time is also apparent in the 90-day averages from the Criterion 6 test. The same patterns as noted in Criterion 2 also appear in the averages for both FV3 and MPAS though as before the values are very small compared to synoptic scale vertical velocities. The grid imprinting in FV3 is considerably larger than it is for MPAS.

References

Stevens, B., et al. (2013), Atmospheric component of the MPI-M Earth System Model: ECHAM6, *J. Adv. Model. Earth Syst.*, 5, 146–172, doi: [10.1002/jame.20015](https://doi.org/10.1002/jame.20015).

2.7 Criterion 7: Code Adaptable to NEMS/ESMF

2.7.1 Overview

NEMS is the infrastructure underlying a coupled modeling system that supports predictions of Earth's environment at a range of time scales. NEMS has been in development at NCEP to streamline the interaction of operational analysis, forecast, and post-processing systems. NEMS is a shared, portable, high performance software superstructure and infrastructure and is built using the Earth System Modeling Framework (ESMF), <https://www.earthsystemcog.org/projects/esmf/>. ESMF provides utilities like generation of interpolation weights and utilities for calendar and time management as well as wrappers that create a standard component calling interface. The atmosphere component is the first component implemented in NEMS and can hold multiple atmospheric models. Currently, the Global Spectral Model (GSM) and the Non-hydrostatic Multiscale Model on B-grid (NMMB) are residing as sub-components under the atmosphere component. Besides atmosphere models, many applications based on Earth system components are currently under development in NEMS. Most of these components like the ocean models (the Hybrid Coordinate Ocean Model (HYCOM), the Modular Ocean Model (MOM)), the wave model (WAVEWATCH-III), and the sea ice model (Los Alamos Sea Ice Model (CICE)), are currently residing in NEMS as separate instantiations, coupled to other Earth system components using a National Unified Operational Prediction Capability (NUOPC) Mediator <https://www.earthsystemcog.org/projects/nuopc/>. The NUOPC mediator adds additional rules about how ESMF models interact, increasing their interoperability; and covers aspects from the level of build dependencies, such as to standardization of initialization phases, and standard names of the exchanged fields. All components of NEMS share a standardized I/O format (NEMSIO), which currently can handle binary data and GRIB1/GRIB2 data. NEMSIO functionality can be extended to handle many other data formats like NetCDF, HDF5 etc. NEMSIO has a serial version and a parallel version (using MPI-II parallel I/O). NEMSIO is more efficient in handling large files, and supports use of IO groups and quilt servers. A schematic of NEMS coupled system is shown in Figure 7.1.

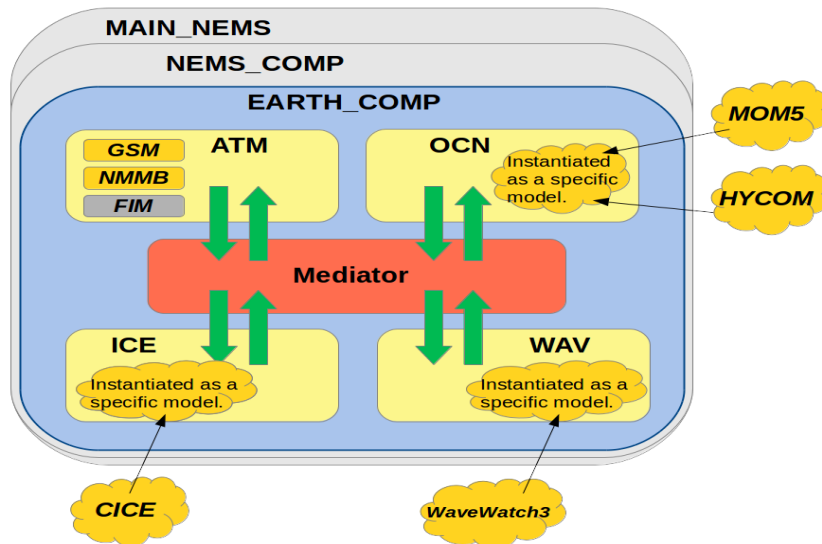


Figure 7.1: Schematic representation of NEMS Coupled System.

The atmosphere component of NEMS is being re-designed to separate the dynamics and physics sub-components to allow for accelerated development of physical parameterizations that are largely independent of the dycore itself. In this revised design, physics will interact with the dycore using a NUOPC interoperable physics driver and will allow for a clean implementation of the NGGPS selected dycore in NEMS. Figure 7.2 shows how, within NEMS, dynamics and physics of atmospheric model can interact using the physics driver. The Global Modeling Test Bed (GMTB) is currently developing a suite of physical parameterization schemes known as Community Common Physics Package (CCPP) starting with the GSM physics and is providing support for development and testing of advanced physics within NEMS.

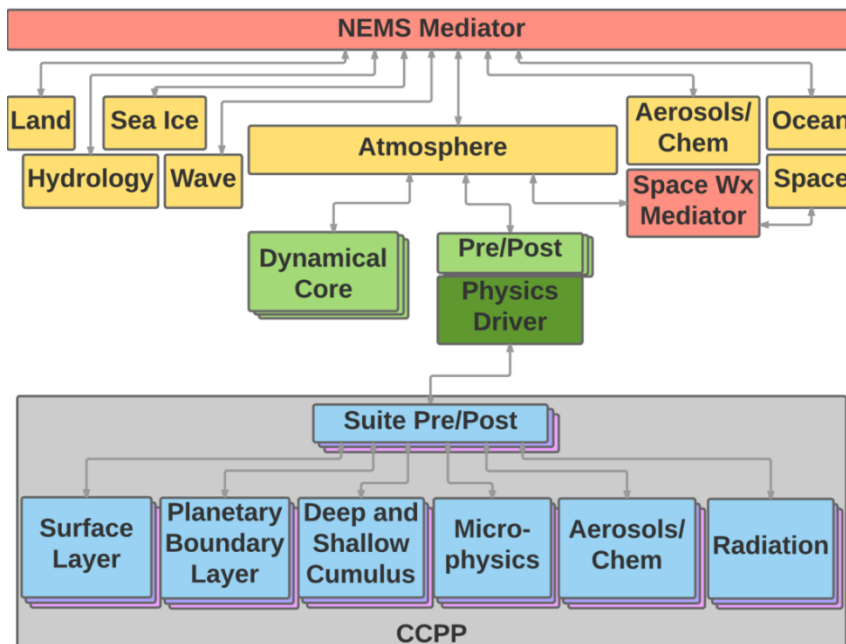


Figure 7.2: Schematic representation of dycore and physics interactions within NEMS using the interoperable physics driver.

In order to replace the GFS spectral dycore (GSM) with the NGGPS selected dycore, it is important to evaluate the differences between code structures of the candidate dycores and compatibility of the candidate modeling system software for NEMS, ESMF and NUOPC architectures. The DTG developed a detailed questionnaire and requested that the GFDL FV3 and NCAR MPAS groups provide subjective information on various aspects of their dycore code structure including:

- a) Interface with NEMS and ESMF, grid components and decomposition, import and export of required fields between various components
- b) Specification of control variables (run-time parameters)
- c) Interface between dynamics and physics
- d) Quilting and interpolation methods
- e) Input and output data formats and compatibility with NEMSIO
- f) Methods of compilation and build mechanism

Only the GFDL modeling group provided responses to all the questions. Dr. Mark Iredell, Software Engineering Team Lead at EMC, reviewed the responses from GFDL and sought some minor clarifications to the original responses. Dr. Iredell found the GFDL responses to be satisfactory, indicating no major issues or show-stoppers for implementing the FV3 dycore in the NEMS/ESMF infrastructure. The questionnaire and the material submitted by GFDL was collected by Dr. Vijay Tallapragada and archived at https://www.earthsystemcog.org/projects/dycore_test_group/NEMS_ESMF_Documentation and http://www.weather.gov/sti/stimodeling_nggps_dycoredocumentation. The DTG found that the submitted documentation from GFDL and evaluation from EMC was complete and sufficient for evaluation of Criterion 7 (code compatibility with NEMS/ESMF).

2.7.2 Conclusion

Responses from GFDL clearly illustrated the suitability of the FV3 dycore for implementation in the NEMS/ESMF superstructure. Since NCAR did not provide a response, it was not possible for the DTG to evaluate the compatibility of the MPAS modeling software for NEMS/ESMF.

Although the subjective evaluation of GFDL responses are adequate for the purposes of the Phase 2 selection process, much more in-depth technical evaluation would ultimately be required for the chosen dycore before implementing in the NEMS/ESMF.

2.8 Criterion 8: Detailed Dycore Documentation, including Documentation of Vertical Grid, Numerical Filters, Time-Integration Scheme and Variable-Resolution and/or Nesting Capabilities

In order to understand the differences between the candidate dycores, the DTG requested that each group 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

The material submitted by each group was collected by Dr. Richard Rood and archived at https://www.earthsystemcog.org/projects/dycore_test_group/Documentation_Introduction and is also available at http://www.weather.gov/sti/stimodeling_nggps_dycoredocumentation. The DTG found that the submitted documentation was suitable for the purposes of the Phase 2 selection process, but that much more documentation, both scientific and user-oriented, would ultimately be required for the chosen dycore.

One important difference between the FV3 and MPAS dycores is the vertical coordinate. The documentation describing the design and implementation of the vertical coordinate systems in the two dycores was analyzed by DTG consultants and their analysis is presented in Appendix F (“Vertical Coordinate Analysis”).

2.9 Criterion 9: Evaluation of Performance in Cycled Data Assimilation

2.9.1 Overview

Confronting a model with observations by running it within a data assimilation system poses a significant challenge to the dycore. Analysis produced by assimilation of observations may destroy some of the model balance, resulting in unbalanced motions when model is integrated from the analysis as the initial condition. These unbalanced motions can potentially be amplified by the dycore numerics and can feedback on the analysis when the background-error covariances are estimated from a forecast ensemble. To maximize the potential for feedback, the ensemble-based NCEP operational data assimilation system is used. Typically, the background-error covariance estimate (which is used in combination with observation error covariances to weight information coming from new observations relative to the background forecast) is a blend of an ensemble-based estimate and a static, offline estimate. In operations, the weight given to the static estimate is 12.5%, here the weight was set to zero so the background error covariance estimate is coming purely from an 80-member ensemble initialized from the previous analysis. Following is a list of differences between the 3D hybrid ensemble-variational system that was operational until May 2016 and the test version used here.

- The resolution of the 80-member ensemble is nominally 50 km (the same resolution used in the Criterion 6 test). The operational version uses a T574 ensemble (nominally 35 km).
- No static background-error covariance component is used (12.5% is used in the operational system).
- No special methods are used to control gravity wave noise (the operational system is using a digital filter finalization step and a tangent-linear normal mode balance constraint on analysis increments).
- No parameterization of model uncertainty is included in the ensemble forecast (the operational model uses stochastic parameterizations of model uncertainty). Instead, the background error covariance is increased using multiplicative inflation to account for model uncertainty, using a relaxation-to-prior spread (Whitaker and Hamill 2012) coefficient of 1.1 (a value of 0.85 is used in operations).
- No high-resolution single control forecast is included. Instead, the variational solver uses the ensemble mean forecast as the background forecast. The operational system uses covariances estimated from a T574 ensemble to update both the T574 ensemble and a T1534 control forecast.

Aside from these differences, the test system is nearly identical to the operational system prior to May 2016, including assimilation of all the satellite radiance observations used in operations. In May 2016, the operational system was upgraded to include a 4D ensemble-variational algorithm and assimilation of cloud-affected radiances.

2.9.2 Fit of First-Guess Forecasts to Observations

Three experiments were run, one with FV3, one with MPAS and a baseline run of the GFS at T382 resolution. The experiments were started from the operational GFS analyses on 1 September 2015 at 00UTC and were run through 29 September 2015 at 00UTC. Figure 9.1 shows the RMS difference between the ensemble mean first-guess forecast and all assimilated in-situ observations starting from 5 September at 00UTC and ending at 29 September at 00UTC, as a function of pressure. Although the quality control decisions are different in each experiment, the total numbers of observations going into the calculation at each pressure level for each experiment are very close (differences are less than 5 percent). The results show that the FV3 forecasts fit the data better than MPAS and the baseline GFS configuration forecasts. This suggests that the skill of the FV3 10-day retrospective forecasts would improve if the FV3 model were initialized from its own analyses, generated from a fully-cycled DA system – and likely would perform better than the operational GFS. The relatively poor performance of MPAS is consistent with the skill assessment of the 10-day retrospective forecasts discussed in the Criterion 3 section.

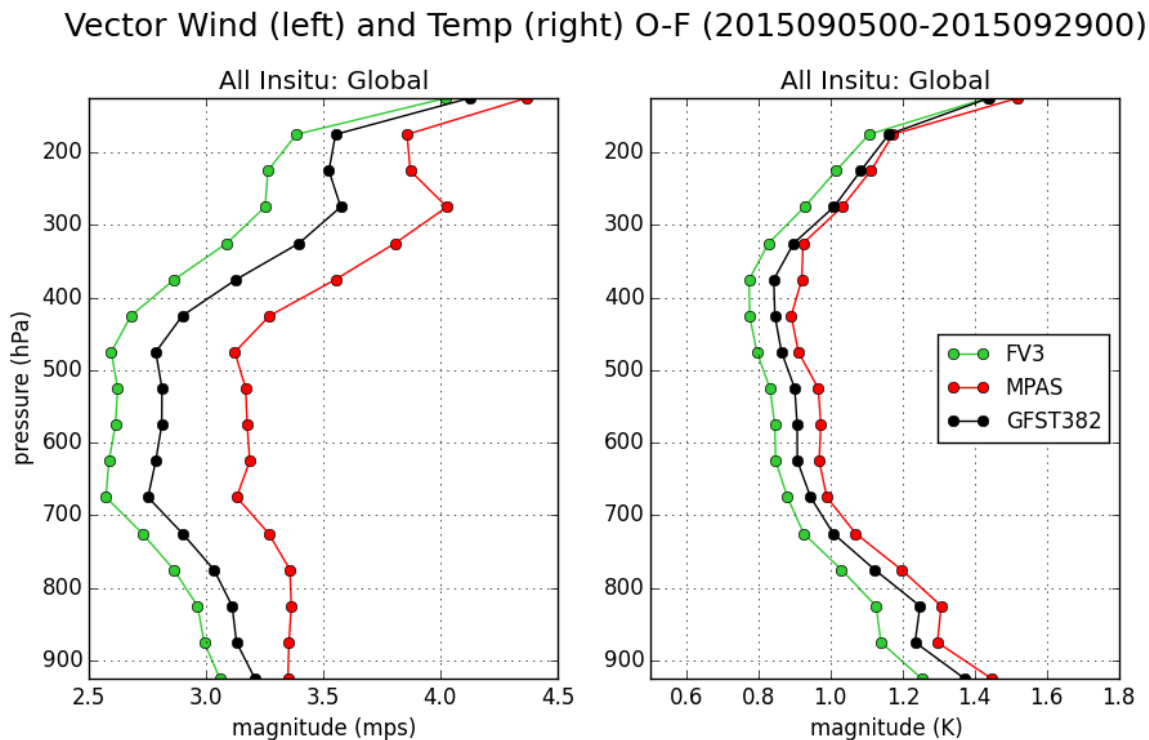


Figure 9.1: RMS difference between ensemble first-guess forecasts from FV3, MPAS and a T382 version of the GFS, and all in-situ wind (left) and temperature (right) observations as a function of pressure.

Figure 9.2 compares the FV3 RMS innovations (observations – forecast) for FV3 and the operational T1534 GFS. The operational GFS fits the temperature observations and low-level wind observations slightly better than FV3, but FV3 fits the wind observations above the boundary layer slightly better than the GFS. This is despite the fact that the FV3 experiment was run at approximately four times lower nominal resolution than the operational GFS, with a modified version of the data assimilation system missing some features that are used in operations (as described in section 2.9.1).

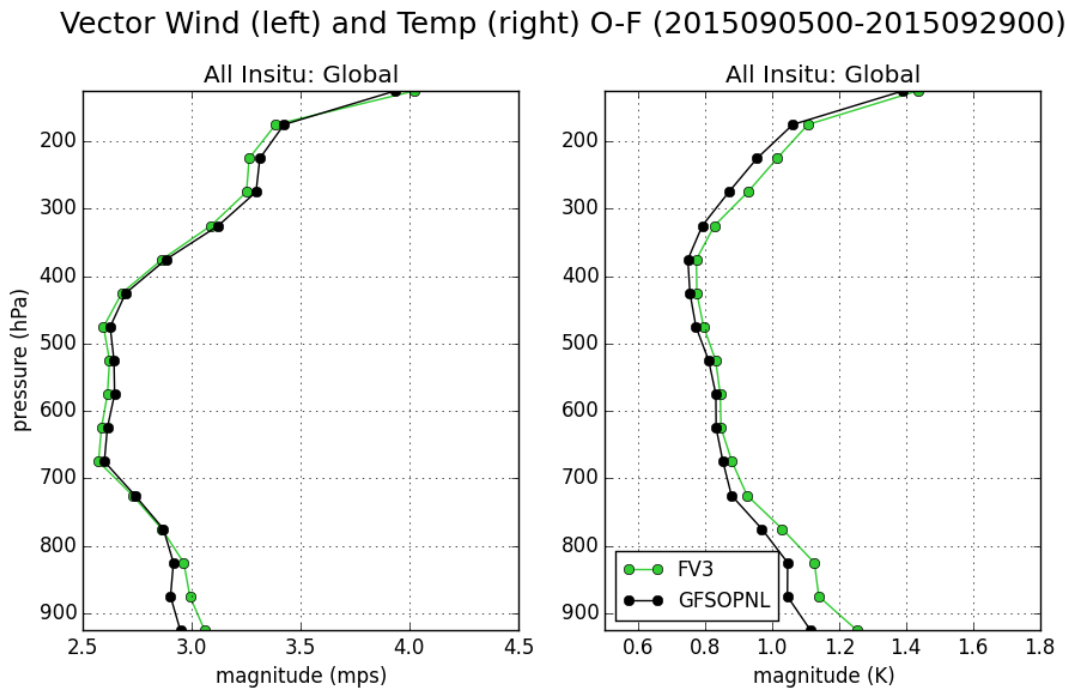


Figure 9.2: RMS difference between ensemble first-guess forecasts from FV3 (at 50 km resolution) and the operational high-resolution GFS (at 13 km resolution), and all in-situ wind (left) and temperature (right) observations as a function of pressure.

2.9.3 Model-Space Verifications Relative to ECMWF Analyses

Figure 9.3 shows the standard deviation of the difference between the 6-hour FV3, MPAS and T382 GFS ensemble mean forecasts and the ECMWF analyses of surface pressure, for the same period as shown in Figures 9.1 and 9.2. The time mean surface pressure difference is removed, since it primarily reflects differences between the ECMWF orography and the NCEP orography. The FV3 forecast surface pressure tracks the ECMWF analyzed surface pressure better than either MPAS or the T382 GFS.

Figure 9.4 shows the RMS analysis increment of surface pressure, as well as the ensemble spread. The FV3 analysis increment is smaller, indicating the data assimilation step has to make smaller corrections to the first-guess surface pressure field. This could be either because the FV3 forecasts are closer to the observations, or because the data assimilation system ‘trusts’ the FV3 forecasts more (relative to the GFS and MPAS forecasts). The fact that the ensemble spread (which the data assimilation system uses as an estimate of forecast error) is nearly identical for all three experiments suggests that the former is more likely.

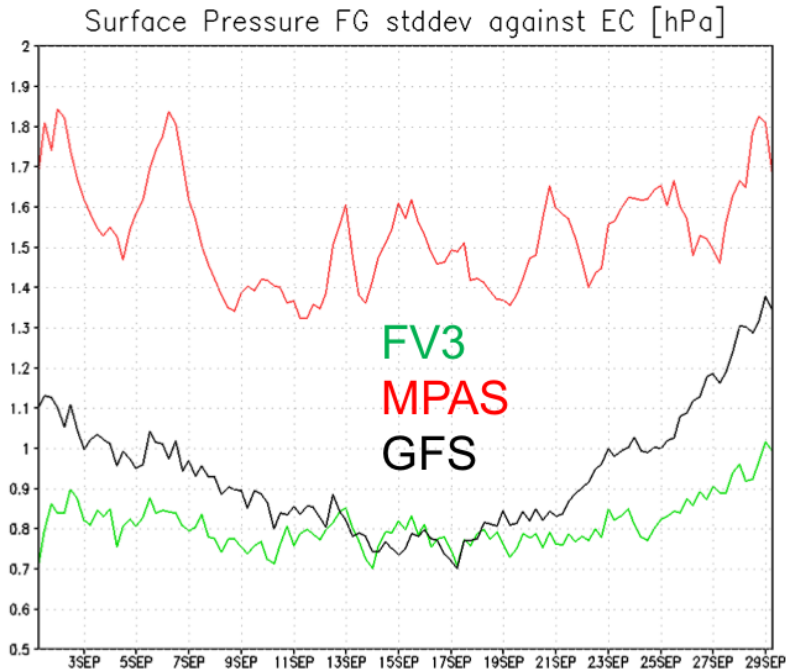


Figure 9.3: Standard deviation of the difference between ensemble-mean first guess forecasts of surface pressure and ECMWF analyses, with the time mean removed and averaged over the globe. Units are hPa.

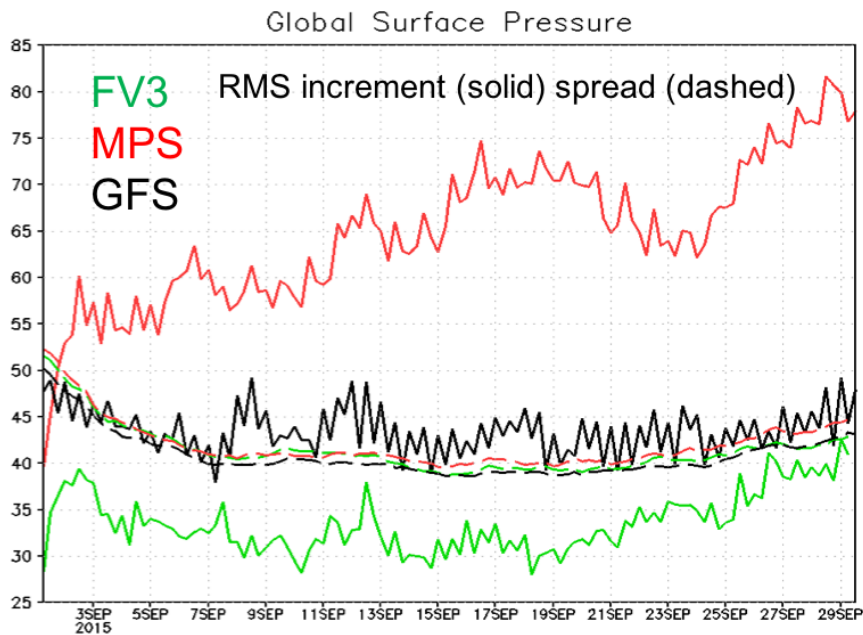


Figure 9.4: Global mean RMS analysis increments (solid) and ensemble spread (dashed) for surface pressure. Units are Pa.

Figure 9.5 shows the time mean analysis increment for surface pressure. If the forecast model were unbiased, the time mean increment should be small. The FV3 and GFS mean increments are much smaller than the RMS increments, but the MPAS time mean increments are large by comparison. This suggests that MPAS forecasts have a significant systematic bias in surface pressure, which is not present in the other two models.

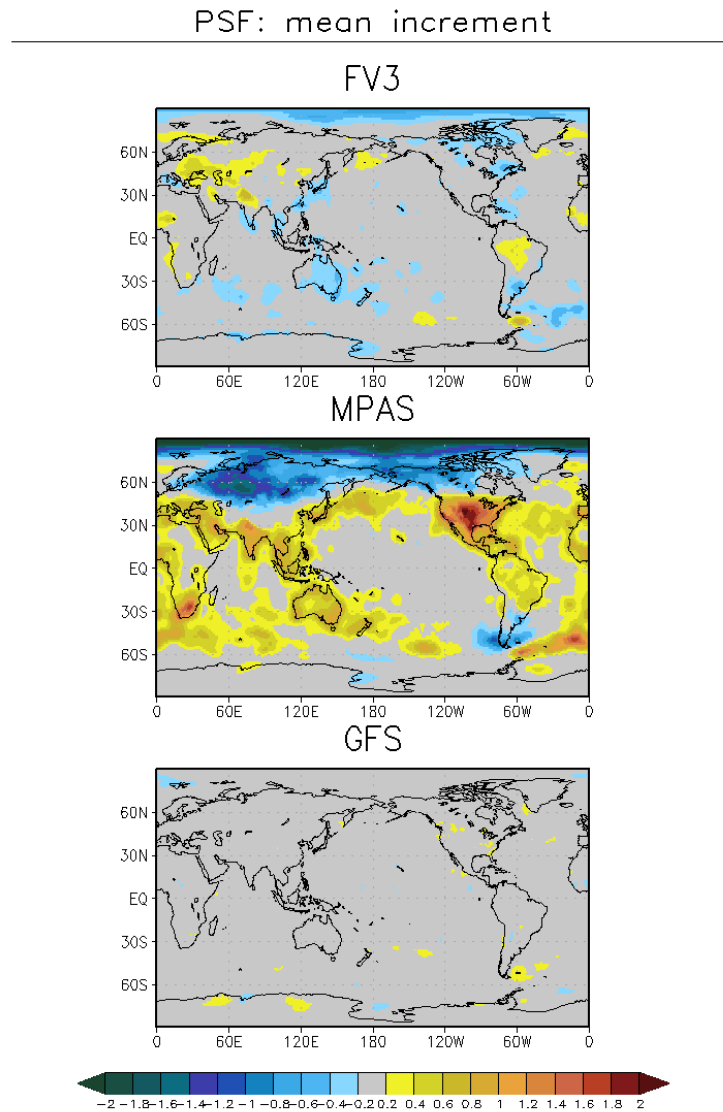


Figure 9.5: Time mean surface pressure analysis increments for FV3 (top), MPAS (middle) and the T382 GFS (bottom). Units are hPa.

2.9.4 Conclusions

Fit of first-guess forecasts to observations - In this test the initial conditions of each model were defined from its own data assimilation cycle rather than using the GFS initial conditions as in Criterion 3. Comparison of these six-hour forecasts with observations shows that FV3 is superior to both MPAS and a reference run of the GFS at reduced resolution. This suggests that if FV3 were initialized from its own

analyses, generated from a fully-cycled DA system (rather than the GFS analysis as in Criterion 3) it would likely perform better than the operational GFS.

The current operational GFS (13 km) and the FV3 (50 km) first-guess forecasts both fit the temperature and wind observations well, with the operational GFS fitting the temperature observations better, and the FV3 fitting the wind observations above the boundary layer better. This is despite the fact that the FV3 experiment was run at approximately four times lower nominal resolution than the operational GFS and with an assimilation system missing some features that were used in operations.

Model-space verifications relative to ECMWF analyses - Comparing the ensemble first guess (6-hour forecasts) for MPAS, FV3 and GFS (50 km) to the ECMWF analysis shows that FV3 does much better than MPAS and somewhat better than the GFS. In addition, FV3 requires less adjustment to the first-guess forecasts than both GFS and MPAS, and there is considerably less surface pressure bias in FV3 forecasts relative to MPAS.

The overall conclusion from Criterion 9 is that FV3 using its own data assimilation system is likely to provide superior forecasts than the current GFS operational system. Significantly more work will be required to improve the data assimilation performance of MPAS to be on par with the current operational GFS.

References

Whitaker, J. S., and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, **140**, 3078-3089. <http://dx.doi.org/10.1175/MWR-D-11-00276.1>

2.10 Criterion 10: Implementation Plan (Including Costs)

2.10.1 Overview

The Global Climate and Weather Modeling Branch (GCWMB) of NCEP EMC is ultimately responsible for transitioning the NGGPS chosen dycore into operations. The procedures for implementing a new model (or upgrades to an existing model) involve various stages of development, testing, evaluation, and optimizing the end-to-end system for real-time operations. Computer and human resources are two major factors that determine projected costs for implementing the GFS with the NGGPS dycore in operations. The DTG requested EMC provide estimates of timelines, computational and human resource requirements for implementing GFS with the NGGPS chosen dycore in operations, with a configuration of the model that matches or exceeds the GFS in operations at the time of transition.

The implementation plan for replacing the Global Spectral Model with the new dycore is designed based on the information provided by the DTG on Criteria 3 and 4 (robust model solutions under a wide range of realistic atmospheric initial conditions using a common (GFS) physics package; and computational performance with GFS physics). The assessment by the DTG for these two criteria was used as a subjective measure of the new dycore's scientific and computational readiness and was factored in estimating the costs for implementation. Figures 10.1 and 10.2 show projected timelines for implementing the FV3 and MPAS dycores in operations.

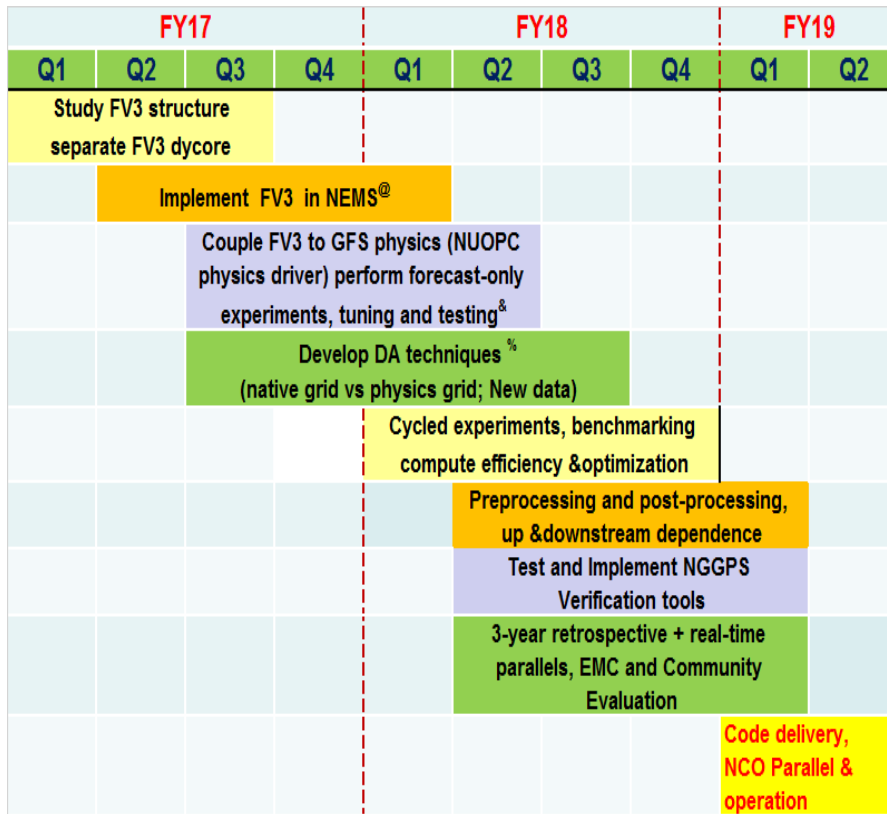


Figure 10.1: Tentative implementation timelines for GFDL FV3 dycore in operations at NCEP.

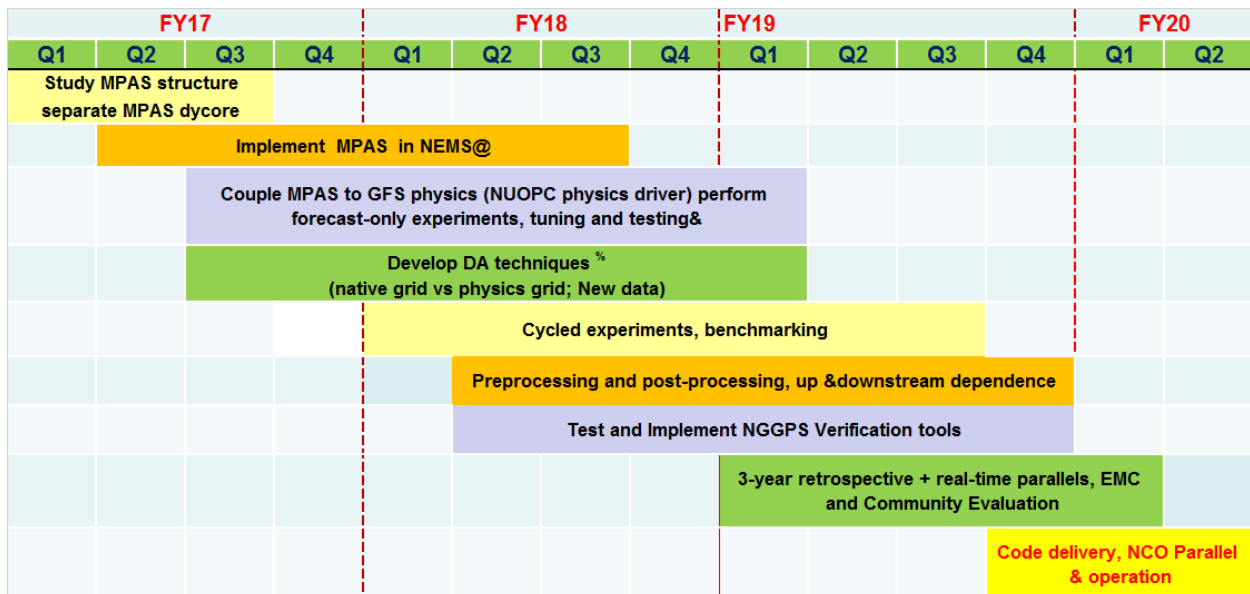


Figure 10.2: Tentative implementation timelines for NCAR MPAS dycore in operations at NCEP.

2.10.2 Implementation Tasks

The new dycore implementation tasks are divided broadly into nine categories, many of which have overlapping timelines, but are dependent on each other for integrating the full end-to-end system for transition into operations:

- 1) **Dycore integration into NEMS:** This is a prerequisite for implementing the NGGPS chosen dycore in operations. A detailed evaluation of NEMS/ESMF readiness of candidate dycores was requested by the DTG (see Criterion 7), however, only the GFDL modeling group has provided the requested information. Based on this information, EMC estimated that it will take about 6 to 9 months for initial implementation of the GFDL FV3 dycore in NEMS (with continuous enhancements afterwards), while the MPAS dycore might take about 12-15 months of effort.
- 2) **Implementation of the physics interface for the dycore in NEMS:** The next task is for developing and implementing the appropriate physics interface for the dycore coupled to GFS physics using the NUOPC Physics Driver in NEMS. Preliminary assessment based on Criterion 3 suggested that the GFS physics was accurately implemented in FV3 while this was not demonstrated for MPAS. Both dycore groups have experimented with the interoperable physics driver and GFS physics. EMC estimated that the amount of work needed to implement GFS physics with the MPAS dycore in NEMS would take 18-21 months compared to about 9-12 months for the FV3 dycore.
- 3) **Integration of data assimilation:** This is by far the most critical and daunting task. Both dycores were tested with limited self-cycled data assimilation systems during the DTG Phase 2 experiment period. While Criterion 9 focused on testing a sub-optimal configuration of FV3 and MPAS with cycled data assimilation similar to that of the current operational Global Data Assimilation System (GDAS) using 4-D Hybrid EnKF-Variational technique, additional development is necessary to satisfy operational needs. EMC anticipates that the GSI based cycled 4D-EnVAR hybrid DA system would need to account for the staggered grid of the FV3 (winds and scalar quantities at different positions), the non-orthogonal vertical coordinate (different for dynamics and physics due to mass adjustment), non-hydrostatic dynamics, physics grid vs. native grid considerations, and ensemble configurations (EnKF) to include stochastic physics. EMC did not extensively evaluate the needs for the MPAS dycore, however, EMC expects similar emphasis will need to be given for treating the vertical coordinate, treatment of analysis grids vs. model/physics grids and stochastic physics for ensembles used in the DA system. Taking into account the assessment provided by the DTG on Criterion 9, FV3 has shown a better fit to observations and better analysis fields compared to MPAS, EMC expects about 18-21 months of development work for MPAS compared to about 9-12 months for FV3 for the DA integration task. This task could overlap with the physics task described earlier.
- 4) **Downstream and upstream dependencies:** GFS is the foundational forecast guidance system that influences majority of the other modeling systems run operationally in the NCEP production suite. Figure 10.3 illustrates upstream and downstream dependencies of GFS on various production suite elements. The technical and scientific evaluation of the impact of the new dycore based GFS on these dependencies is critical, and it will require working with developers of all production suite applications that are dependent on GFS in some form. One example is the impact of GFS on regional hurricane models like Hurricane Weather Research and Forecasting model (HWRF) and GFDL models that use GFS initial and boundary conditions for

- 6) **Verification, validation and visualization:** This is a non-trivial task that will involve developing/modifying verification and diagnostic tools that are used in evaluation of operational and parallel GFS extended for the new dycore based GFS. EMC estimated about 9-12 months of work for full integration of end-to-end system that includes verification, diagnostics and visualization of model outputs. There will be continuous need for adjusting these tools based on the progress made with rest of the tasks, hence the extended timelines for MPAS compared to FV3.
- 7) **Benchmarking, testing and evaluation:** Scientific evaluation and validation of forecast skill improvements of the GFS with the new dycore is by far the most laborious task in preparing a new model or model upgrades for transition to operations. While EMC has well-established procedures to conduct these tests systematically, it will still require careful evaluation of scientific results by the model developers in collaboration with various partners and stakeholders. For a major model upgrade, GFS requires three years of retrospective forecasts combined with at least one year of real-time parallel runs with frozen model codes and configuration. Furthermore, added significant value must be demonstrated when compared to the operational model at that time. This process is generally independent of the choice of the dycore, however, based on the results available so far from FV3 and MPAS as evaluated by the DTG, MPAS has much lower forecast skill compared to FV3 or operational GFS (Criterion 3), EMC estimated a much longer period of testing for MPAS when compared to FV3. This will also result in requiring additional computational and human resources to conduct more experiments for MPAS.
- 8) **Workflow, code optimization and fine tuning:** Operational requirements for implementing model upgrades include code optimization for operational HPC, integrating into operational workflow, and fine tuning the system for robustness, accuracy and on-time delivery of forecast products. Based on the AVEC evaluation of computational performance of FV3 and MPAS (Criterion 4), EMC expects that substantially larger software engineering effort will be required to optimize computational efficiency of MPAS than that of FV3. Integrating the entire modeling system into a workflow framework that is consistent with current operational and parallel GFS setup would also require optimizing the codes to fit into the allocated computational resources. As such, EMC expects about 12-18 months of effort for MPAS compared to about 6-9 months for FV3.
- 9) **Operational implementation:** The final step in the implementation plan is to transition the entire modeling system in its final form to NCEP Central Operations (NCO), who is responsible for running the model in operations in real-time. Transition to operations will involve several stages of interactions of model developers and code managers with NCO to help NCO setup and evaluate the 30-day parallel runs for IT stability, reliability, data flow, upstream and downstream dependencies, product generation and distribution, and robustness of the model solutions. This process is independent of the dycore and would take about 4-5 months from code hand-off to final operational implementation. However, EMC is of the opinion that there will be additional work required for MPAS by NCO to fit into operational resources on the production machine.

2.10.3 Human and Computational Resource (Implementation Cost) Requirements

EMC has estimated the anticipated level of effort both in terms of human capital and compute resources to complete the model development, conduct testing and evaluation, and meet the implementation timelines based on the analysis of various tasks described in the previous section. Table 10.1 summarizes the subjective and rather conservative estimates of additional resources required at NCEP EMC to implement the NGGPS chosen dycore into operations. It is anticipated that GFS, with the FV3 dycore, would be implementable in operations by the second quarter of 2019, while implementation would take an additional year for GFS with the MPAS dycore (see Figures. 10.1 and 10.2). The level of effort required is based on the information made available by the DTG during the Phase 2 evaluation of both dycores. Computational costs are estimated based on AVEC reports made available by the DTG.

Initial Implementation (transition to operations) Cost in FTEs (in addition to existing personnel managing O&M for operational GFS)										
Activity	FY17		FY18		FY19		FY20		Total	
	MPAS	FV3	MPAS	FV3	MPAS	FV3	MPAS	FV3	MPAS	FV3
Dycore integration into NEMS	3	3	2	2	2	2	2	0	9	7
Physics implementation	2	1	2	1	1	1	1	0	6	3
Physics Driver implementation	1	1	2	1	1	1	1	0	5	3
DA integration	4	2	3	2	3	2	2	0	12	6
Pre/Post	2	2	2	2	1	1	1	0	6	5
Benchmarking	0	0	4	3	4	4	5	0	13	7
Code Management	2	2	2	2	2	2	2	2	8	8
Computational efficiency	2	1	2	1	2	1	2	0	8	3
Transition to operations	0	0	0	0	0	3	3	0	3	3
Total	16	12	19	14	16	17	19	2	70	45

Computer Resource Requirements for Initial Implementation (FY17-FY19 for FV3 and FY17-FY20 for MPAS)						
	CPU*	CPU Hours**	Disk	Period		% change w.r.t. GFS
GFS	5,150,880	399,840	10 PB	FY17-FY18		0
FV3	6,565,620	509,660	30 PB (2 streams)	FY17-FY19		28%
MPAS	19,959,660	1,549,380	45 PB (3 streams)	FY17-FY20		288%

CPU = Y x 4 cycles x 365 days x 3 years, Y is number of cores required for 8.5 min/day
Y = 1176 (GFS), 1499 (FV3), 4557 (MPAS) based on current operational resolution (~13 km). 1176 1499 4557
Computational requirements for intended implementation configuration TBD
**CPU hours = Y x 8.5 min/day x 10 days x 4 cycles
HPC resources for Data Assimilation is not included
Availability of computational resources will require development/testing of FV3 in two parallel streams while MPAS would require three parallel streams
Summary Implementation Costs (Human Resources) for MPAS are 55% more compared to FV3
Implementation Costs (computational resources) for MPAS are 204% more compared to FV3

Table 10.1: Cost (human and computational) estimates for the MPAS and FV3 dycores for initial implementation into operations.

Human capital (FTE) requirements are reported for each fiscal year for the activities listed in the table. FV3 development and implementation timelines (FY17-FY19) would require about 45 FTE, while for MPAS the timelines extend to FY20 and would require an additional 25 FTE to complete the development and implementation.

Computational resource requirements were based on the AVEC benchmark results for running FV3 and MPAS using GFS physics at current operational model resolution of ~ 13 km (Criterion 4). Operational requirements for GFS are to produce a 10-day forecast in about 85 minutes of wall-clock time and run three years of retrospective experiments on the development machine (WCOSS). Hence, the number of CPUs required for producing a one-day forecast in 8.5 minutes is used for computing total CPUs and total CPU hours for two sets of experiments for FV3 and three sets of experiments for MPAS. These

experiments will be configured to run in fully cycled mode, with production look-alike model products archived for evaluation and downstream model support. Since MPAS is more expensive and takes more time to run compared to FV3, EMC anticipate running three streams of experiments to complete all the necessary retrospective runs. EMC did not include data assimilation resource requirements in these computations, assuming that the resource requirements would be independent of the choice of the dycore (although ensemble component of data assimilation requires significantly higher computational resources for MPAS compared to FV3). The disk usage estimates will most likely remain constant and are estimated at 10 PB of disk and archive space on WCOSS and HPSS.

Summary of the cost estimates provided in Table 10.1 indicate that MPAS would require 55% more human capital and 204% more computational resources compared to FV3, assuming no significant improvements in the efficiency of MPAS.

2.10.4 Conclusion

Both cores, FV3 and MPAS, can be implemented into operations at NCEP following the plan outlined above. However it is estimated that the cost of implementing MPAS will be considerably more than for FV3 primarily because MPAS will require more effort to bring its current skill, as shown in Criterion 3, up to the same level as the current GFS. FV3 on the other hand was shown to have comparable skill to the GFS in the Criterion 3 tests. In addition, in the Criterion 4 tests, MPAS was shown to be considerably more computationally costly to produce the same forecast as FV3.

Implementing the FV3 dycore into operations will be an easier, less costly process than for MPAS.

Chapter 3 Phase 2 Conclusions

The various criteria that are part of the Phase 2 Test Plan can be generally divided into three groups. First, there are the criteria that do not involve actual testing of the dycores, but rather involve an informed analysis of a particular dycore aspect. These include Criterion 1 (suitability for space weather applications), Criterion 7 (adaptability to the NEMS/ESMF infrastructure), Criterion 8 (dycore documentation) and Criterion 10 (cost to implement in operations). For the criteria 1, 7, and 8, both dycores were deemed to have passed these tests. Criterion 10 will be an ongoing process for NCEP, though a clear initial plan for implementation for FV3 has been presented by NCEP/EMC.

The second group of criteria involved conducting actual tests of the FV3 and MPAS codes, and both models performed equally well in those tests. These include Criterion 2 (conservation of mass, tracers, entropy, and energy), Criterion 5 (simulation of moist convection and demonstration of variable-resolution and/or nesting capabilities) and Criterion 6 (stability in long-term integrations).

For the conservation test (Criterion 2), both dycores were shown to accurately conserve energy, mass, tracers, and entropy better than the GFS. This test (and also Criterion 6) was used to assess the impact of the computational grid on the solutions and, although grid imprinting was apparent, the impact on the solutions was quite small. Both dycores demonstrated a clear ability to accommodate variable-resolution and/or nesting with no apparent problems (Criterion 5). The GFS physics that was used for this test to isolate the impact of the dycores is only marginally suitable for the convection-permitting resolution of this test so a direct assessment of the accuracy of dycores in correctly forecasting deep convection was not possible. The long term (90-day) integrations with both dycores (Criterion 6)

showed no problems and compared well with GFS though some bias errors particularly with 2 m temperature were noted for both MPAS and FV3.

The third group of criteria involved tests that showed a distinct difference between the two dycores. These include Criterion 3 (retrospective forecasts from GFS conditions using GFS physics), Criterion 4 (computational performance) and Criterion 9 (performance in cycled data assimilation).

The Criterion 3 test involved running 10-day forecasts with both cores using the same GFS physics package and GFS initial conditions every five days for a calendar year. These forecasts were then analyzed using the standard NCEP verification package used to assess operational forecasts. Figure 3.5 in the Criterion 3 section shows the main message derived from these tests; that the FV3 clearly outperformed MPAS with considerably more skill over the year of forecasts. Similar differences were noted in the other verification statistics computed by NCEP. Furthermore, the FV3 skill was comparable to that of the operational GFS which the DTG considers remarkable given that the FV3 was run with no tuning of the GFS physics to the FV3 core while the GFS has benefited from many years of tuning.

The Criterion 3 test results were also used to estimate the effective resolution of the two dycores. That result is shown in Figure 3.2 of the Criterion 3 section. Both FV3 and MPAS have comparable effective resolution of around four times the grid spacing, and GFS has an effective resolution of approximately 10 times the grid spacing.

The Criterion 4 test involved comparing the computational efficiency of the dycores. This test used the models configured the same way they were for the retrospective forecasts presented in Criterion 3 including using the same GFS physics package and a nominal resolution as close as possible to the resolution of the operational GFS. The main result is displayed in Figure 2 of the Criterion 4 section - FV3 was approximately three times faster than MPAS. The stretched grid/nesting strategy employed by FV3 was also found to be more efficient than the variable-resolution mesh used by MPAS.

The Criterion 9 test ran both dycores with GFS physics within the NCEP operational data assimilation system. The data assimilation system setup was very similar to what is used operationally by NCEP though with a few differences as noted in the Criterion 9 section. Figures 9.1 and 9.3 which assess the accuracy of six-hour forecasts (which forms the first guess for the next data assimilation cycle) provide the gist of the message; FV3 forecasts initialized within a cycling data assimilation system are more accurate than MPAS forecasts and the reduced resolution baseline GFS configuration. This suggests that once fully operational with a properly tuned physics package, a global model configured with the FV3 core and cycled within the data assimilation system should outperform the current operational GFS.

With the evidence outlined above the DTG has no hesitation in recommending that NWS adopt the FV3 atmospheric dycore for the Next Generation Global Prediction System.

Chapter 4 Next Steps

Phase 3 of the dycore selection process will be the operational integration and implementation of the new core. In parallel with the dycore integration, other modeling components within the NGGPS will be upgraded. For example, the NGGPS program team is working in collaboration with the Physics Implementation Team, EMC, GMTB, and other modeling community members, to implement a CAPP. Additional efforts include the implementation of improved data assimilation capabilities (4DnVar with 4D incremental analysis update and stochastic physics). A community modeling environment is also

planned for future end-to-end model system development. Figure 4.1 reflects NOAA’s current plans and timelines, to implement FV3 in the evolution of the GFS.

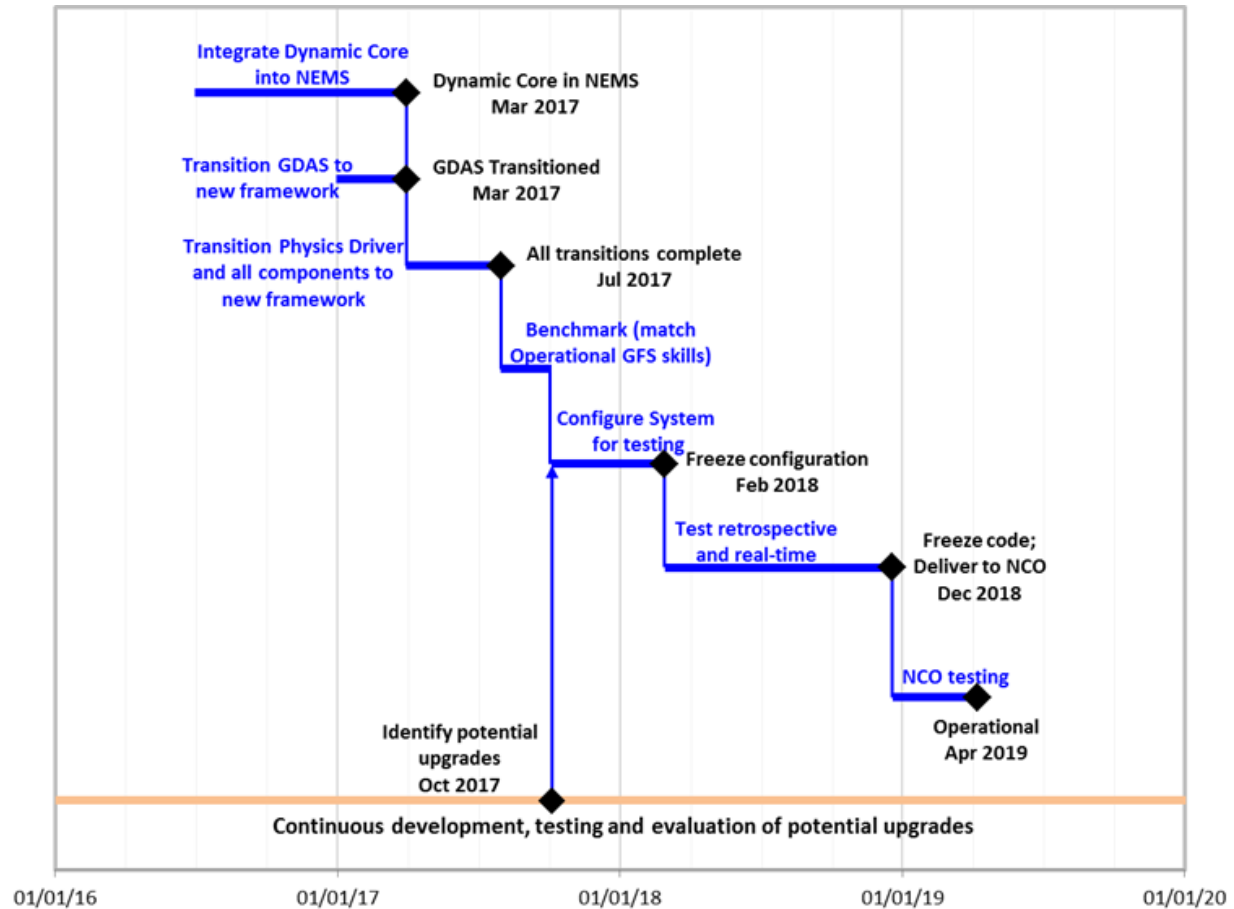


Figure 4.1: Proposed Phase 3 Implementation Plan.

Additionally, NOAA is moving to strengthen our community engagement on global modeling. Beginning in FY17 NOAA will establish a community associated with the implementation of the new GFS. While these are yet to be determined and/or finalized (at the time of this writing), a variety of activities are planned which will likely include one or more community workshops, rapidly spinning up “core users” with the new dycore to build and expand the developer and user bases, and working out an infrastructure to support community engagement. As these activities are identified and planned they will be announced on the NGGPS web page: http://www.weather.gov/sti/stimodeling_nggps.


Appendix A: Director, NWS Approval Memorandum

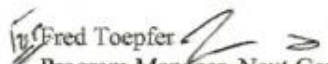


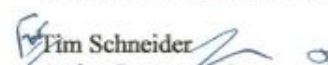
U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
NATIONAL WEATHER SERVICE
1325 East-West Highway
Silver Spring, Maryland 20910-3283

July 8, 2016

MEMORANDUM FOR: Dr. Louis Uccellini
Assistant Administrator for Weather Services
and Director, National Weather Service

FROM: Dr. Ming Ji 
Director of the Office of Science and Technology Integration
and Chair, Next Generation Global Prediction System Dynamic
Core Test Group

 Fred Toepfer
Program Manager, Next Generation Global Prediction System

 Tim Schneider
Acting Program Manager, Next Generation Global Prediction System

SUBJECT: NGGPS Phase 2 Dynamic Core Selection Recommendation

The National Weather Service Office of Science and Technology Integration recommends the selection of the Geophysical Fluid Dynamics Laboratory (GFDL) Finite Volume Cubed-Sphere (FV3) dynamical core for use in Next Generation Global Prediction System (NGGPS) Phase 3 dynamic core integration and implementation.

The NGGPS Phase 2 evaluation of candidate dynamic cores for the purpose of selecting a non-hydrostatic atmospheric dynamic core has been completed. The NGGPS Dynamic core Test Group (DTG) reviewed and assessed the findings of the Phase 2 evaluation and has determined that the GFDL FV3 dynamic core represents the lowest risk, lowest cost alternative for the new NGGPS atmospheric model. Compared to the National Center for Atmospheric Research (NCAR) Model for Prediction Across Scales (MPAS) dynamic core, FV3:

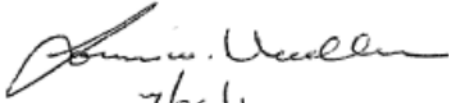
- Meets all technical needs
- Less expensive to implement
- Higher readiness for implementation
- Significantly better technical and computational performance
- Lower risk

Based on this assessment, we recommend selection of the GFDL FV3 dynamic core for use in NNGPS Phase 3 dynamic core integration and implementation.

Approved ✓

Disapprove

Let's Discuss



7/26/16

Date

Date

Date

Appendix B: Dynamical core Test Group Charter

Dynamical Core Test Group (DTG) Charter

29 May 2015

Prepared by:

NGGPS Project Management Team

1. BACKGROUND

As part of its Research to Operations (R2O) Initiative, the National Weather Service (NWS) plans to develop and implement a state-of-the-art Next Generation Global Prediction system (NGGPS) which will be readily adaptable to and scalable on evolving High Performance Computing (HPC) architectures. The NGGPS will be designed to produce useful forecast guidance to 30 days, as well as become the foundation for the operating forecast guidance system (Global Forecast System) for the next several decades. Current research and development efforts both inside and outside NWS, including the Navy, NOAA laboratories, National Center for Atmospheric Research (NCAR), the university research community, and other partnership efforts, will contribute to the development of this prediction system.

The current operational Global Forecast System is based on the Global Spectral Model. The Global Forecast System is an operational, very mature hydrostatic dynamic model with known limitations regarding scalability and adaptability to future computing architectures. The future NGGPS will require an atmospheric dynamical core (dycore) that is non-hydrostatic, highly scalable and architecturally compatible with projected high performance computing architecture. Six dycores currently being developed and modified from a variety of institutions are viewed as potential candidates to be evaluated for the new system. The NGGPS Dycore Testing Plan will guide the testing of these dycores and leverage ongoing High Impact Weather Prediction Project (HIWPP) activities in the evaluation of the dycores.

Objective and unbiased assessment of the test and evaluation results is essential to the selection of the future atmospheric model dynamical core for the NGGPS. A Dynamical core Test Group (DTG) is hereby established to conduct this assessment. The DTG will evaluate the test plan, conduct of the test, and results of the NGGPS evaluation and provide an assessment, either individually or collectively of their evaluation to NWS management. This assessment, along with business considerations will be used in the development of the business case supporting the selection of the next dycore by NWS management.

2. ROLES

The role of the DTG is to review the technical aspects of all dycore testing and provide an assessment of test results in written reports to NWS management for each of the candidate dynamical core codes. The DTG will provide guidance on outstanding issues relayed from the Advanced Computing Evaluation Committee (AVEC) or the NGGPS Project Management Team regarding the preparation for and conduct of dycore performance testing and will advise on resolution of conflicts on testing procedures, scoring or ranking

Initial Phase 1 and Phase 2 testing is described in the NGGPS Testing Plan, however, the DTG will assess the evaluation criteria and provide feedback to the NGGPS Program Manager as applicable. Overall Phase 1 testing

results will be compiled by the NGGPS Project Management Team for presentation to the DTG for review. The DTG will complete a review of the Phase 1 testing data and provide an assessment to the NGGPS Program Manager prior to Phase 2 testing. Upon completion of Phase 2 testing, evaluation of all test result data will be performed by the DTG. Results of this evaluation will be included in a final report prepared for the NGGPS Program Manager.

The DTG will meet as needed to review test and evaluation procedures and to conduct assessments of dycore test data.

DTG deliberations and products are confidential until released publicly by the NWS.

3. PARTICIPANTS

Each candidate dycore shall have one representative on the DTG. Technical consultants will also be included in the group. Other technical representatives, including the NGGPS Test Manager and the Advanced Computing Evaluation Committee Chair, will participate as needed.

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

Staff: Steve Warren/Sherrie Morris

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. Venkatachalam 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

4. FUNDING

There is no dedicated funding associated with the committee for Federal Employees. Consultants will be compensated in accordance with their contract.

5. PERIOD OF AGREEMENT AND MODIFICATION/TERMINATION

This Agreement shall become effective on the date of approval by the NGGPS Program Manager. The remaining signatures constitute the agreement by the participating organizations and or consultants to participate. It is anticipated that the charter will be terminated once the tasks of the DTG are completed. Completion of DTG tasking is expected to coincide with the delivery and acceptance of a written dycore test assessment (individually and/or collectively) to/by NWS management. Any extension of the agreement will be proposed, as necessary, by NWS management.

Reviews of this charter may be conducted as deemed necessary by the DTG or NGGPS Project Team at any time. The latest date of amendment constitutes the new effective date unless some later date is specified.

6. SIGNATURE – *signatures are on file*

_____ Date _____
Chair

_____ Date _____
NGGPS Program Manager

7. ACKNOWLEDGMENT

_____ Date _____
Consultant

_____ Date _____
GFDL – FV3

_____ Date _____
Consultant

_____ Date _____
EMC – NMMB and GSM-NH

_____ Date _____
Consultant

_____ Date _____
NCAR/MMM – MPAS

_____ Date _____
Navy/NRL - NEPTUNE

_____ Date _____
ESRL/GSD - NIM

Appendix C: SME Variable-Resolution

Discussion with Subject Matter Experts (SME) on the Variable-Resolution Tests

As the DTG began to discuss and evaluate the results from the variable-resolution tests, in particular the high-resolution, real-data hindcast of the Moore Tornado, May 19-21, 2013 and the supercell Idealized case described by Klemp et al. (2015), a number of questions were noted that the DTG felt needed help from experts not on the DTG. It was noted early on in DTG discussions of the Moore Tornado Case that there were major problems with the GFS physics used in the simulations. The DTG felt external guidance was needed to disentangle the physical and dynamical core impacts on the simulation.

The GFS physics was chosen for these tests because it works well at larger scales, it allows direct comparison between the cores themselves, and it gives neither core an obvious advantage over the other in the tests. GFS physics will also be used in the initial testing of NGGPS at NCEP. If each core used different physics or a physics package specifically tuned to one of the cores, direct comparison of the dynamical cores would not be possible⁶. The GFS physics were ported to both cores and configuration of physics was a collaborative effort of the DTG participants. The GFS physics has been carefully tuned for resolutions coarser than 13 km but not for resolutions approaching 3 km where non-hydrostatic effects become important. As noted below, particular concern was voiced by the SME on whether or not the microphysics used in the GFS was appropriate for 3 km simulations.

The idealized supercell case used a highly simplified physics where the microphysics was the Kessler warm rain approximation. However, the DTG had concerns that even though both models showed that they converge on a solution shown by going to ever-higher resolution, they were not the same solution.

Given these questions and concerns a SME group from outside the DTG was consulted. The group and our specific questions are shown below.

Members of the Subject Matter Expert Group

Jack Kain - NOAA/NSSL
Louis Wicker - NOAA/NSSL
David Stensrud – Penn State Univ
Derek Posselt – Univ of Michigan
Paul Markowski – Penn State Univ
Yvette Richardson – Penn State Univ
William Putman - NASA/GSFC

⁶ In Phase 1 of the tests, forecasts were performed with native physics at 3 km global resolution. In that test, FV3 and MPAS both produced realistic fine scale detail in tropical cyclones and regions of severe convection.

The questions for the SME:**Idealized Supercell Tests**

1. Should this idealized supercell test (from Klemp et al. 2015) be definitive for a storm-scale mesoscale model application for the non-hydrostatic global model, as we have assumed?
2. Should it be possible to obtain a converged solution on the idealized supercell test by going to higher resolution?
 - Is obtaining a converged solution necessarily possible for both cores given their different formulations?
 - Is it possible that the converged solution for each core will be different?
 - What are the key characteristics of a supercell to look for in results from this idealized test?
 - Do you see key differences between the FV3 and MPAS simulations that concern you?
3. What do you see in the 0.5 to 1 to 2 to 4 km resolution?
 - Are we correct in looking for as similar a solution as possible between these resolutions for our NGGPS supercell test? Do we need to add a 3 km test?
 - Is the ability to produce a supercell at marginal resolution a key characteristic to look for (another possible reason for the 3 km test)?
4. What other tests or graphics would be desirable to answer the questions above?
5. What should be the role of horizontal and vertical diffusion in these tests in:
 - Convection evolution using realistic values?
 - Obtaining a converged solution

Moore tornado case

1. Is the GFS physics (as it stands, with parameterized convection either on or off) good enough for the purpose of evaluating the performance of the two dynamical cores at 3 km resolution for this case?
2. What aspects of the simulations might be robust [useful?] indicators of the capabilities of the dycores in simulating convective systems at non-hydrostatic scales?
3. We know that we cannot expect absolute forecast accuracy at such lead times. We also know that certain features will be misrepresented because of the current microphysics scheme. Can we nevertheless expect certain qualitative aspects of the solution, such as the spatial and temporal structure of the storms, to be well captured?
 - If so, which aspects?
 - What diagnostics should we look at in order to examine such aspects?
4. We feel we need to quantify consistency between divergence, vertical motion and precipitation for example are we seeing cell-size precipitation or is it organizing into multiple cell complexes.
5. Given that we have, perhaps, fundamental issues with physics, if we focus on the initiation of "events" rather than the evolved convective system (because as the events evolve the physics deficiencies are likely to make the evolution of convection odd—cold pools etc.), do we see consistent differences between the cores that are a concern?

The process with the SME began with a conference call involving the members of the SME group listed above, the DTG consultants, the DTG Test Manager, Jeff Whitaker and DTG members, Jim Doyle and Stan Benjamin. The purpose was to outline to the SME the NGGPS process and results to date related to the variable-resolution tests. The questions shown above were e-mailed to the SME group at the same

time. All members of the group responded with written comments, and we followed up the written comments with a conference call. Comments were discussed during the call, along with the possibility of further tests. Additional tests included runs of the supercell test with different diffusion and additional diagnostics of the Moore Tornado case including hourly output and a careful comparison of the grid resolution in the high-resolution region. Results from these tests and additional diagnostic output are mentioned in other portions of this report.

The comments and discussions from the SME were extensively considered by the full DTG both at its face-to-face meeting in Silver Spring and again its weekly conference calls.

There was concern among the SME that there wasn't enough information in either of the tests they were shown to evaluate, definitively, the core's ability to forecast deep convection. This was, in significant part, due to the common physics that the DTG chose not being optimal for resolutions of around 3 km. Further, the GFS microphysics did not have the completeness of state-of-the-art convection models. The SME noted that at this scale the solutions were very sensitive to the microphysics. Indeed, a number the SME pointed to deficiencies in the microphysics that led to problems forecasting formation of cold pools associated with moist convection.

Still, with this limitation, the DTG believed enough was learned to help in the overall evaluation of the cores. In particular, the DTG was looking for impacts from a particular core that may preclude its use in future development of the core into a model fully capable of simulating and forecasting deep convection. This would, of course, require adopting of a physics package that would work at convection resolving resolutions—such a physics package is under development at NCEP. The DTG was looking for “show stoppers,” and did not find any in this regard for the MPAS and FV3.

The SME recommended the DTG look at hourly output from the Moore real data case, where the focus would be on the early development of the convection before the defects in the microphysics severely impact the convection evolution. The DTG was concerned in the supercell case whether the “converged” simulation from each core should be expected to be the same. The SME advice was that is probably not the case due to differences in numerical formulation and grid between the models (see also, Williamson, 2008; Scott et al., 2016). The SME also gave the DTG good suggestions on other tests that might be performed with the idealized system— sensitivity to diffusion for example. These results are discussed elsewhere.

Working with the SME on the question of the deep convection was good and useful and helped the DTG in making recommendations for the final choice of a core for the NGGPS. Particularly, neither core has any obvious defects that will preclude its development into a model fully capable of forecasting and simulating moist deep convection.

References

Klemp, J. B., Skamarock, W. C., and Park, S.-H., 2015: Idealized global nonhydrostatic atmospheric test cases on a reduced-radius sphere, *J. Adv. Model. Earth Syst.*, 07, doi: 10.1002/2015MS000435.

Scott, R. K., Harris, L. M., and L. M. Polvani, 2016: A test case for the inviscid shallow-water equations on a sphere, *Q. J. R. Meteorol. Soc.*, 142, 488-495.

Williamson, D. L., 2008: Equivalent finite volume and Eulerian spectral transform horizontal resolutions established from aqua-planet simulations, *Tellus*, 60A, 839-847.

Appendix D: AVEC Phase 2 Report

AVEC Report:

NGGPS Phase-2 Benchmarks and Software Evaluation

Advanced Computing Evaluation Committee

Initial draft: 6/15/16;

Final version: 7/1/2016

I. Executive Summary

Phase 2 of the Advanced Computing Evaluation Committee (AVEC) was formed in the fall of 2015 to evaluate HPC performance, suitability and readiness to inform final selection of a new non-hydrostatic, dycore to meet National Weather Service's operational forecast requirements for Next Generation Global Prediction System (NGGPS). The two dycores evaluated were NOAA/GFDL's FV3 and NCAR's MPAS, finalists from the Phase 1 NGGPS dycore evaluation. This report describes methodology, cases, model configurations, and results of benchmarks conducted during dedicated access to Cori, a 52-thousand processor core supercomputer at the U.S. Department of Energy's National Energy Research Scientific Computing Center (NERSC)⁷. AVEC's testing addressed Criteria 4 and 5 in the NGGPS Test Plan: computational performance with then-current operational GFS physics, and computational efficiency of variable-resolution and/or nesting capabilities of the two models. The dycores' software implementations were also evaluated for suitability on next-generation HPC architectures (Criterion 10).⁸

For Criterion 4, the AVEC tested the number of computational cores (processors) needed to achieve a speed of 8.5 minutes per day at 15 km, 13 km and 11 km nominal horizontal resolution. MPAS required between 2.5 and 3 times more processors than FV3 at the three nominal resolutions tested. Moreover, FV3 at the finest horizontal resolution (11 km) required fewer cores than MPAS at the coarsest resolution (15 km). FV3 required 26 percent more processors than the 13 km hydrostatic GFS running operationally at NCEP.

For Criterion 5, the AVEC measured how efficiently each dycore was able to focus computational resources over non-uniform (nested or mesh-refined) resolution domains relative to the cost of a uniform 3 km domain. FV3's nesting scheme was 97 percent efficient compared with 64 percent efficiency for MPAS's in-place refinement.

⁷ <https://www.nersc.gov>

⁸ See NGGPS Dycore Test Plan: http://w2.weather.gov/sti/stimodeling_nggps_implementation_atmdynamics.

Evaluation of the models' software implementation uncovered no unusual risks or incompatibilities for next-generation HPC architectures that would be a concern for Criterion 10 or the NGGPS implementation plan.

The following is a chronology of the Phase 2 testing conducted by AVEC.

- The NGGPS Phase 2 Benchmarking Test Plan was created in November, 2015, with concurrence on workloads, tests and evaluation methods, completed in January 2016.
- The code and workload configurations were finalized February, 2016.
- Benchmark codes, data and verification programs were delivered 11 April 2016.
- Benchmarks were conducted during an eight-hour session at NERSC, 28 April 2016.
- A second one-hour benchmark session was conducted at NERSC, 24 May 2016.
- Full agreement by AVEC members on the contents of this report, 1 July 2016.

AVEC was chaired by John Michalakes (UCAR). AVEC members were Rusty Benson (NOAA/GFDL), Mark Govett (NOAA/ESRL), Mike Young (NOAA/NCEP), and Michael Duda (NCAR). Michael Duda participated fully in AVEC Phase 2 discussions and activities but ceased participation in AVEC after 20 May 2016, when NCAR formally withdrew MPAS from consideration as a dynamical core for NGGPS and ceased participation in the Dycore Test Group. The remaining members of the AVEC, working in accord, completed and approved this report.

The remainder of this report provides details on the benchmark workloads, methodologies, and results summarized above.

II. Performance with GFS Physics (Criterion 4)

Performance of the two candidate dynamical cores running with GFS physics was measured as the number of processor cores needed by the model to achieve the current operational threshold of 8.5 minutes of wall clock time per day of forecast, disregarding initialization and I/O costs. The modeling groups agreed to three workload configurations with nominal horizontal resolutions of 15 km, 13 km and 11 km. The groups then provided AVEC with codes, datasets, and verification scripts. The workloads and configurations are shown in Table 1. AVEC ran each workload on several different numbers of processing cores that gave model performance above and below the target simulation rate of 8.5 minutes per day. These results are shown in Figure 1.a. Figure 1.b shows the time spent in the dynamical core alone. The number of processing cores needed was then estimated by interpolation between the core-counts that straddled the 8.5 min/day target simulation rate. These results are shown in Figure 2.

The workloads used were based on the test cases used in the NGGPS Criterion 3 test: "Robust model solutions under a wide range of realistic atmospheric initial conditions using a common (GFS) physics package." The workload used initial conditions for the ten-day retrospective case starting at 00Z on 1 August 2015. To limit the amount of machine time needed while capturing the full diurnal cycle, only the first 24 hours were benchmarked. As it turned out, what little variation there was in the measured time-per-time step over the course of a run was almost entirely from the GFS physics. AVEC was able to isolate the cost of GFS physics using a special purpose timing package.

Additional technical detail. Timing data was collected using a set of low-overhead timers developed for the AVEC tests⁹ that, when inserted around sections of the code, generated a per time-step series of timings from each MPI task for each invocation of the instrumented section of code. The FV3 and MPAS modeling groups inserted calls to the AVEC timers into their codes to measure the overall time for each time step and the time for calls to GFS physics. The dynamics-only cost of the runs was the difference between the cost of a time step minus the cost of GFS physics for that step. Cost of model initialization and I/O was disregarded.

Benchmarks were conducted on an otherwise empty Cori system to avoid unwanted run-time variation caused by contention with other jobs running on the system. In addition, the AVEC timer data from each run was post-processed to filter other sources of run-time variability (e.g. periodic background system tasks). Figure 3 shows a sample of benchmark data that was collected for each run, before and after filtering.¹⁰ Each plotted result shows the cost per time step and cost per time step without physics (i.e. dynamics) as a time-series over the course of a one-day forecast. The cost per step was the maximum time over all MPI tasks in the run. The cost per step without physics was the cost per step (above) minus the maximum physics time over all MPI tasks. The filtered plots were produced by computing the standard deviation of the series of times per step minus physics and then truncating any value that exceeded that value by a given factor of the standard deviation. This “clipping” factor is listed in the legend of each plot. On a few occasions, results that showed excessive system-dependent noise were discarded and rerun during the benchmarking session.

Performance with advection of additional tracers. Each group was asked to provide two workloads based on the 13 km GFS physics benchmarks above, one with 15 and one with 30 additional artificial tracers, to measure the rate at which computational cost increased as a function of additional tracers. The benchmarks were run on the number of processor cores that was close to the number needed to run at 8.5 mins per day without additional tracers. Results are shown in the third, fourth and fifth columns of the table below. The factor of increase from three tracers to the highest number of tracers is shown in the last column.

	Cores	Number of tracers / Minutes			Factor (lowest to highest)
MPAS	4800	3 / 8	18 / 14.6	33 / 19.8	2.5
FV3	1536	3 / 8.14	15 / 9.8	30 / 12.0	1.5 (1.53 adjusted)

Note: it was only discovered after all benchmarks were completed 24 May that the FV3 group interpreted the instructions to mean the workloads should have 15 and 30 tracers *total*, whereas the MPAS group interpreted the instructions to mean 15 and 30 *additional* tracers, as was stated in the AVEC Phase 2 Test Plan. Therefore, the table also shows the actual number of tracers run for each code. The FV3 results show a roughly linear increase in cost with additional tracers; thus, had the benchmark been done with 33 instead of 30 tracers, the factor of increase would be 1.53, as indicated in parentheses. AVEC regrets this methodological error, but believes it does not impact the finding that FV3 was more efficient than MPAS with additional tracers.

⁹ https://michalakes.svn.cloudforge.com/rtrtmic/avec_timer

¹⁰ http://www.esrl.noaa.gov/gsd/ato/AVECPhase-2Benchmarks20160428_adjusted.pdf

III. Computational efficiency with non-uniform resolution

As part of the Criterion 5 evaluation, “Demonstration of variable-resolution and/or nesting capabilities, including physically realistic simulations of convection in the high-resolution region”, AVEC conducted benchmarks to determine how efficiently the candidate models were able to focus computational resources over a higher resolution region of interest compared with the cost of running uniformly high-resolution over the full global domain. The top half of Figure 4 shows the definition used to calculate “refinement efficiency”. Ideally, the best improvement possible should be the cost in operations to compute the uniform high-resolution domain divided by the lower number of operations needed to compute the case where only part of the domain is high-resolution. The refinement efficiency E was the ratio of measured ($S_{measured}$) versus ideal speedup (S_{ideal}) of the non-uniform resolution code over a uniform 3 km workload using the same number of processing cores. The benchmark measured inefficiency resulting from additional communication, smoothing and interpolation, and computations over duplicated or transitional parts of the domain.

The bottom half of Figure 4 shows the uniform and non-uniform resolution configurations that were benchmarked for each model, the benchmark timings, and the resulting refinement efficiencies for the two dynamical cores, excluding the cost of GFS physics. FV3 was 97 percent efficient; MPAS was 64 percent efficient. Figure 5 shows the distribution of grid cell sizes used in the non-uniform resolution workloads run for FV3 and MPAS. A possible explanation for the marked difference in efficiency is that the models used different approaches to implementing non-uniform resolution. MPAS’s in-place grid refinement varied spatial resolution but not temporal resolution, so that the small time step needed at the finest resolution was used over the full domain. There may also have been inefficiency resulting from the cells of intermediate resolution in the transition zone. FV3 implemented non-uniform resolution by overlaying a two-way, high-resolution nest onto a moderately-stretched global grid and was able to apply different dynamics time steps to the global grid versus the nest.

IV. Readiness for next-generation HPC

AVEC was directed by the NGGPS program manager to evaluate and report on the readiness of MPAS and FV3 software for next-generation HPC as follows:

AVEC will review available evidence and provide a consensus report on specific serious or otherwise significant weaknesses (if any) uncovered in the design and implementation of a candidate model’s algorithms, data structures, or code that, in AVEC’s opinion, present unusual or unreasonable risk for NGGPS on next-generation HPC architectures. Given the uncertainty about still-evolving HPC technology, AVEC limits on time and resources, and the limited breadth and diversity of HPC subject matter expertise available for a thorough and objective evaluation, the AVEC is not asked to determine which candidate model is “better” than the other for next-generation HPC at this moment in its development; only that there are no foreseeable “show-stoppers.” In the event issues are found to exist, the AVEC’s report can be used by NGGPS program management to inform its business-case analysis. The report should be reasonably brief and at a level that is readable and understandable by NGGPS program management, the DTG and their consultants. The AVEC may use external SENA and associated resources to conduct this short analysis.

The following list of potential concerns for performance or usability of the dycores on current and next generation software. The points below are based on AVEC’s experience working with the codes during

the setup and running of the benchmarks at NERSC, and from AVEC's review of more detailed reports produced by Mark Govett, James Rosinski and Tom Henderson at NOAA/ESRL.^{11,12}

- MPAS
 - MPAS grids are defined and decomposed over processors using off-line grid generation software that has not been parallelized. Generating and decomposing large grids need only be done once per configuration; however, the cost in terms of time and limits on available memory is a concern.
- FV3
 - The cost of vertical remapping, while small in the workloads evaluated by ESRL, can become significant if vertical remapping needs to be called more frequently for different configurations. Certain inefficiencies relating to loop nesting and data organization in the vertical remapping were also identified. Effort to improve computational efficiency of vertical remapping is recommended.
 - The ESRL team identified a potential for inefficiencies from load imbalance in FV3 shared-memory parallelism where threading is over both transverse and vertical dimensions. GFDL responded that care should be taken to configure the model optimally and that such information will be included in the documentation.

Otherwise, AVEC found no serious or otherwise significant weaknesses in the candidate models that present unusual or unreasonable risk for NGGPS on next-generation HPC architectures.

Acknowledgements

We wish to thank the U.S. Department of Energy's National Energy Research Scientific Computing Center for the allocation of time, access to the Cori supercomputer and invaluable assistance setting up and conducting the NGGPS Phase-2 benchmarks. We especially wish to thank NERSC director Dr. Sudip Dosanjh and NERSC staff members Rebecca Hartman-Baker, Clayton Bagwell, Richard Gerber, Nick Wright, Woo-Sun Yang, and Helen He.

¹¹ http://www.esrl.noaa.gov/gsd/ato/FV3_Analysis-final.pdf

¹² <http://www.esrl.noaa.gov/gsd/ato/MPAS-Analysis-final.pdf>

Table 1: Model configurations for benchmarking FV-3 and MPAS with GFS physics.

Eval. Criterion #4 -- Performance with GFS Physics		
	FV-3	MPAS
Nominal resolution (km)	13.03 (equat.), 12.05 (avg.)	13
Grid Points	3,538,944	3,504,642
Vertical Layers	63	63
Time Step (sim. sec)	112.5 (dyn.), 18.75 (acous.)	75 (transport), 37.5 (dynamics), 18.75 (acoustic)
Radiation Time Step	3600	3600
Physics (other) Time Step	225	225
Tracers	3	3
Coarser than nominal resolution (km)	15.64 (equat.), 14.46 (avg.)	15
Grid Points	2,547,600	2,621,442
Vertical Layers	63	63
Time Step	225 (dyn.), 22.5 (acous.)	90 (transport), 45 (dynamics), 22.5 (acoustic)
Radiation Time Step	3600	3600
Physics Time Step	225	180
Finer than nominal resolution (km)	11.72 (equat.), 10.34 (avg.)	11
Grid Points	4,816,896	4,858,092
Vertical Layers	63	63
Time Step	112.5 (dyn.), 16.07 (acous.)	60 (transport), 30 (dynamics), 15 (acoustic)
Radiation Time Step	3600	3600
Physics Time Step	225	180

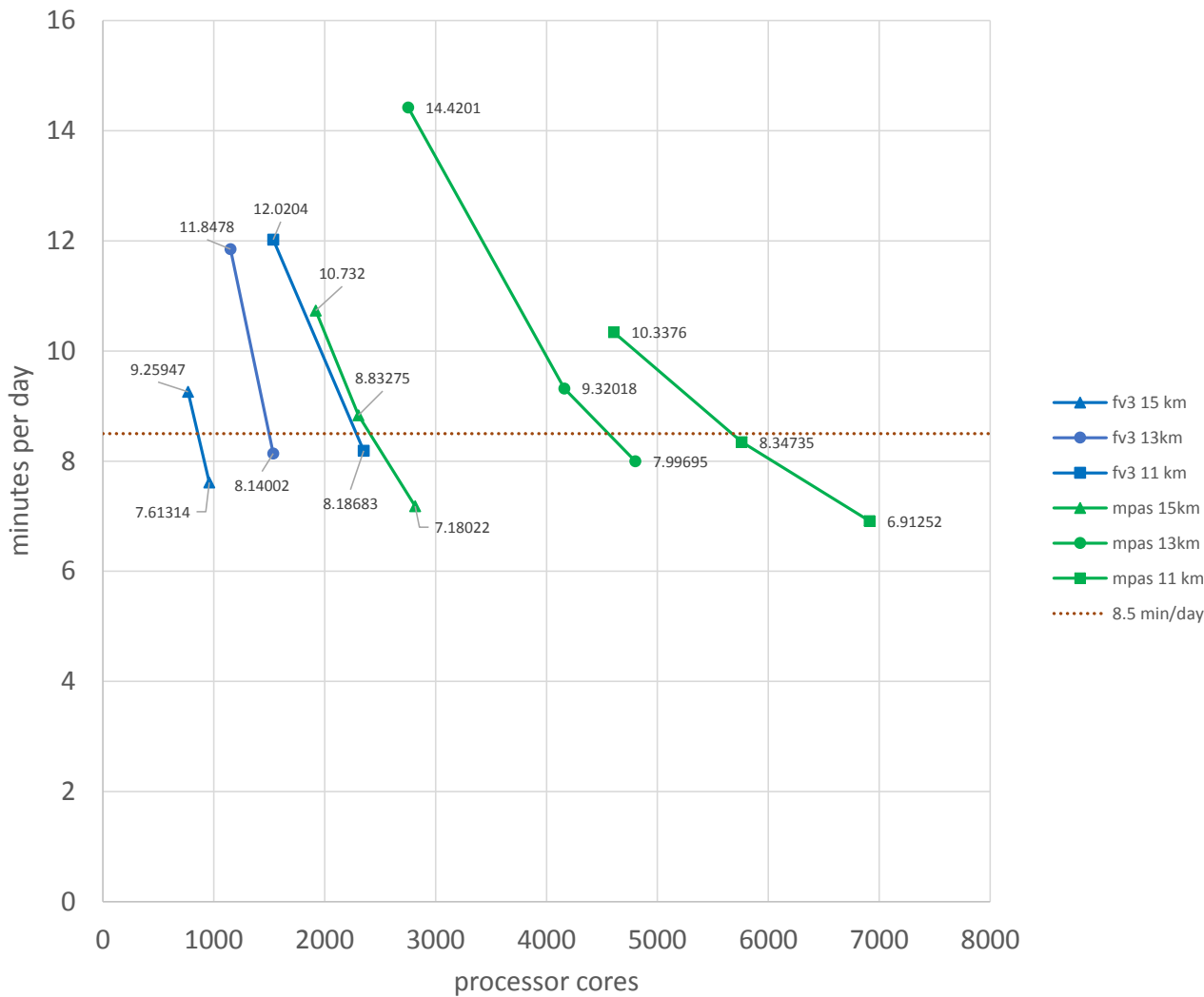
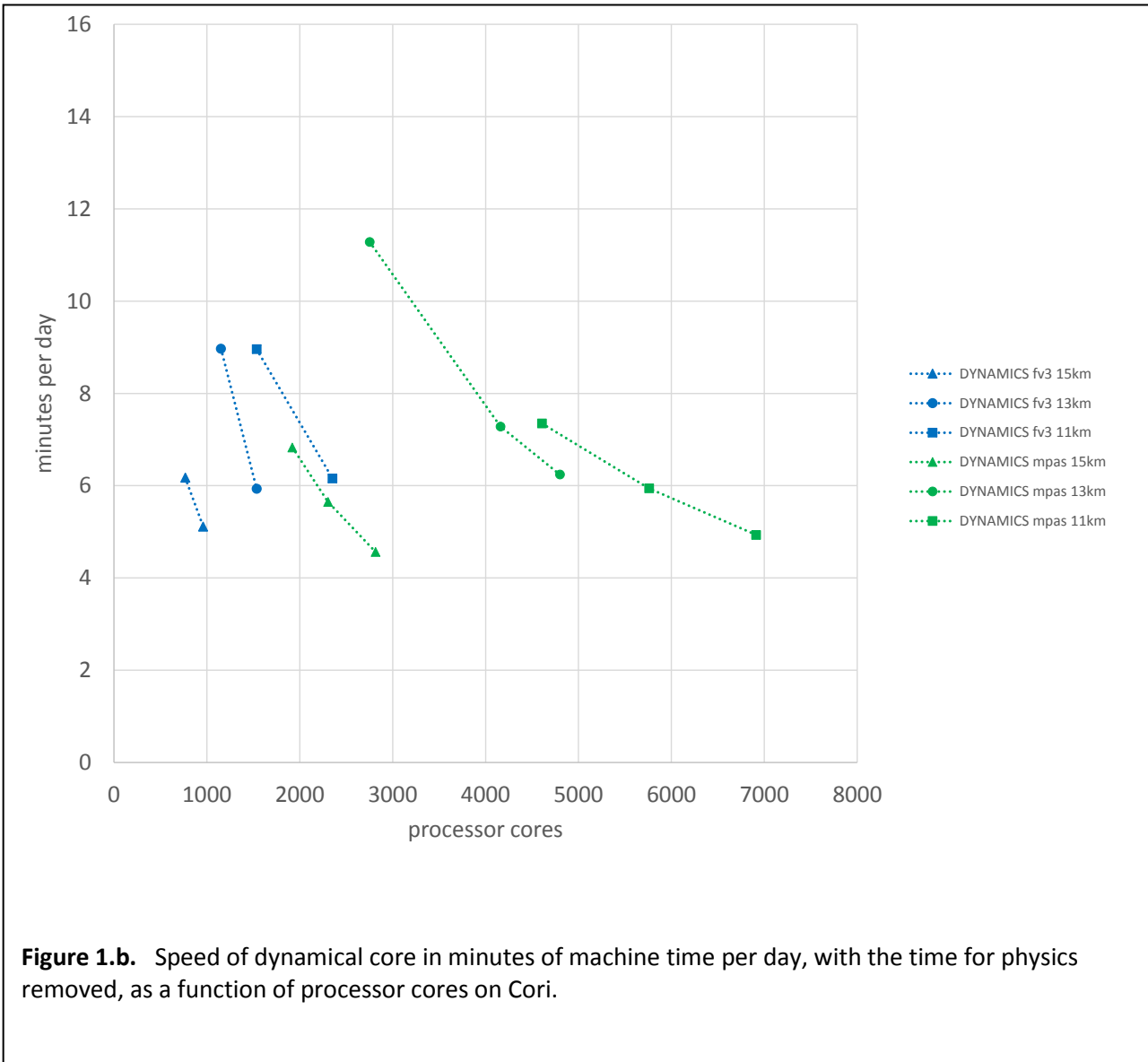


Figure 1.a. Speed in minutes per day as a function of number of processor cores on Cori. Dotted horizontal line indicates operational speed requirement of 8.5 minutes per forecast day. The intersection with the plotted lines is used to estimate the number of processor cores required to meet the operational speed requirement shown in Figure 2.



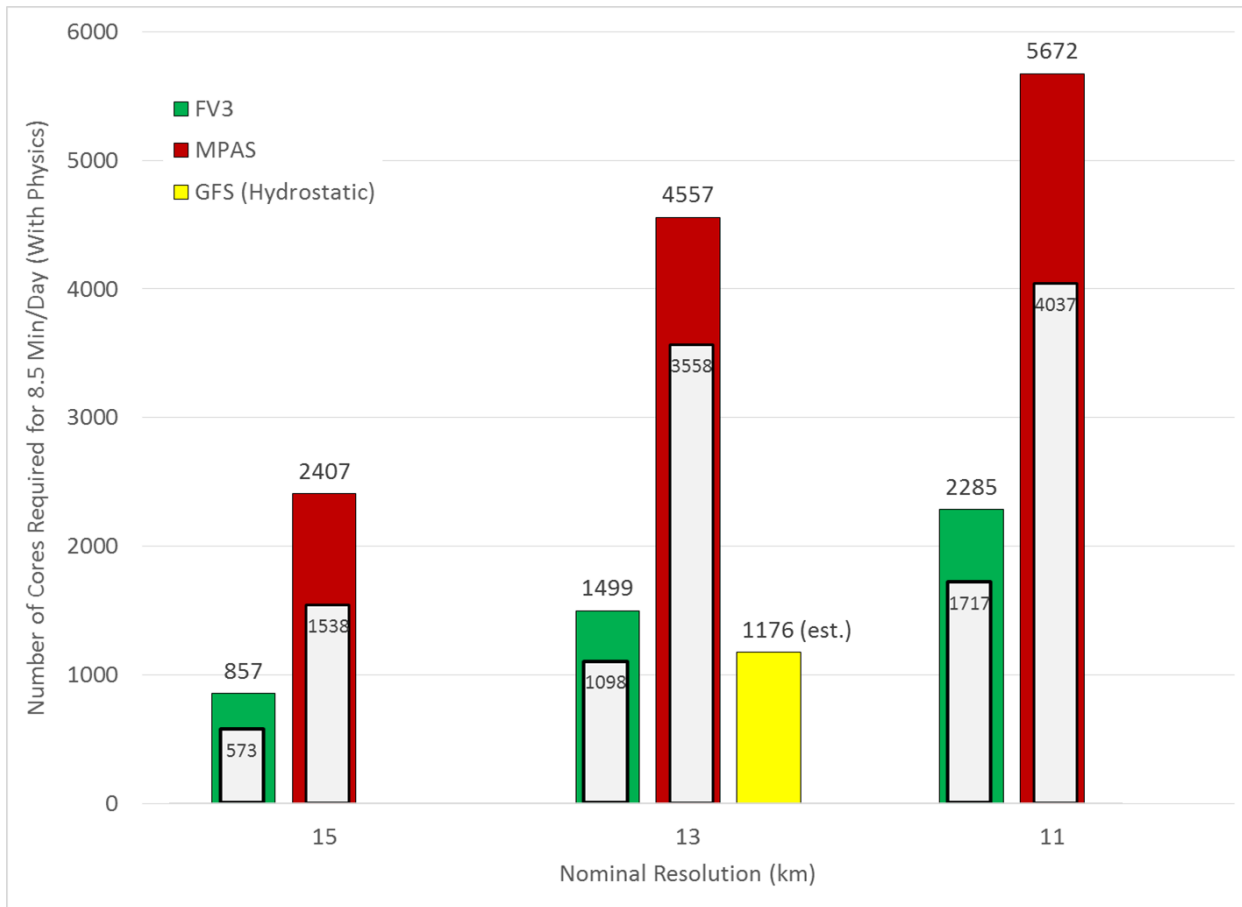


Figure 2: Cores required to meeting 8.5 minutes per day forecast speed requirement for operations at 15, 13, and 11 km horizontal resolution. All cases used 63 vertical levels. Colored bars show time with GFS physics; insets show the fraction of cores required by the dycore alone. The estimated number of cores required to run the 13 km operational GFS in 8.5 minutes on NCEP’s WCOSS Cray XC40 is shown for comparison.

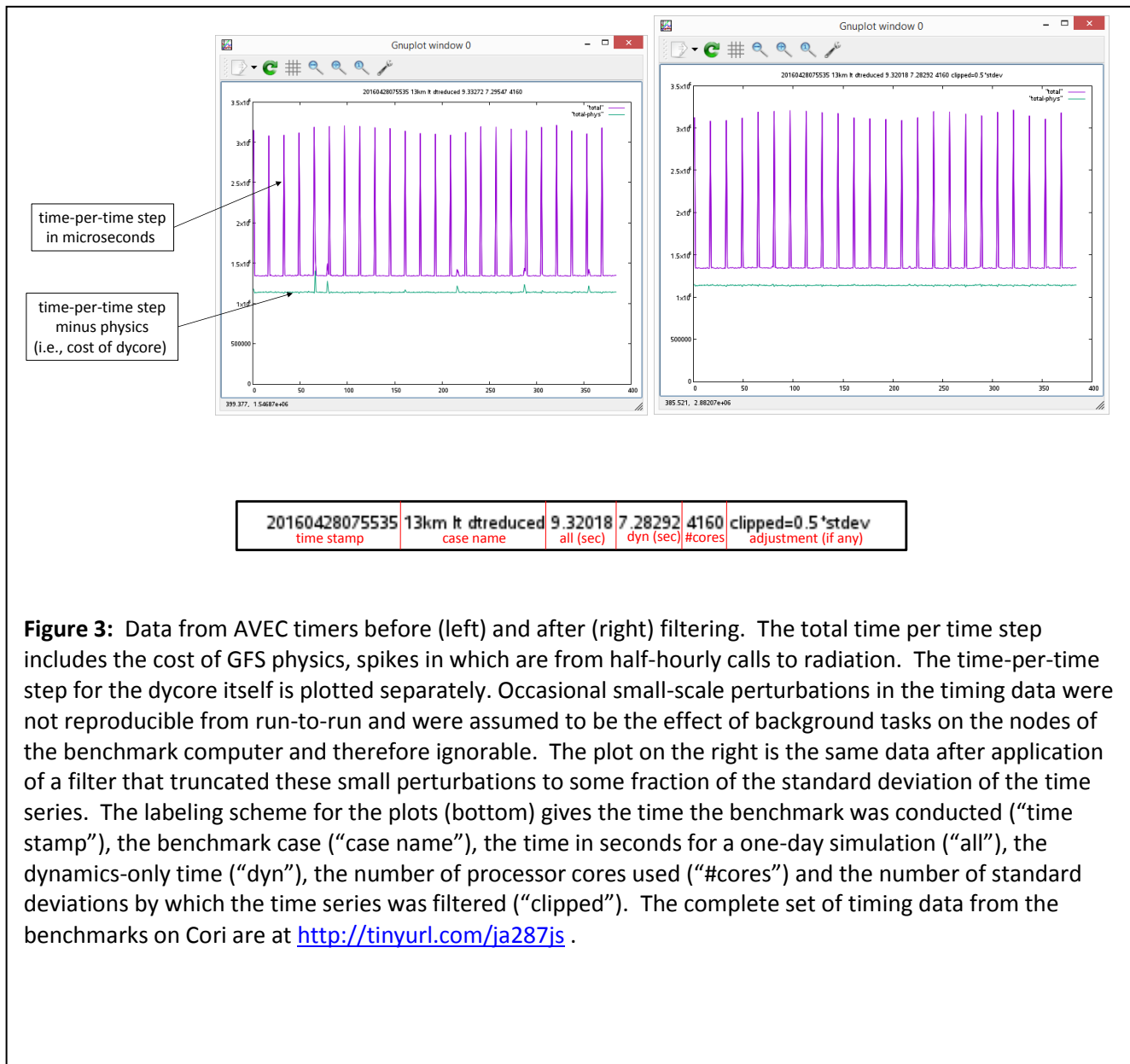


Figure 3: Data from AVEC timers before (left) and after (right) filtering. The total time per time step includes the cost of GFS physics, spikes in which are from half-hourly calls to radiation. The time-per-time step for the dycore itself is plotted separately. Occasional small-scale perturbations in the timing data were not reproducible from run-to-run and were assumed to be the effect of background tasks on the nodes of the benchmark computer and therefore ignorable. The plot on the right is the same data after application of a filter that truncated these small perturbations to some fraction of the standard deviation of the time series. The labeling scheme for the plots (bottom) gives the time the benchmark was conducted (“time stamp”), the benchmark case (“case name”), the time in seconds for a one-day simulation (“all”), the dynamics-only time (“dyn”), the number of processor cores used (“#cores”) and the number of standard deviations by which the time series was filtered (“clipped”). The complete set of timing data from the benchmarks on Cori are at <http://tinyurl.com/ja287js>.

Definition of nesting efficiency E:

a_g = area of domain ($5.101e14 \text{ m}^2$)

a_h = area of refinement (FV3: $2.52e13 \text{ m}^2$; MPAS: $2.82e13 \text{ m}^2$)

$r = a_h / a_g$ ← fraction of domain at high resolution (for uniform resolution domain, $r = 1$)

dx_L ← lowest resolution in non-uniform resolution run

dx_H ← highest resolution in non-uniform resolution run

$C = r (dx_L / dx_H)^3 + (1 - r)$ ← idealized cost for a run, assuming constant cost per cell step

$$S_{ideal} = \frac{(dx_L / dx_H)^3}{r (dx_L / dx_H)^3 + 1 - r} \left\{ \begin{array}{l} \leftarrow C_{uniform} \\ \leftarrow C_{refined} \end{array} \right.$$

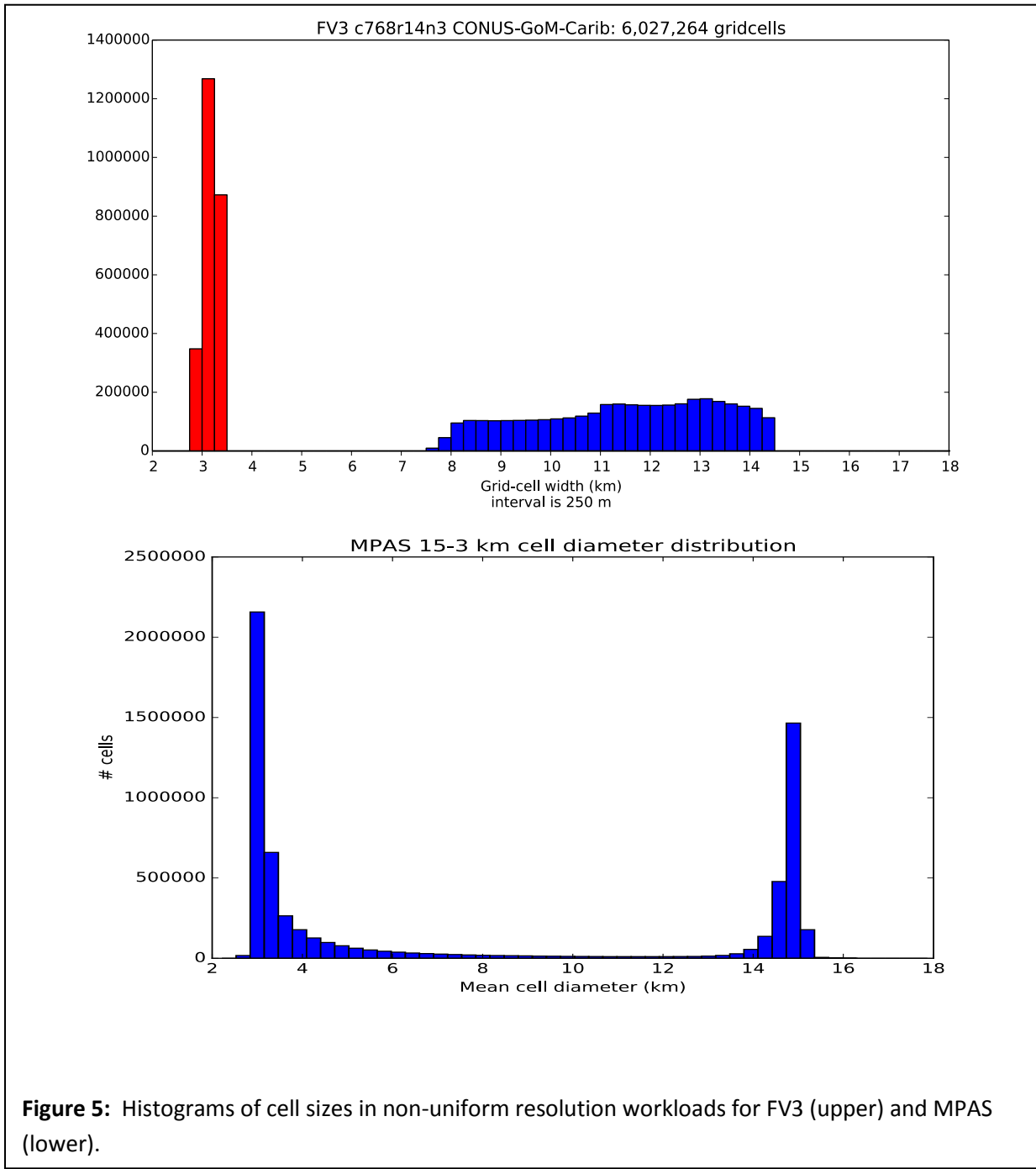
$T_{uniform}$ ← measured time for uniform 3 km resolution run

$T_{refined}$ ← measured time for non-uniform resolution run

$$E = S_{measured} / S_{ideal}$$

	FV3	MPAS
a_g (global domain area m^2)	5.101E+14	5.101E+14
a_h (high res area m^2)	2.52E+13	2.82E+13
$r = a_h/a_g$ (fraction of domain in high res)	0.0494	0.0553
dx low	14	15
dx high	3	3
dx_l / dx_h	4.67	5.00
$(dx_l / dx_h)^3$	101.63	125.00
$C_{uniform}$ (ideal)	101.63	125.00
$C_{refined}$ (ideal)	5.97	7.86
S_{ideal} , speedup from refinement	17.02	15.91
$T_{uniform}$ (measured)	345.93	344.65
$T_{refined}$ (measured)	20.98	34.10
$S_{measured}$, speedup from refinement	16.49	10.11
Efficiency	96.9%	63.5%

Figure 4: Definition of nesting efficiency and calculation using measured speed of non-uniform domain (nested or mesh-refined) domain and speed for a globally-uniform 3 km domain. The FV3 uniform and non-uniform resolution runs used 3072 processor cores. The MPAS uniform and non-uniform runs used 8192 processor cores.



Appendix E: Effective Resolution

Effective and Equivalent Resolution including Grid Considerations: Analysis

Dynamical core Test Group (DTG) Consultants

June 21, 2016

The resolution of a model is often stated as the size of the grid cells. However, the ability to resolve a feature, fully, requires several grid cells to represent the feature. For example, a wave of wavelength 100 km is much better represented by 10 cells of 10 km size than by two cells of 50 km. It is unrepresented by a single cell of 100 km. One definition of the effective resolution of a dynamical core is the smallest spatial scale that is fully resolved by the method. What is meant by “fully resolved” will generally depend on the context, but this spatial scale is usually significantly larger than the size of a grid cell. Effective resolution is often expressed in terms of an integer number of grid cells of size Δx . Many modern numerical methods are said to have an effective resolution of, approximately, $6\Delta x$ - $8\Delta x$.

The importance of effective resolution can be exemplified by assuming that there are two methods. If Method A has an effective resolution of $4\Delta x$ and Method B has an effective resolution of $8\Delta x$ then Method B requires twice as many points, hence a smaller grid cell, to represent the same geometrical feature. Presume that Method A achieves its $4\Delta x$ effective resolution through a more complex calculation that requires more computer time. Then, the evaluation of the two methods should examine the additional calculation per grid cell cost for Method A versus the need for more grid cells for Method B.

Equivalent resolution is a useful concept for expressing this idea. Equivalent resolutions of Methods A and B occur at the geometric resolutions of the two schemes where their respective effective resolutions match. Since the cost of a numerical method typically scales like $1/\Delta x$ to the third or fourth power, an advantage of around 20-25% in effective resolution would be enough to compensate for a method being twice as expensive as its competitor at the same Δx .

Effective resolution is difficult to define precisely and to quantify in practice. It will generally depend on the flow regime and the types of features to be simulated, so it is important to look at a variety of measures.

Linear wave dispersion analysis gives information on how accurately wave propagation characteristics such as phase velocity and group velocity are captured for different wavelengths. Shallow water versions of both the MPAS and FV3 schemes have been analyzed in the literature (Thuburn, 2008; Skamarock, 2008). MPAS uses a hexagonal version of the C-grid. Gravity wave phase speeds are captured quite accurately down to the shortest resolvable wavelength. However, it does support a set of small-scale ‘computational’ Rossby modes. DTG tests, as well as other investigations (e.g., Thuburn et al., 2014; Weller, 2012), have not revealed any adverse effects of these computational Rossby modes.

The FV3 scheme uses a D-grid placement of prognostic variables. C-grid-located horizontal winds are constructed diagnostically during the time step, but to leading order the wave dispersion properties are those of a D-grid: short wavelengths ($<4\Delta x$) are significantly slowed and there are $2\Delta x$ computational modes that do not propagate. However, the FV3 time integration scheme provides some selective

damping of the shortest gravity waves, which is supplemented by other explicit damping terms. These are enough to control the behavior of the scheme. DTG tests have not revealed any poor behavior that could be attributed to the use of a D-grid.

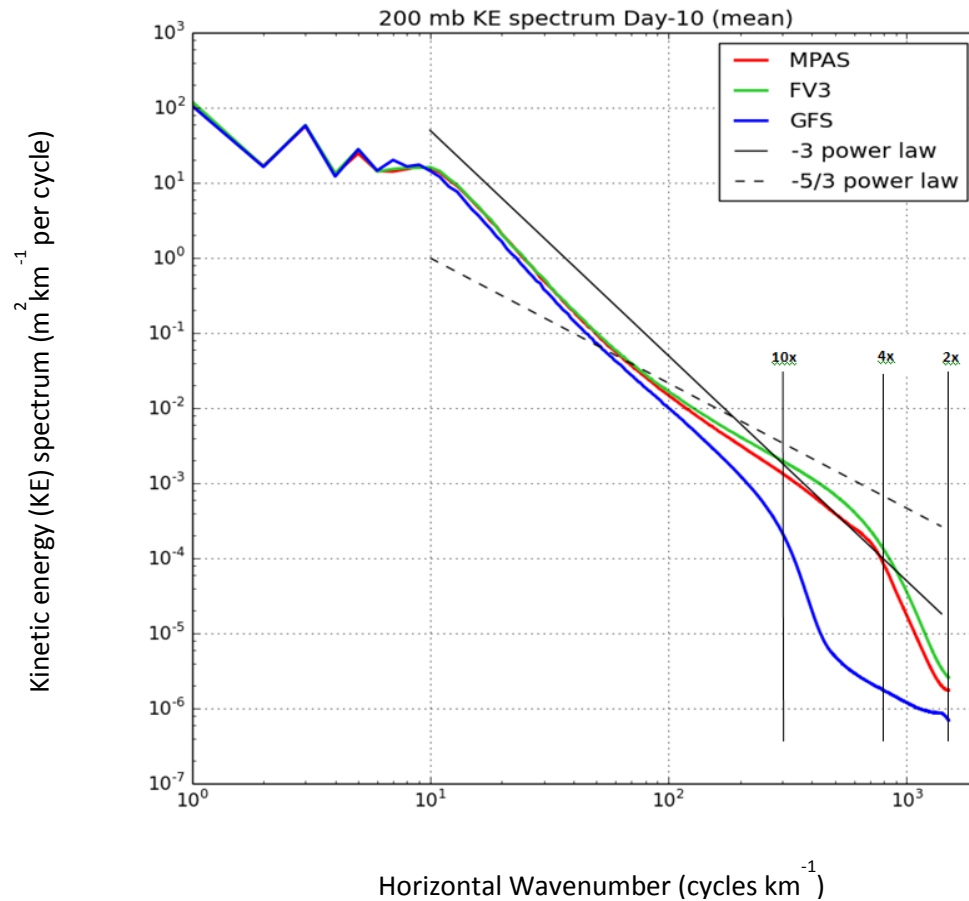
It is possible for constant velocity tracer advection to define effective resolution in terms of the ability of the method to transport a pure wave of a given wavelength (Kent et al., 2014). The quality of the method can be described in terms of errors in phase speed and wave amplitude. Direct measures of error can be calculated for a given amount of time. As a rule of thumb for credible numerical methods, features of $2\Delta x$ are unresolved and features of $10\Delta x$ are resolved. Formally, however, depending upon the measure that defines resolved, it might take 20 or more grid cells to represent a feature. Evaluation of the effective resolution of FV3 and MPAS directly in this way was not tried. However, the conservation test and other tests confirm that both dycores have accurate advection schemes.

This simple linear analysis of wave propagation and advection is a useful starting point for consideration of MPAS and FV3. However, effective resolution is *not* uniquely defined, and hence, measurement is open to interpretation and controversy. The complexity of MPAS and FV3, for example, the use of flux limiters, complex grid-scale management of physical-numerical behavior, etc., complicate the definition of effective resolution, especially for the complex flows that the schemes are expected to represent. Coupling of the dynamical core with parameterized physical parameterizations makes simple quantitative measures of effective resolution far more difficult.

Consideration and evaluation of effective resolution are longstanding practices in meteorology (e.g., Kreiss and Olinger, 1972). Effective resolution and equivalent resolution have been examined for idealized geophysical flows (e.g., Jablonowski and Williamson, 2006). Methods to document effective and equivalent resolution have been explored for dynamical cores coupled with sub-grid physical parameterizations (e.g., Williamson, 2008). One evaluation approach is to examine statistical properties relative to reference solutions. Reference solutions are often high resolution, nominally converged, solutions from the same method as being evaluated. Evaluation criteria, also, include measures, some qualitative, of how well fronts and closed vortices are represented. Effective resolution is a bulk measure of diffusivity. Non-physical diffusion appears in numerical advection methods. Numerical diffusion can be intrinsic to a numerical method. Diffusive filters are also frequently added to models to manage instabilities and contaminating noise. Both the MPAS and FV3 numerical methods have implicit diffusion, supplemented by explicit diffusive filters. As well, both MPAS and FV3 have made significant efforts to manage and minimize non-physical diffusion.

A primary metric used by the DTG to consider effective and equivalent resolution is the method suggested by Skamarock (2004) to evaluate models with both dynamics and physics. This approach examines the deviations of the kinetic energy spectra from a reference power law. The reference power law is based on observations and theory-based interpretations of those observations. The figure shows an example of the kinetic energy (KE) spectrum at 200hPa from MPAS, FV3, and GFS. The plot is from a Test 3 simulation – the 13 km global forecasts. Both FV3 and MPAS capture a shallowing of the spectrum for wavelengths shorter than a few hundred km, similar to observations, and the wavelength at which spectra steepen rapidly suggests an effective resolution of around $4-6\Delta x$. GFS, on the other hand, does not capture the mesoscale shallowing of the spectrum and has an effective resolution of around $10\Delta x$. We note that this approach combines the role that physics will play on the dycore's effective resolution. We also note that spectra are not determined by a unique set of physical and numerical attributes, and that realistic spectra can be obtained for incorrect reasons (For example,

insufficient vertical resolution can cause an apparently realistic but spurious shallowing of the spectrum; Waite 2015.). Therefore kinetic energy spectra enter as an informative metric; however, spectra are not uniquely definitive. It is noted that amongst current operational weather forecasts models, e.g. ECMWF and NWS GFS, excellent weather forecasts are realized even though there are large deviations from the reference power law.



In addition to the examination of the kinetic energy spectra, the DTG performed numerous tests at different resolutions. These tests were the tropical cyclone and supercell tests from Phase 1 of the testing and single realizations of Phase 2, Test 4, global forecasts using physics for the GFS. The tests with the GFS physics determined full-system computational performance at three resolutions: a reference resolution, plus an incremental increase and decrease of the reference resolution. The entire suite of forecasts at the reference resolution was evaluated relative to verifying analyses.

The results of most of these tests suggest that the effective resolutions of MPAS and FV3 are very similar. For example, marginally resolved features such as fronts in the baroclinic wave test and updrafts in the supercell test appear equally well resolved by the two dycores.

In the high-resolution region (nominally $\Delta x = 3$ km) of the variable-resolution test case, there is some indication that convective storms have slightly smaller horizontal scale in the MPAS solutions than in the

FV3 solutions, and this is supported by autocorrelation length scale diagnostics. Careful examination of the grid densities near the storm location showed that the MPAS grid was, in fact, slightly more dense than the FV3 grid, and that this could account for some of the difference in storm scale. To test the hypothesis that the diffusion in the two dycores might explain the remaining difference, the supercell test – a more carefully controlled test setup - was run at 3 km resolution using the same diffusion settings as in the variable-resolution test. However, this test showed no significant difference between the dycores in the scale of the simulated updrafts.

In summary, the results of the DTG's experiments do not reveal any definitive effective-resolution advantage of either MPAS or FV3. Looking across tests and diagnostics, some narrow examples could be made in favor of each method. Collectively, including the kinetic energy spectra, the two methods would be viewed as having an effective resolution of, approximately, $6\Delta x$, at a 13 km grid scale.

References

- Jablonowski, C., and Williamson, D. L., 2006: A baroclinic instability test case for atmospheric model dynamical cores, *Q. J. R. Meteorol. Soc.*, 132, 2943-2975.
- Kent, J., Whitehead, J. P., Jablonowski, C., and Rood, R. B., 2014: Determining the effective resolution of advection schemes. Part 1: Dispersion analysis, *J. Comput. Phys.*, 278, 485-496.
- Kreiss, H.O., and Oliger, J., 1972: Comparison of accurate methods for the integration of hyperbolic equations, *Tellus*, 24, 199-215.
- Skamarock, W. C., 2008: A linear analysis of the NCAR CCSM Finite-Volume dynamical core, *Mon. Wea. Rev.*, 2112-2119.
- Thuburn, J., 2008: Numerical wave propagation on the hexagonal C-grid, *J. Comput. Phys.*, 227, 5836-5858.
- Thuburn, J., Cotter, C. J., and Dubos, T., 2014: A mimetic, semi-implicit, forward-in-time, finite volume shallow water model: comparison of hexagonal-icosahedral and cubed sphere grids, *Geosci. Model Dev.*, 7, 909-929.
- Waite, M. L., 2015: Dependence of model energy spectra on vertical resolution, *Mon. Wea. Rev.*, 144, 1407-1421.
- Weller, H., 2012: Controlling the computational modes of the arbitrarily structured C-grid, *Mon. Wea. Rev.*, 140, 2734-2755.
- Williamson, D. L., 2008: Equivalent finite volume and Eulerian spectral transform horizontal resolutions established from aqua-planet simulations, *Tellus*, 60A, 839-847.

Appendix F: Vertical Coordinate Analysis

Dynamical core Test Group (DTG) Consultants

June 16, 2016

The choice of vertical coordinate for a model is a design decision and as with any choice there are advantages and disadvantages. For both MPAS and FV3, the vertical coordinates have been well tested and work well at both global and mesoscale resolutions.

The vertical coordinates used in MPAS and FV3 are different:

MPAS uses a hybrid height (z) coordinate. In the lower atmosphere height is normalized by topography so that the coordinate follows the Earth's surface at the resolution of topography. Above the surface, there is a transition to a z coordinate. The vertical coordinate is constant with time.

FV3 uses a hybrid pressure (p) coordinate. In the lower atmosphere pressure is normalized by surface pressure (σ , $\sigma = p/p_{\text{surface}}$) so that the coordinate follows the Earth's surface at the resolution of topography. Above the surface, there is a transition to a pressure coordinate. FV3 implements the hybrid pressure coordinate as a Lagrangian coordinate so that the pressure on model levels varies with time. The model level pressure is periodically reset and a remapping procedure is used to place the model variables onto the reset pressure levels.

Both MPAS and FV3 are implemented as a terrain-following coordinate. Terrain-following coordinates are expected to face challenges at very high resolution. At high resolution, the orography steepens, and the terrain-following coordinate is no longer (quasi)-orthogonal; accuracy is lost. Hence, both dynamical cores will likely require research and development related to the vertical coordinate at cloud-resolving resolutions.

Height (z) coordinates have been widely utilized in non-hydrostatic mesoscale models, and are increasingly used in global models. Intuitively, height coordinates accommodate hydrostatic and non-hydrostatic flows without special consideration.

Pressure (p) coordinates have been widely used in global hydrostatic models. Pressure coordinates provide numerous simplifications to the equations of motion for large-scale hydrostatic flows. Non-hydrostatic flows require relaxing the hydrostatic approximation used in these equations and their implementation in dycores. The Lagrangian implementation in FV3 provides numerous advantages to advective transport and reduces spurious numerical mixing.

Within the test suite for this evaluation, the conservation test (Test 2) was motivated, largely, by the benefits realized by the use of an isentropic vertical coordinate in research models. The results of Test 2 confirm that both MPAS and FV3 conserve mass to within round-off error, and both have excellent

conservation properties for energy and entropy. Both dycores have much better conservation properties than the current operational Global Spectral Model.

Another consideration associated with the vertical coordinate is the implementation of physics, chemistry, and data assimilation. Using either vertical coordinate, these capabilities can be successfully implemented. This is well established by the use of these dycores in numerous applications. However, given the intent by NCEP to reuse existing physics, chemistry, and data assimilation algorithms, it is incumbent upon NCEP to do a cost and effort analysis for the vertical grid used in the dycore. Likewise, there are planning and design issues that should be coordinated with other NGGPS working groups.

For both MPAS and FV3, the development teams have made well-founded design decisions about the vertical coordinate. The testing by the DTG reveals that the two teams have established the fidelity and robustness of the dycores across a wide range of the scales and applications. Both dycores are likely to require research and development of their vertical coordinate as resolutions advance to cloud-resolving scales and the application suite is expanded beyond global weather prediction. The DTG testing reveals no specific scientific performance results favoring one vertical treatment over the other. Computationally, the FV3 Lagrangian approach with remapping should offer advantage over a full three-dimensional calculation of advection. Based largely on NCEP's desire to reuse physics, chemistry, and data assimilation algorithms, there are potential implementation differences.

Acronym List

3-DVAR	Three-Dimensional Variational
AA	Assistant Administrator
AVEC	Advanced Computing Evaluation Committee
CCPP	Common Community Physics Package
CICE	Los Alamos Sea Ice Model
DA	Data Assimilation
DCMIP	Dynamical Core Model Intercomparison Project
DTG	Dynamical core Test Group
Dycore	Dynamical core
ECMWF	European Center for Medium Range Weather Forecasting
EMC	Environmental Modeling Center
ESMF	Earth System Modeling Framework
ESRL	Earth System Research Laboratory
FTE	Full-time equivalent
FV3	Finite-Volume Cubed-Sphere Dynamical Core
GCWMB	Global Climate and Weather Modeling Branch
GDAS	Global Data Assimilation System
GFDL	Geophysical Fluid Dynamics Laboratory
GFS	Global Forecast System
GFS-NH	Non-Hydrostatic extension of Semi-Lagrangian Spectral model
GMTB	Global Modeling Test Bed
GPS-RO	Global Positioning System-Radio Occultation
GSM	Global Spectral Model
HPC	High Performance Computing
HIWPP	High Impact Weather Prediction Project
HWRF	Hurricane Weather Research and Forecasting model
MOM	Modular Ocean Model
MPAS	Model for Prediction Across Scales
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NCO	NCEP Central Operations
NEMS	NOAA Environmental Modeling System
NEMSIO	NEMS standardized I/O format
NEPTUNE	Navy's Environmental Prediction System Using the NUMA core
NIM	Non-hydrostatic Icosahedral Model
NGGPS	Next Generation Global Prediction System
NMMB	Non-hydrostatic Multi-scale Model on B-grid
NMM-UJ	Non-hydrostatic Mesoscale Unified Jacobian Model
NOAA	National Oceanic and Atmospheric Administration

NRL	Naval Research Laboratory
NUMA	Non-hydrostatic Unified Model
NUOPC	National Unified Operational Prediction Capability
NWP	Numerical Weather Prediction
NWS	National Weather Service
OAR	Office of Oceanic & Atmospheric Research
OMB	Office of Management and Budget
OST	Office of Science and Technology
R2O	Research to Operations
RTMA	Real-Time Mesoscale Analysis
SST	Sea Surface Temperature
SON	September-October-November
SWPC	Space Weather Prediction Center
TRMM	Tropical Rainfall Measuring Mission
UPP	NCEP Unified Post-Processing software
WAM	Whole Atmosphere Model