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, dynamical core 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)[1]. 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 dynamical cores' software implementations were also evaluated for suitability on next-generation HPC architectures (Criterion #10).[2]

For Criterion #4, we 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, we measured how efficiently each dynamical core 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.

[1] https://www.nersc.gov
[2] 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 1-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). 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. This report is the consensus of the remaining members of AVEC.

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 tests[3] 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.[4] 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

[3] https://michalakes.svn.cloudforge.com/rrtmmic/avec_timer
[4] http://www.esrl.noaa.gov/gsd/ato/AVECPhase-2Benchmarks20160428_adjusted.pdf

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 we believe it does not impact the finding that FV3 was more efficient than MPAS with additional tracers.

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, our 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.[5,6]

- MPAS
  - MPAS grids are defined and decomposed over processors using off-line grid generation software that has not been parallelized. Generating and

[5] http://www.esrl.noaa.gov/gsd/ato/FV3_Analysis-final.pdf
[6] http://www.esrl.noaa.gov/gsd/ato/MPAS-Analysis-final.pdf

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.

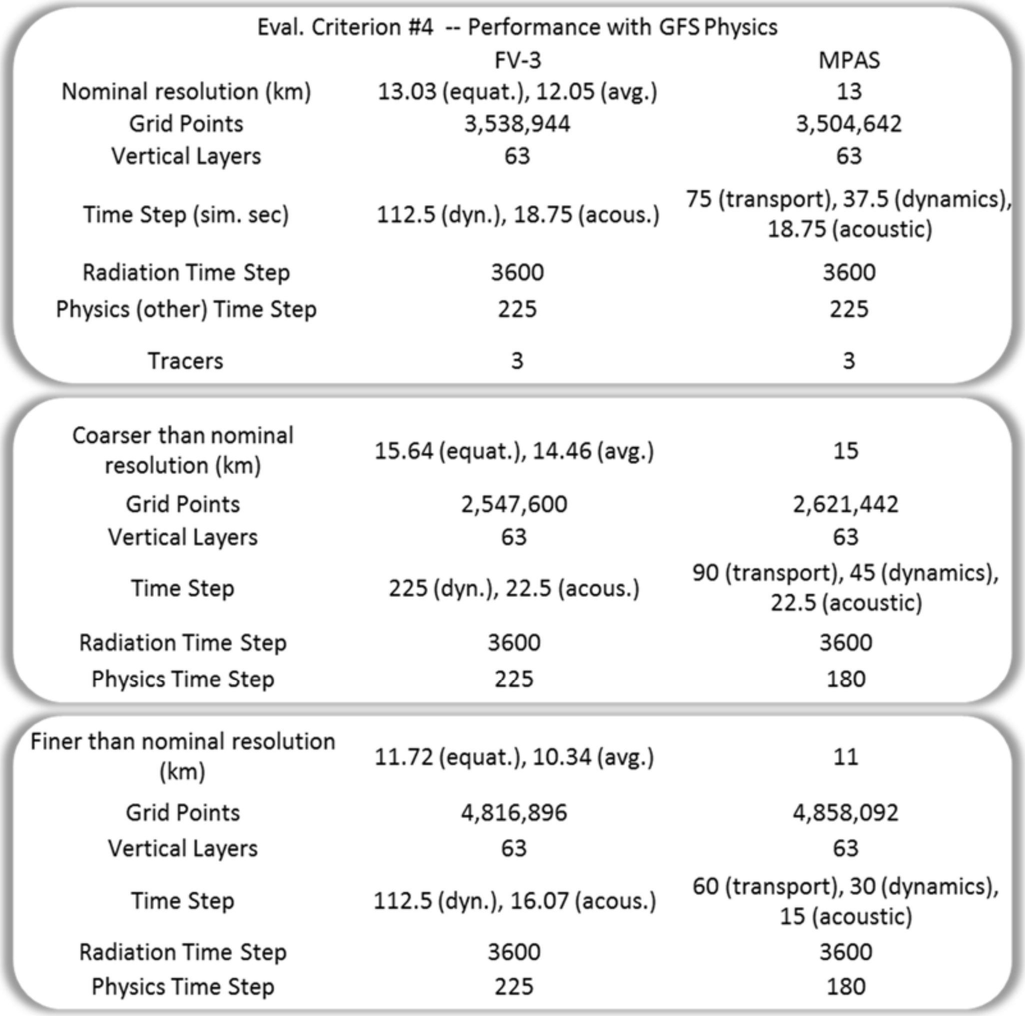
- FV3
  - The cost of vertical remapping, while small in the workloads evaluated by ESRL, can become significant if vertical remapping needs to be called more frequently for different configurations. Certain inefficiencies relating to loop nesting and data organization in the vertical remapping were 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.

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

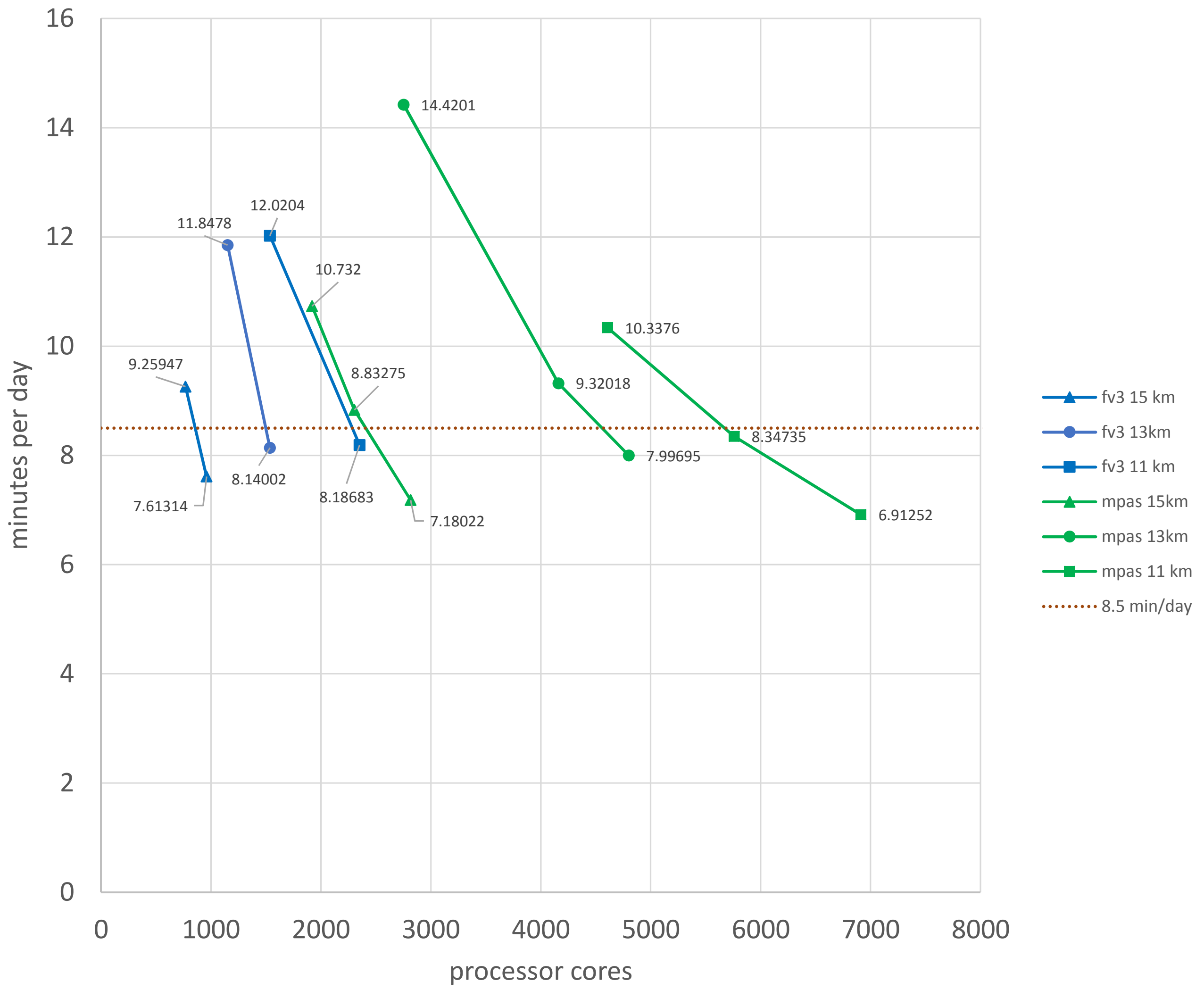


Figure 1.a. Speed in minutes per day as a function of number of processor cores on Cori. Dotted horizontal line indicates operational speed requirement of 8.5 minutes per forecast day. The intersection with the plotted lines is used to estimate the number of processor cores required to meet the operational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

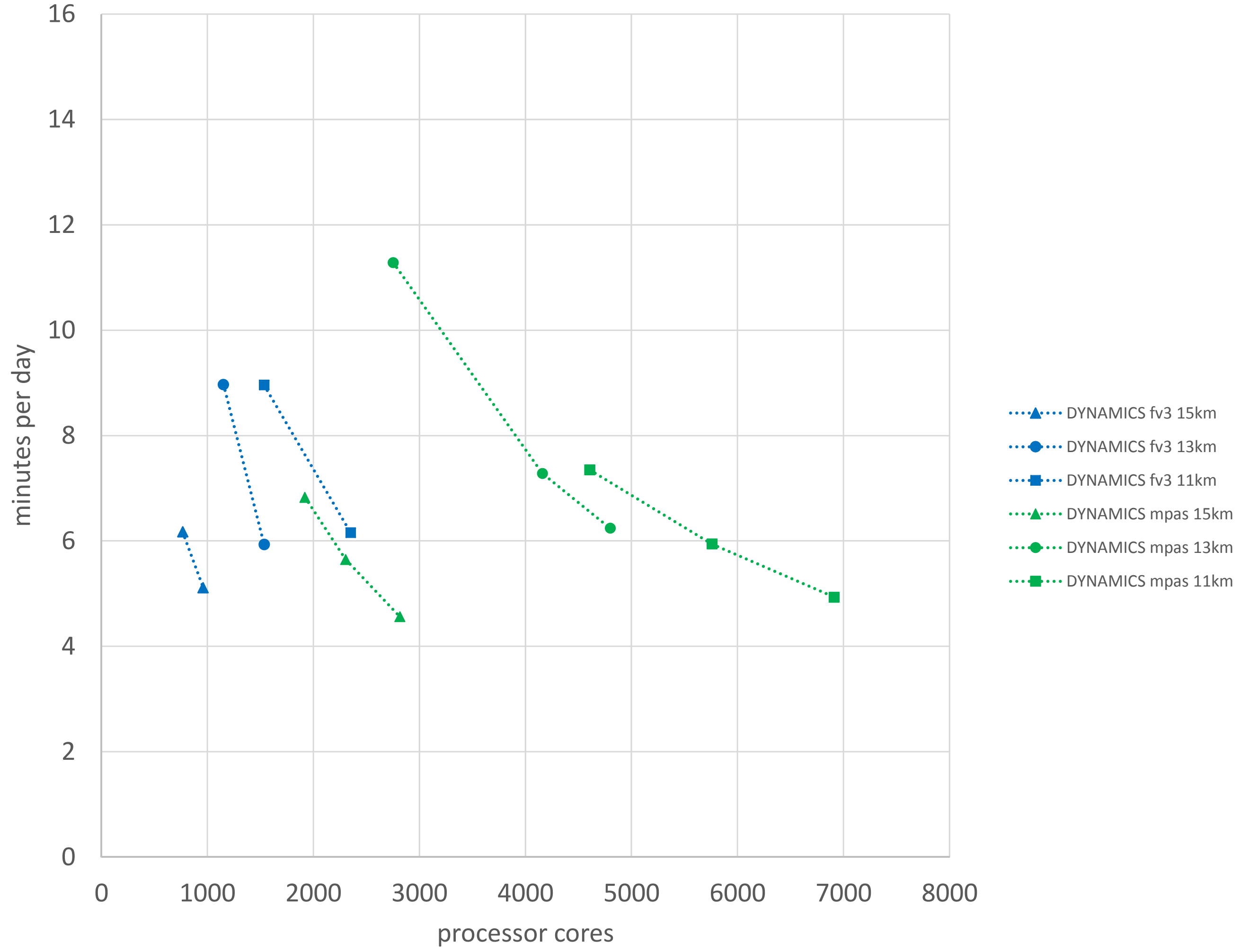


Figure 1.b. Speed of dynamical core in minutes of machine time per day, with the time for physics removed, as a function of processor cores on Cori.

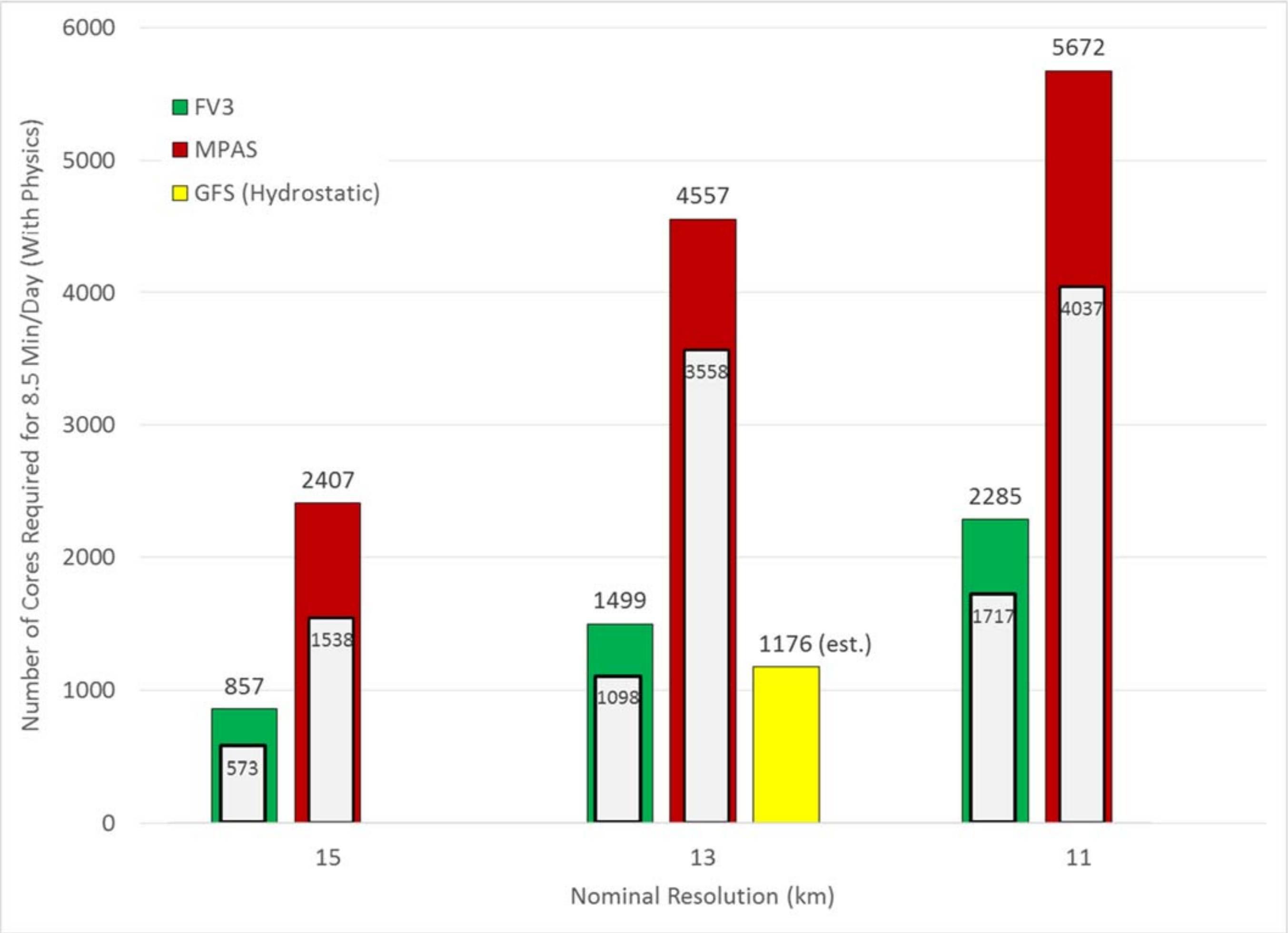


Figure 2: Cores required to meeting 8.5 minutes per day forecast speed requirement for operations at 15, 13, and 11 km horizontal resolution. All cases used 63 vertical levels. Colored bars show time with GFS physics; insets show the fraction of cores required by the dycore alone. The estimated number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

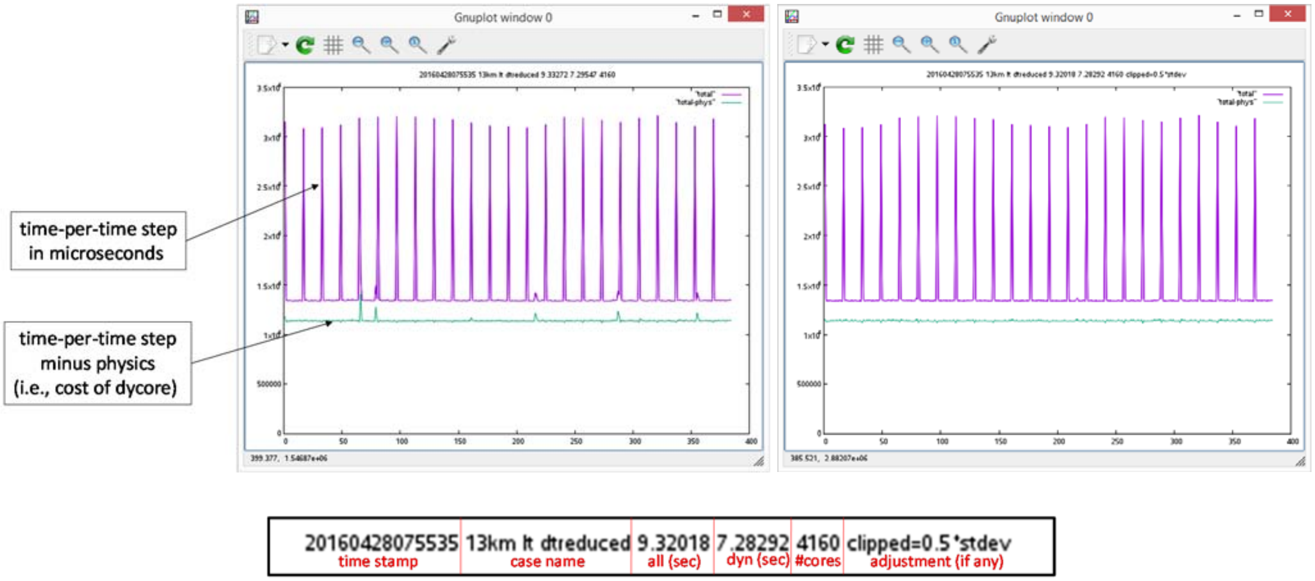
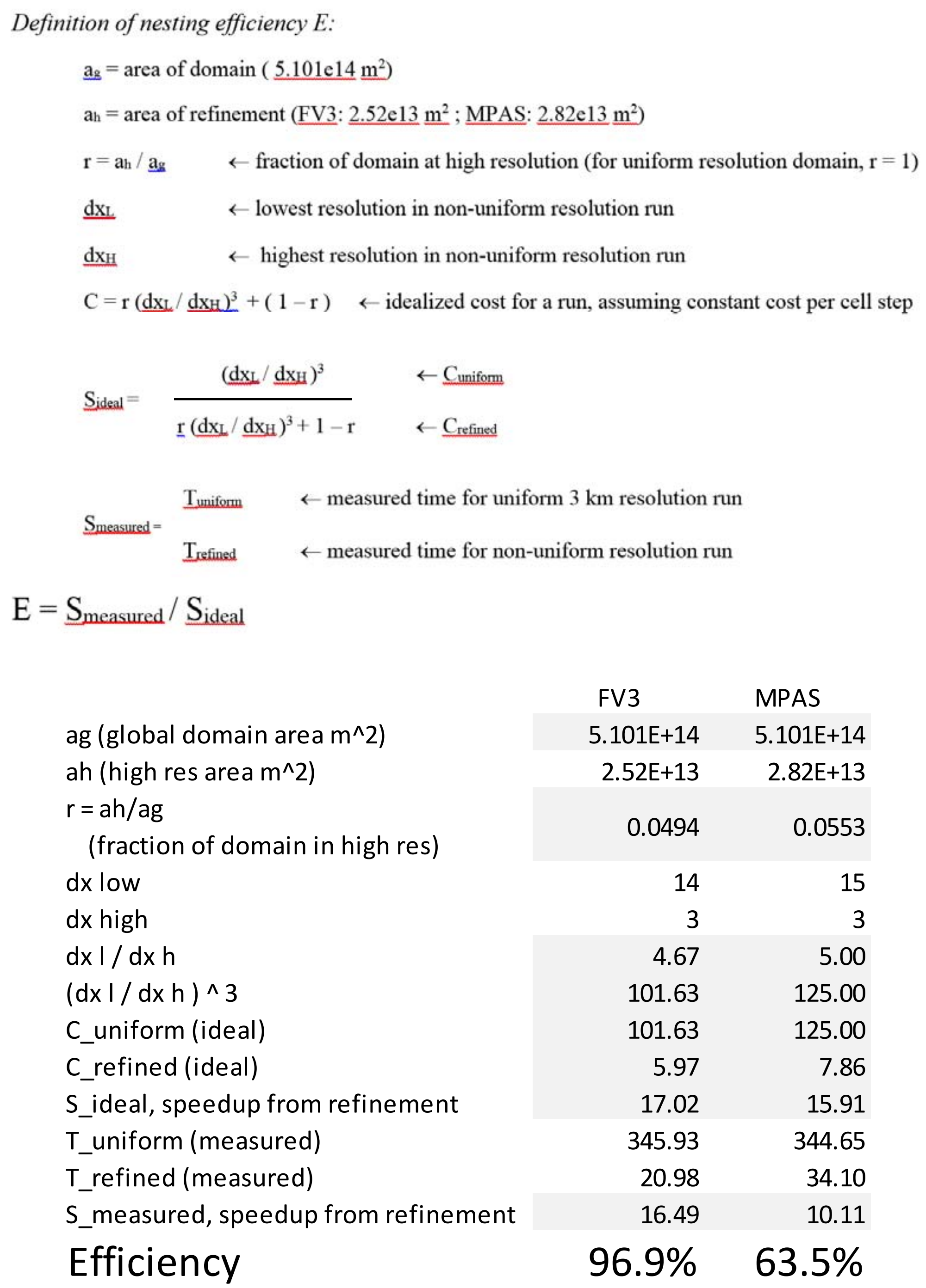


Figure 3: Data from AVEC timers before (left) and after (right) filtering. The total time per time step includes the cost of GFS physics, spikes in which are from half-hourly calls to radiation. The time-per-time step for the dycore itself is plotted separately. Occasional small-scale perturbations in 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 m^2)

a_h = area of refinement (FV3: 2.52e13 m^2 ; MPAS: 2.82e13 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

$$S_{ideal} = \frac{(dx_L / dx_H)^3 \quad \leftarrow C_{uniform}}{r\,(dx_L / dx_H)^3 + 1 - r \quad \leftarrow C_{refined}}$$

$$S_{measured} = \frac{T_{uniform} \quad \leftarrow \text{measured time for uniform 3 km resolution run}}{T_{refined} \quad \leftarrow \text{measured time for non-uniform resolution run}}$$

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

	FV3	MPAS
ag (global domain area m^2)	5.101E+14	5.101E+14
ah (high res area m^2)	2.52E+13	2.82E+13
r = ah/ag (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.

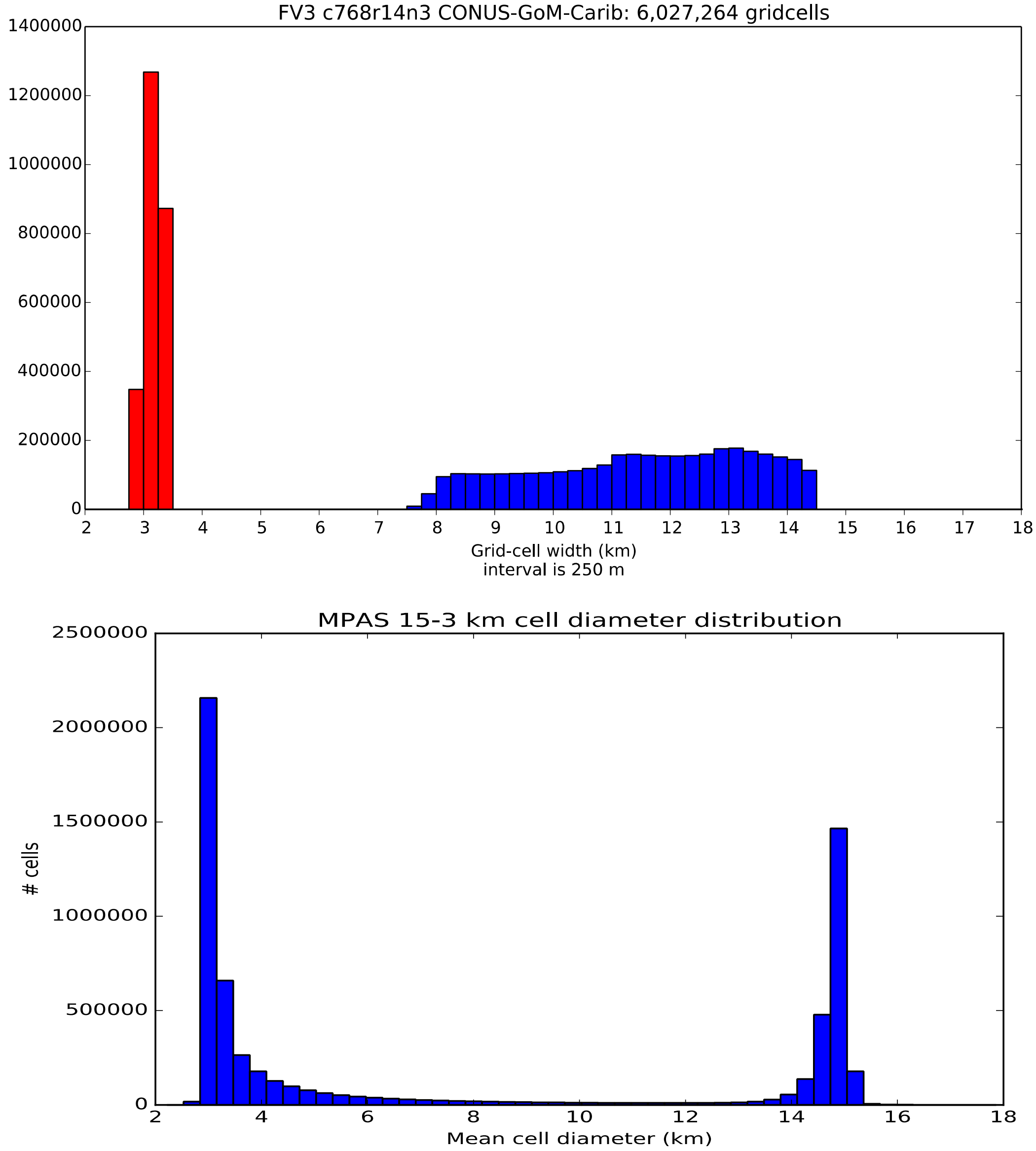


Figure 5: Histograms of cell sizes in non-uniform resolution workloads for FV3 (upper) and MPAS (lower).