Characterizing Differences in Model Output Induced by Changes in High Performance Computing Platform

Christopher W. Harrop[1,2], Ligia Bernardet[1,2]

Cooperative Institute for Research in Environmental Sciences[1],
NOAA - Earth System Research Laboratory - Global Systems Division[2]

Introduction

Scientists and their collaborators often need to use multiple High Performance Computing (HPC) systems to conduct their research and perform numerical modeling experiments. These various HPC systems are usually hosted by different institutions and may be comprised of vastly different hardware and software development environments. Because simulations generally do not produce identical results on different HPC platforms, many scientists question the validity of simulation experiments and comparisons of model results that span multiple HPC systems. To address this concern we are investigating the differences in model output that arise solely from changes to the HPC platform. Our goal is to characterize those differences to provide diagnostic information that scientists can use to gauge whether or not their model is working when ported to a new platform.

Experiment Design

We ran three numerical models on NOAA's three largest R&D HPC systems. Two of the models are large, complex, numerical weather prediction (NWP) codes. The other model is a small, simple, code that simulates the motion of a double pendulum in two dimensions. For a given model, the same configuration and initial conditions were used on all platforms. The two NWP models were initialized at the same set of times, though the two models did not share initial conditions.

The Numerical Models

WRF NMM	FIM	Double Pendulum
V3.2.1	Revision 1278	Fortran 90 4th order Runge-Kutta solver adapted from [1]
13 Km CONUS domain	30 Km Global domain	Point masses with rigid, massless, rods and no friction
21 cases, Jan 2009	20 cases, Jan 2009	1 non-linear case
48 hour forecast	120 hour forecast	60 second forecast

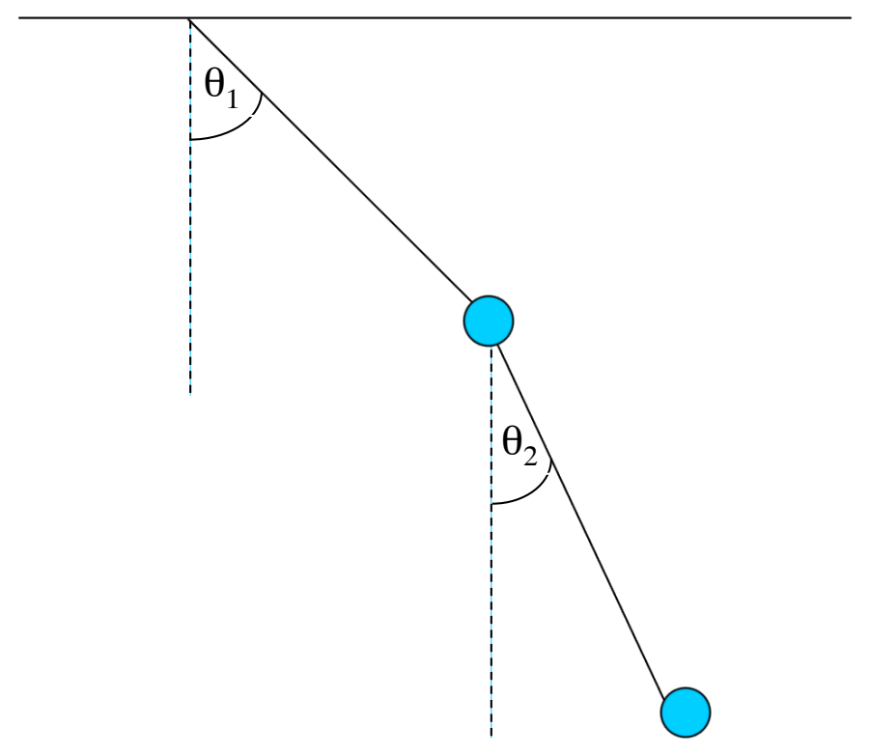
The HPC Hardware

Jet	Vapor	Gaea
ESRL	NCEP	GFDL
Linux	AIX	Cray Linux Environment (CLE)
Intel Nehalem x86_64 cluster	IBM Power6 P575 cluster	AMD Mangy-Cours Cray XT6

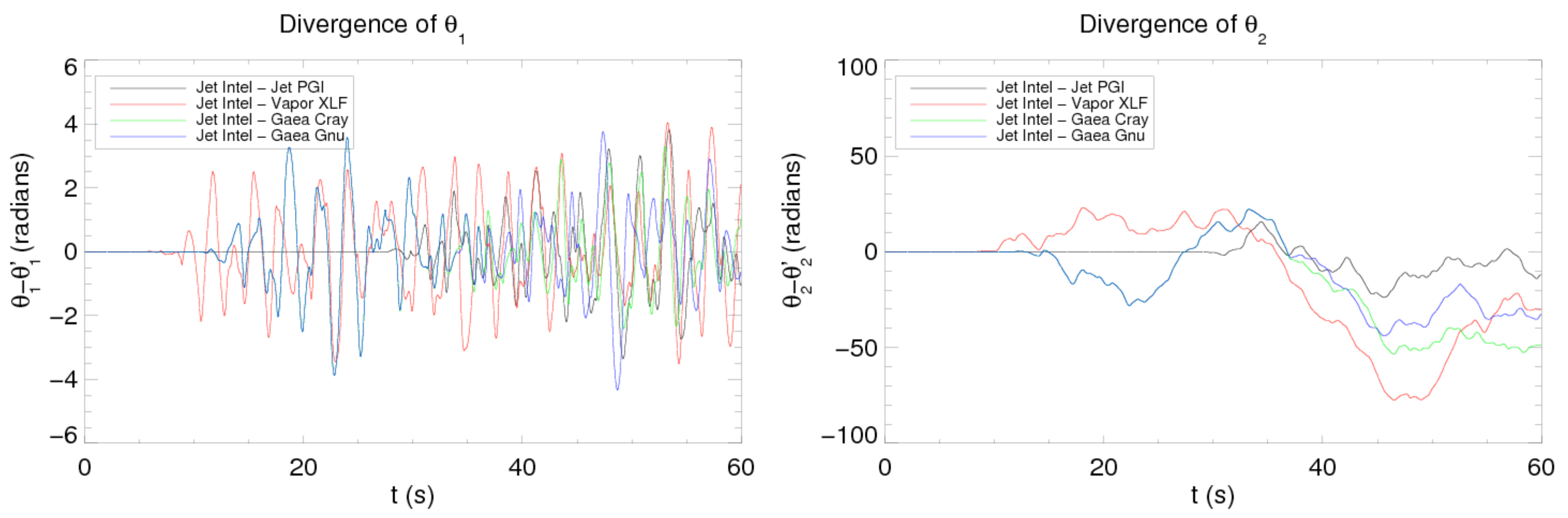
The HPC Software

Compiler	Jet	Vapor	Gaea
Intel 11.1	X		X
PGI 10.6	X		X
XLF 12.1		X	

Divergence of Output With a Double Pendulum



The differential equations of motion for a double pendulum can be solved using a 4th order Runge-Kutta numerical method. The simplicity of both the equations and the numerical method means that the motion of a double pendulum can be simulated with less than 100 lines of Fortran. This means there is much less chance that differences in output are caused by coding errors rather than platform differences. When this simulation is run with initial conditions that result in non-linear behavior, the solutions on various platforms diverge radically, even when double precision is used, all compiler optimization is turned off, and the simulation time step is very small.

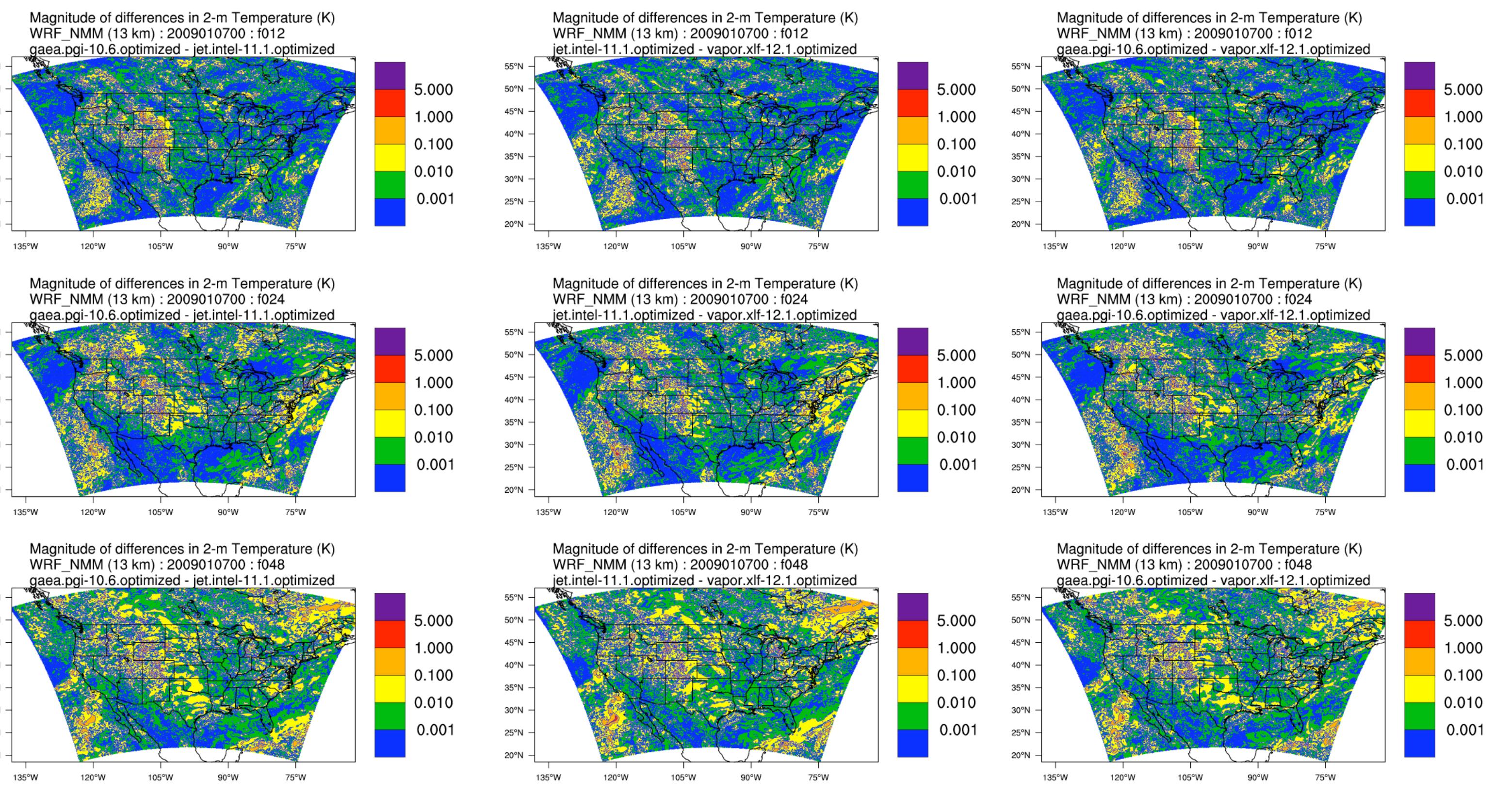


The plots above demonstrate that simulations of non-linear physical systems can indeed yield wildly different results when run on different hardware, or even when compiled with different compilers on the same hardware. The solutions diverge at different times during the simulation, depending on the platforms being compared.

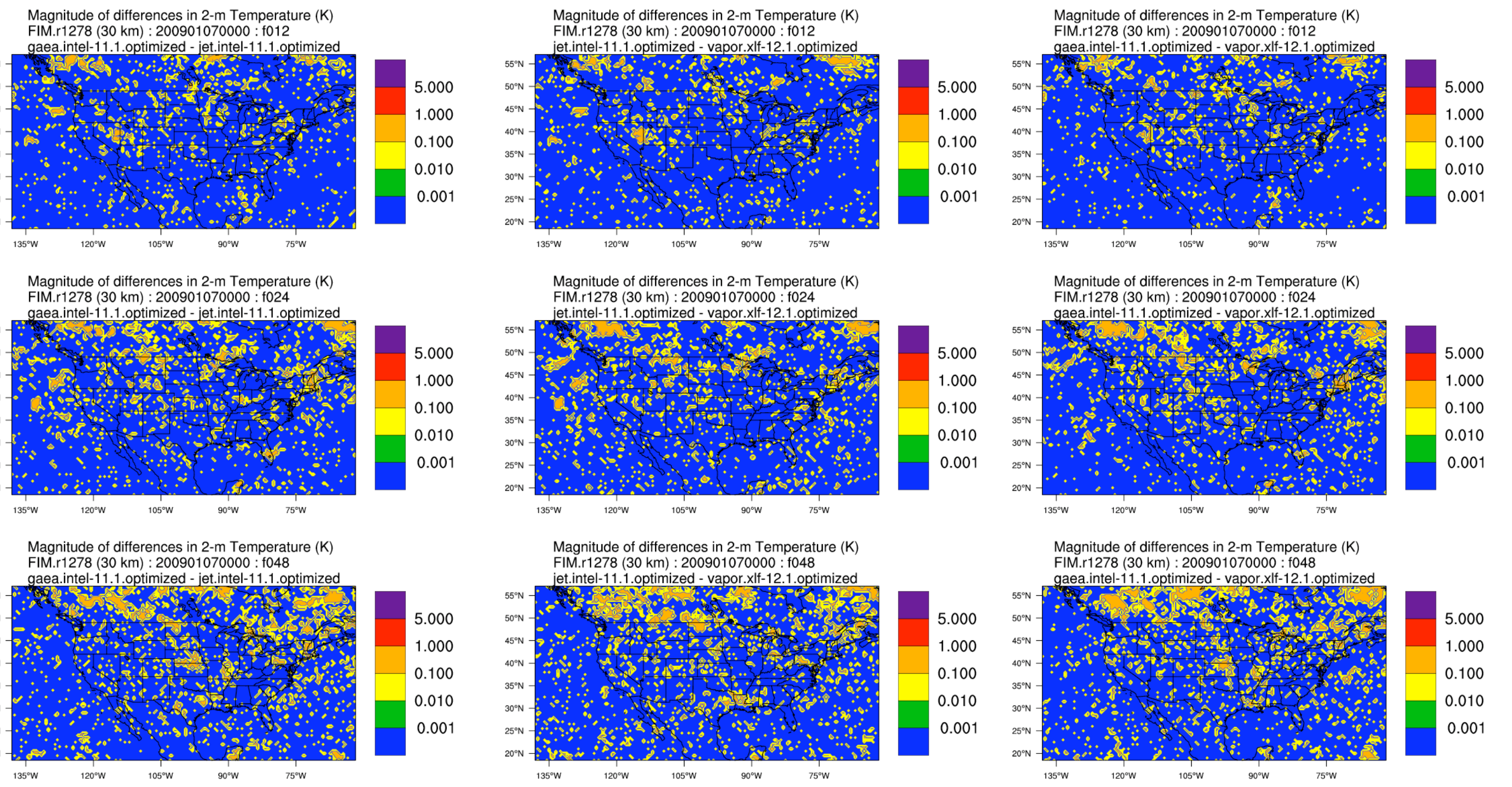
Divergence of Model Forecasts

Computations of differences in model output is a complex problem. To keep things simple for our initial investigation, we restricted our examination of model results to the 2m temperature field and calculated difference fields by computing point-wise differences. In our analyses we computed 2m temperature difference fields for all permutations of platforms. This was done for each case that was run and for each forecast lead time.

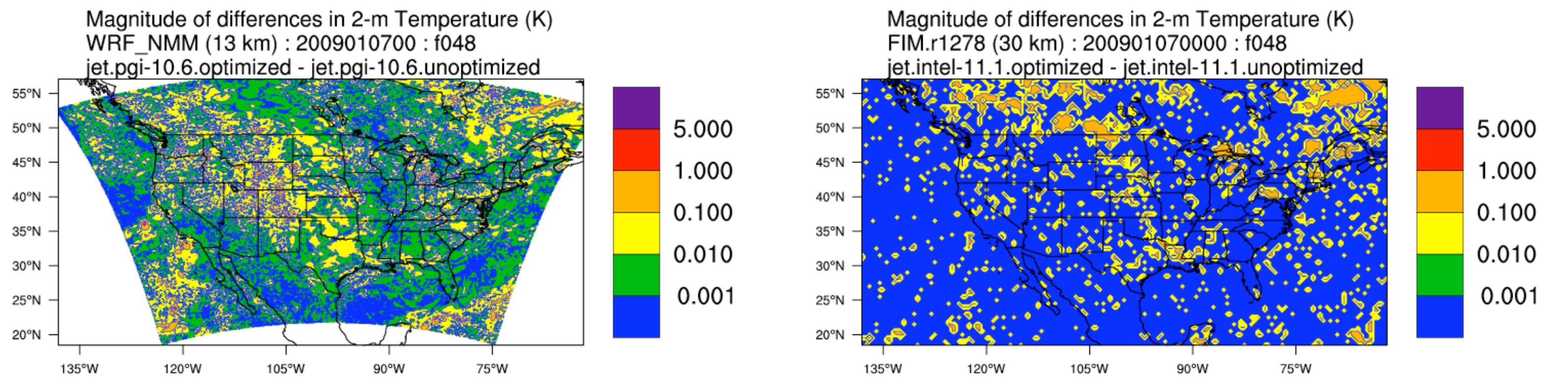
Typical WRF NMM Differences for Different Hardware and Software


Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f012
gaea.pgi-10.6.optimized - jet.intel-11.1.optimized
Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f012
jet.intel-11.1.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f012
gaea.pgi-10.6.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f024
gaea.pgi-10.6.optimized - jet.intel-11.1.optimized
Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f024
jet.intel-11.1.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f024
gaea.pgi-10.6.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f048
gaea.pgi-10.6.optimized - jet.intel-11.1.optimized
Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f048
jet.intel-11.1.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f048
gaea.pgi-10.6.optimized - vapor.xlf-12.1.optimized
5.000
1.000
0.100
0.010
0.001
55°N
50°N
45°N
40°N
35°N
30°N
25°N
20°N
135°W
120°W
105°W
90°W
75°W

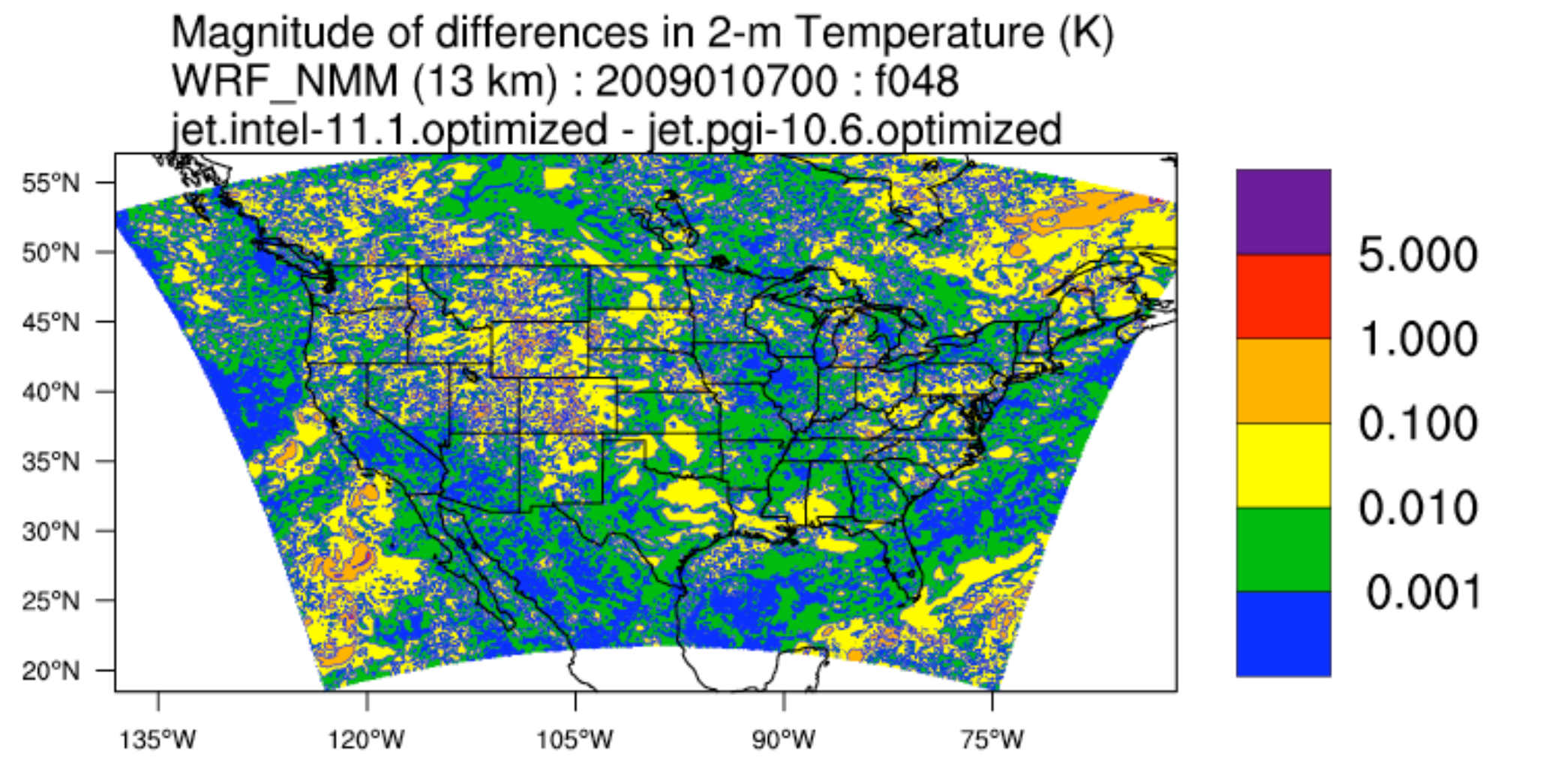
Typical FIM Differences for Different Hardware and Software


Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f012
gaea.intel-11.1.optimized - jet.intel-11.1.optimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f012
jet.intel-11.1.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f012
gaea.intel-11.1.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f024
gaea.intel-11.1.optimized - jet.intel-11.1.optimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f024
jet.intel-11.1.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f024
gaea.intel-11.1.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f048
gaea.intel-11.1.optimized - jet.intel-11.1.optimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f048
jet.intel-11.1.optimized - vapor.xlf-12.1.optimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f048
gaea.intel-11.1.optimized - vapor.xlf-12.1.optimized
5.000
1.000
0.100
0.010
0.001

Differences for Different Optimization Levels on the Same Hardware

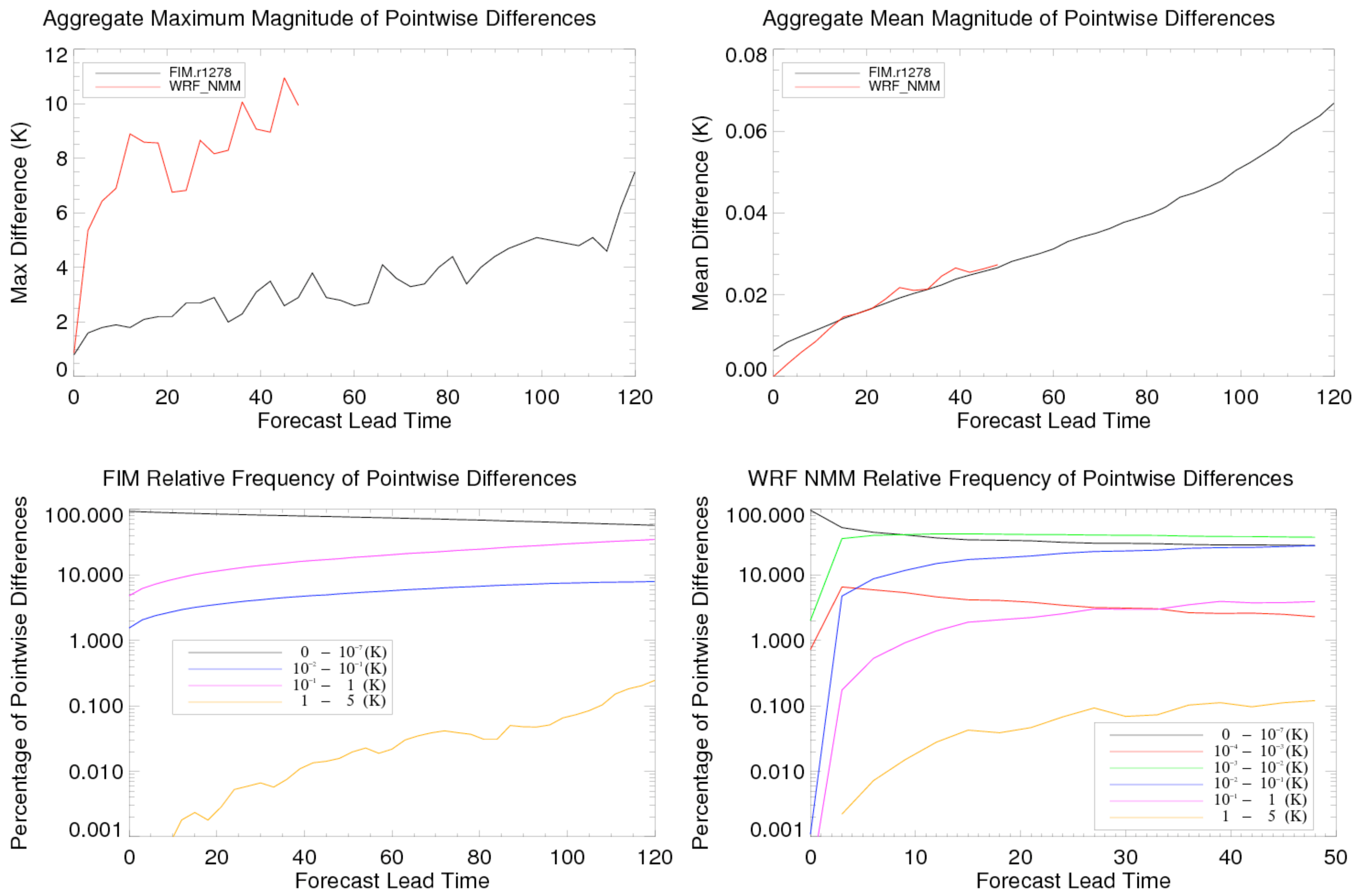

Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f048
jet.pgi-10.6.optimized - jet.pgi-10.6.unoptimized
Magnitude of differences in 2-m Temperature (K)
FIM.r1278 (30 km) : 200901070000 : f048
jet.intel-11.1.optimized - jet.intel-11.1.unoptimized
5.000
1.000
0.100
0.010
0.001
55°N
50°N
45°N
40°N
35°N
30°N
25°N
20°N
135°W
120°W
105°W
90°W
75°W

Differences for Different Compilers on the Same Hardware


Magnitude of differences in 2-m Temperature (K)
WRF_NMM (13 km) : 2009010700 : f048
jet.intel-11.1.optimized - jet.pgi-10.6.optimized
5.000
1.000
0.100
0.010
0.001
55°N
50°N
45°N
40°N
35°N
30°N
25°N
20°N
135°W
120°W
105°W
90°W
75°W

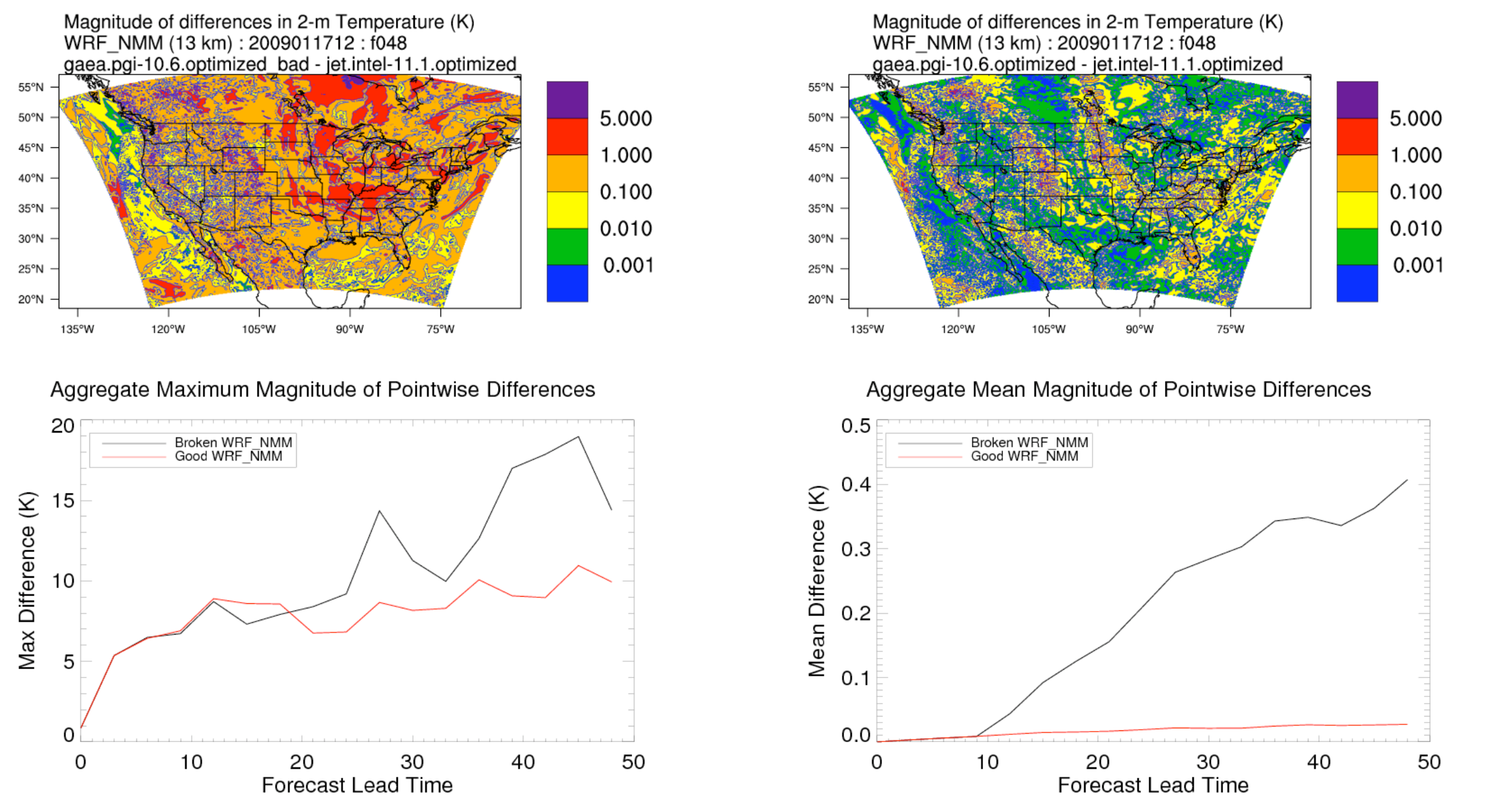
Characterizing Model Output Differences

The differences in model output that arise for our runs on various platforms and test cases appear to have roughly the same magnitude for a given forecast lead time. Interestingly, the relative positions and shapes of the differences also appear to be roughly the same. Plots for all our other initialization times look very similar to the ones shown. We also collected statistics to further characterize the resulting differences.



A Forecast Gone Bad

During our experiment we accidentally produced bad results on Gaea. The plot below on the left shows the difference between a bad run on Gaea and that same case run on Jet. The plot on the right shows the differences for that same case after a problem with missing boundary conditions was corrected. The bad runs are easily identified by looking at the entire set of plots. Plots of the aggregate maximum and mean point-wise differences also show a huge difference between the bad Gaea results and the other runs.



Conclusions

The differences in model output induced by changes in HPC platform for our test cases appear to be quite consistent in magnitude, position, and shape for each individual model. Although the maximum point-wise differences for WRF NMM are larger than those for FIM, the mean point-wise difference for both models is roughly the same. The distribution of magnitudes of differences for the two models is a little different, but the relative frequency of large point-wise differences for both models is extremely small and most differences for both models are near the limits of single-precision representation.

Future Work

The most pressing question we would like to investigate is how changes in HPC platform impact the differences in output when a model is cycled. We plan to use HWRF to investigate that question. We also plan to look at other fields to see if, and how, the magnitude of differences changes, and to see if differences of fields are correlated.

Acknowledgments

We thank the HPCC Incubator program for funding this work; Craig Tierney, Tressa Fowler, Tara Jensen, Jamie Wolff, and Tom Henderson for their helpful advice; Julien Lynge, George Carr, and Judy Henderson for performing model runs; Kevin Brundage for advice and invaluable help with Python.