Advancing U.S. Operational Weather Prediction Capabilities (in the next decade) with Exascale HPC

Mark Govett

NOAA Earth System Research Laboratory

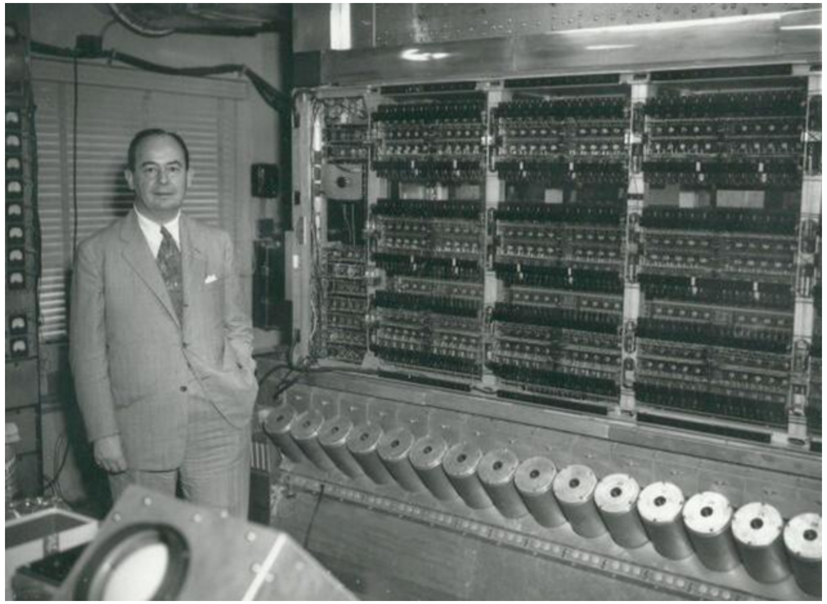
Global Systems Division

Boulder Colorado

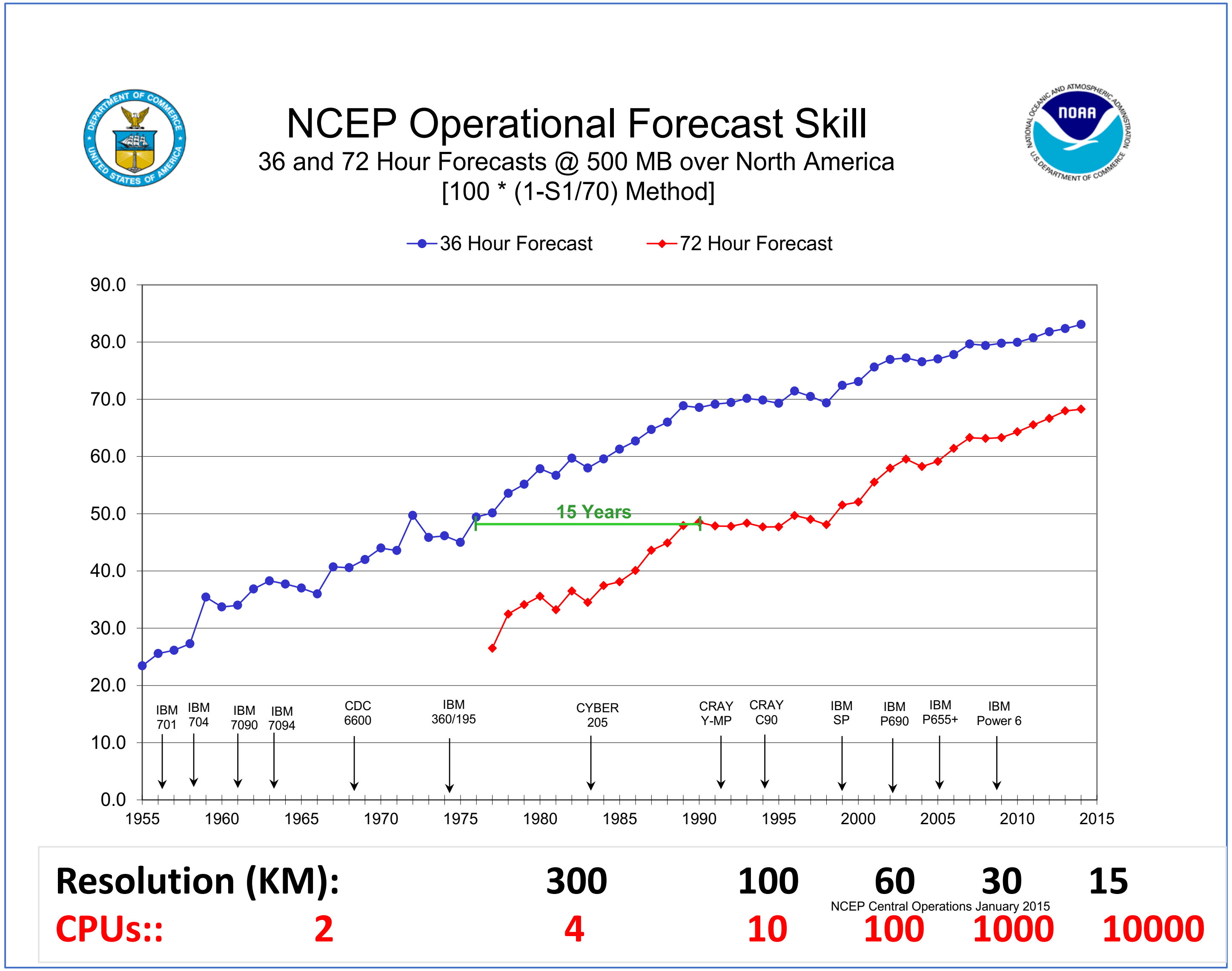
NVIDIA GTC 2019

March 19, 2019

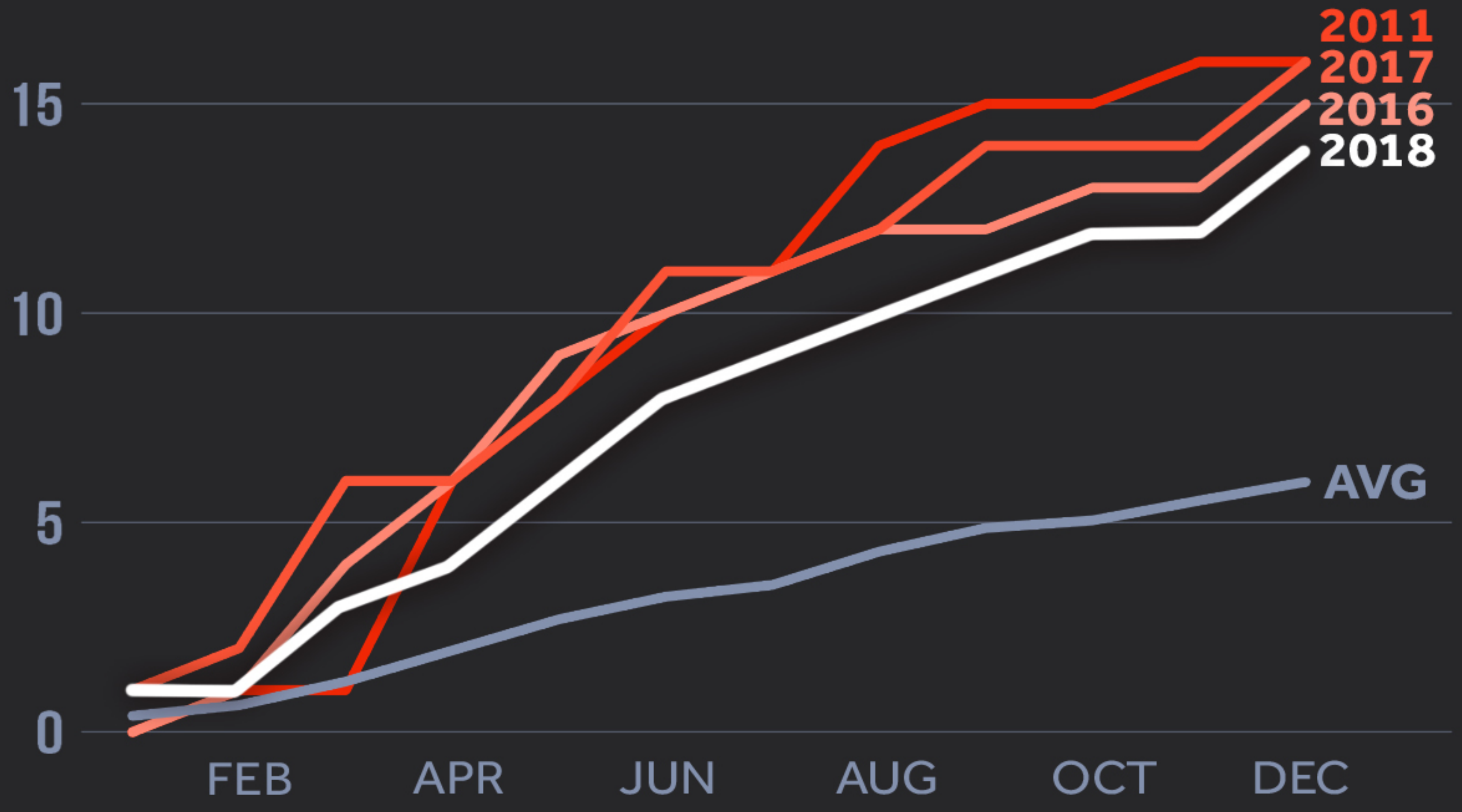
HPC & NWP



John von Neumann posing with the ENIAC computer, 1946 (photo courtesy of NOAA)



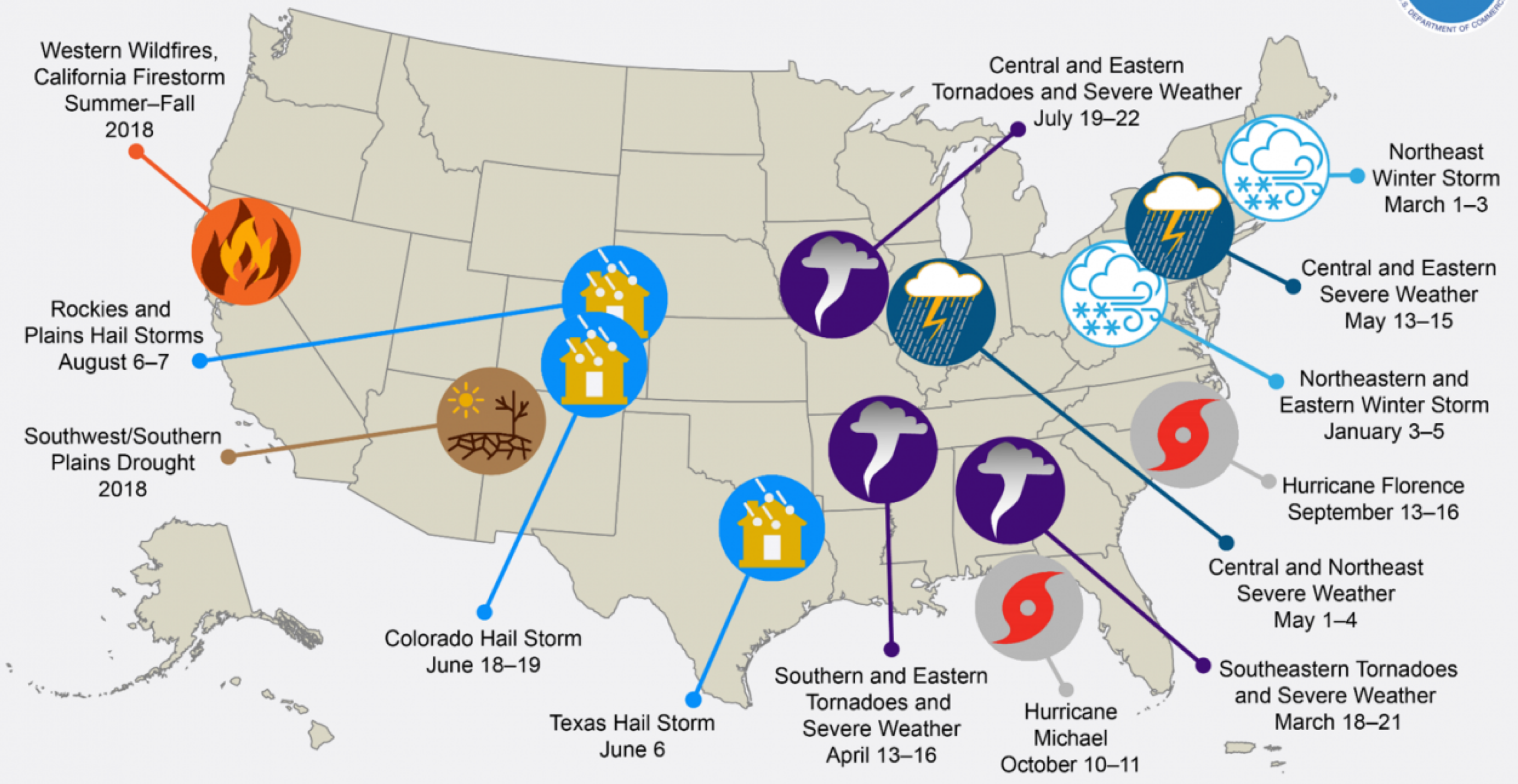
2018 BILLION-DOLLAR DISASTERS
WEATHER AND CLIMATE EVENTS



Cumulative CPI adjusted billion-dollar disaster frequency. 1980-2018 average.
Data as of 2/6/2019. Source: NOAA/NCEI

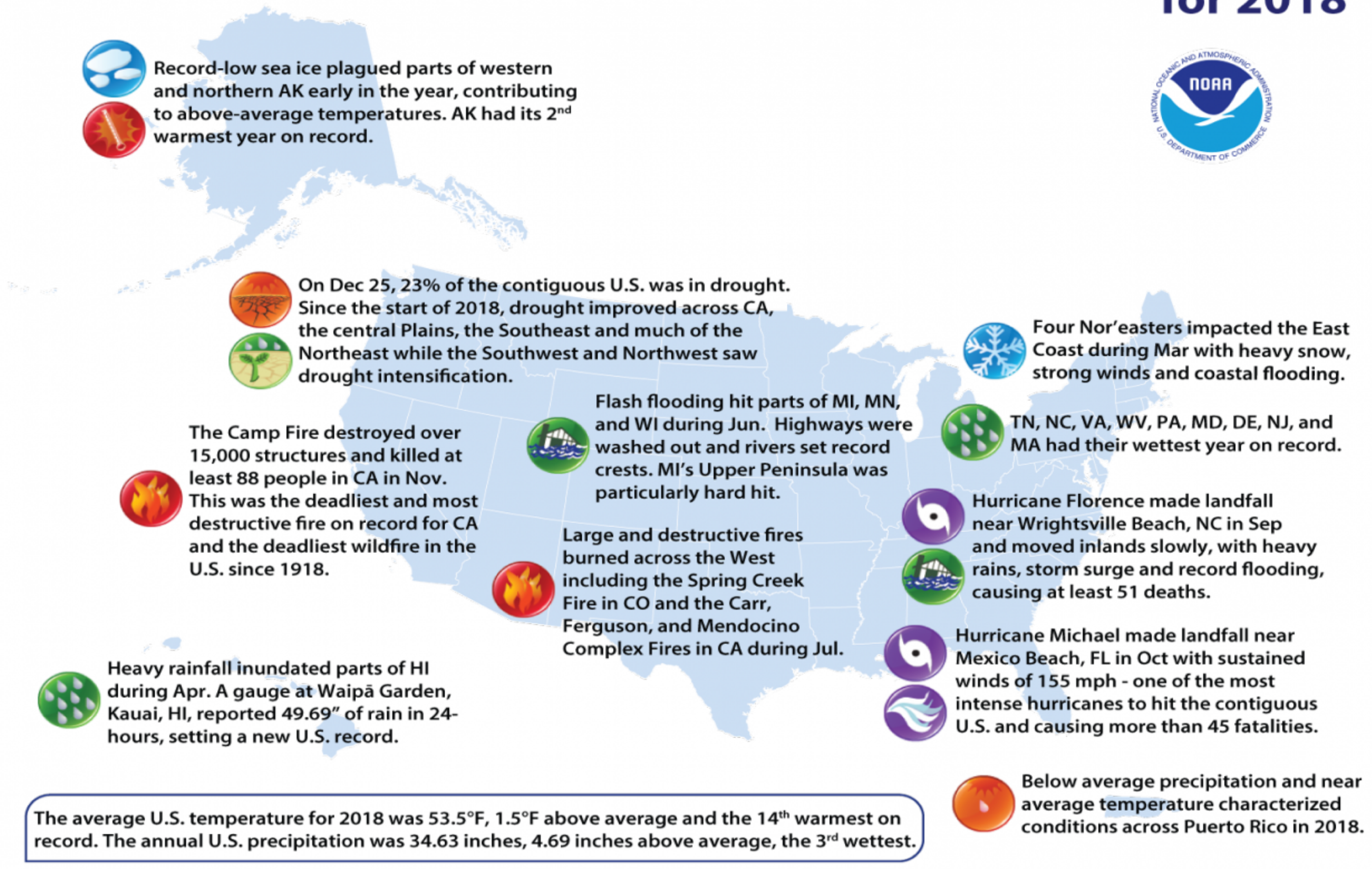
CLIMATE CENTRAL

U.S. 2018 Billion-Dollar Weather and Climate Disasters



*This map denotes the approximate location for each of the **14** separate **billion-dollar weather and climate disasters** that impacted the United States **during 2018**.*

U.S. Selected Significant Climate Anomalies and Events for 2018

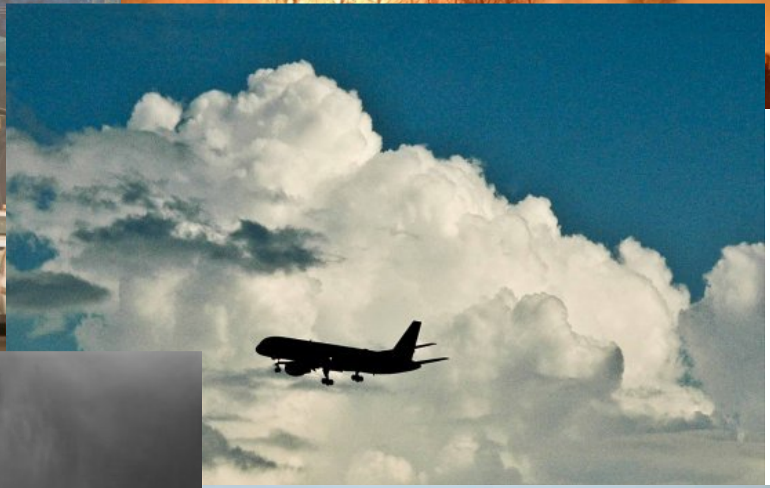


Please Note: Material provided in this map was compiled from NOAA's State of the Climate Reports. For more information please visit: http://www.ncdc.noaa.gov/sotc

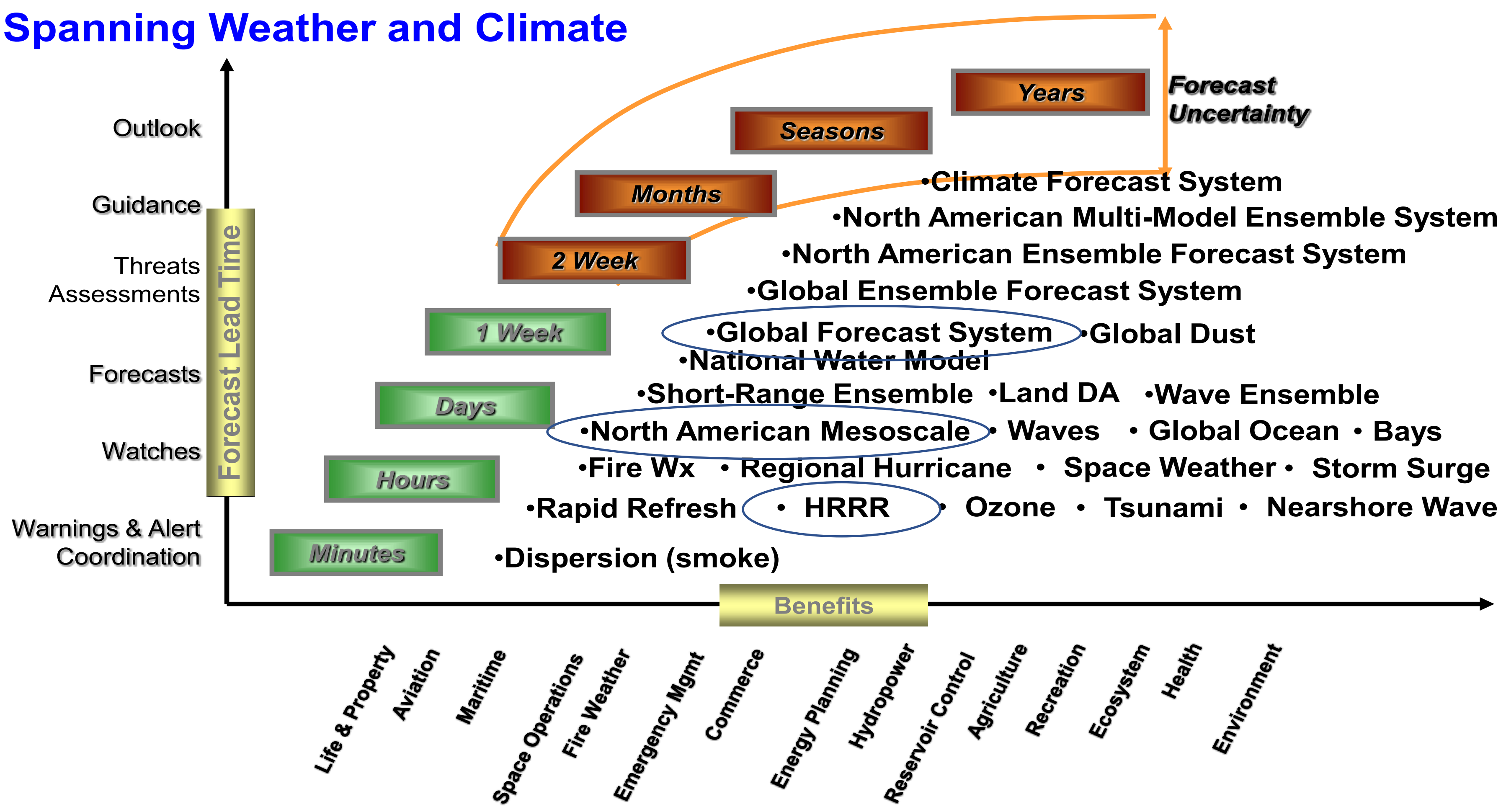
Mitigating Impacts

- Detection
- Prediction
- Dissemination
  - Forecast Offices
  - Fire weather centers
  - Aviation
  - Air quality
  - Transportation
  - Water centers





Seamless Suite of Operational Numerical Guidance Systems

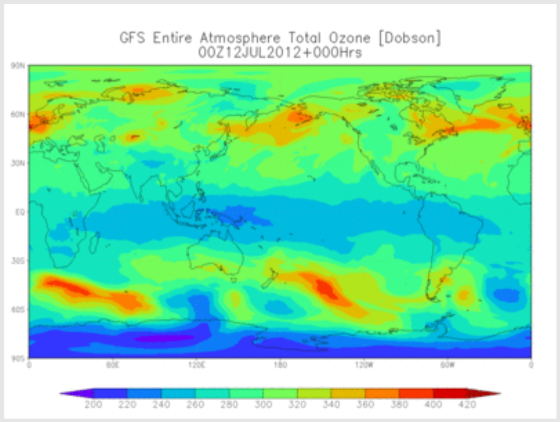


Slide from M.Farrar, EMC Modeling Strategy, 2017

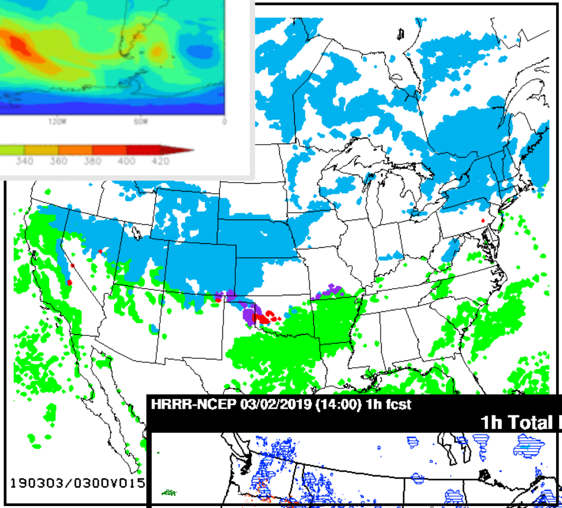
NWS Weather Forecast Models (2019)

constrained by HPC

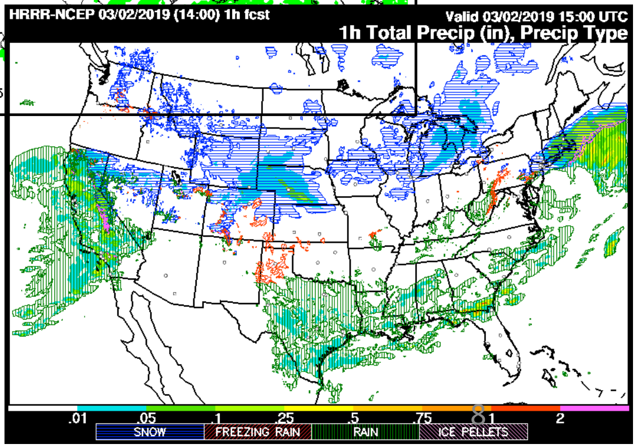
GFS



NAM



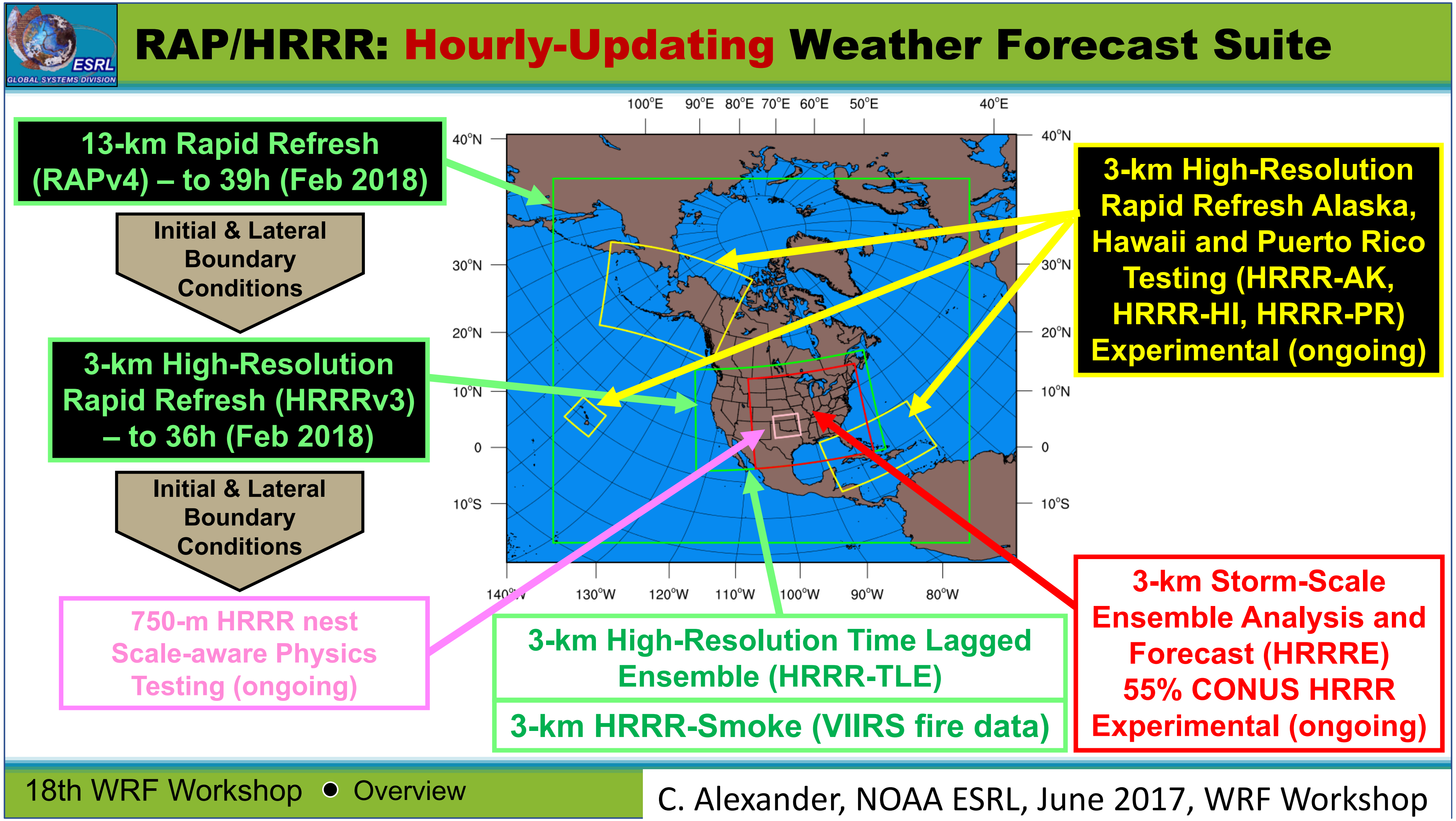
HRRR



Higher resolution *means*
smaller area *and*
shorter forecasts

- Global: Global Forecast System (GFS) (28 KM)
  - Weeks: 0 - 16 day forecasts, 4x / day
- Regional: North American Model (NAM) (12KM)
  - Days: 84 hours, 4x/day
- Regional: High Resolution Rapid Refresh (3KM)
  - Hours: 36 hours, 24x/day

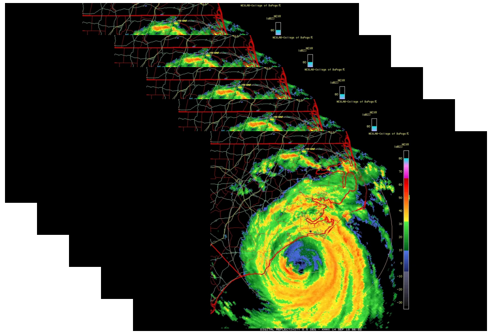
Nesting: GFS (global) + RapidRefresh + HRRR



Improved Weather Prediction

is a tradeoff between
- Computing
- Accuracy
- Time-to-solution

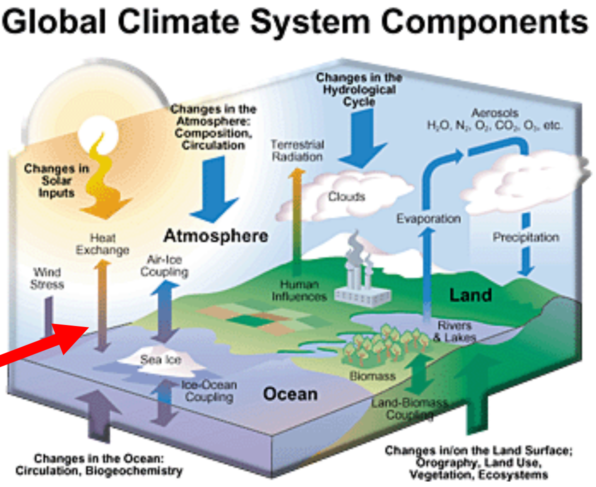
10-100s of members



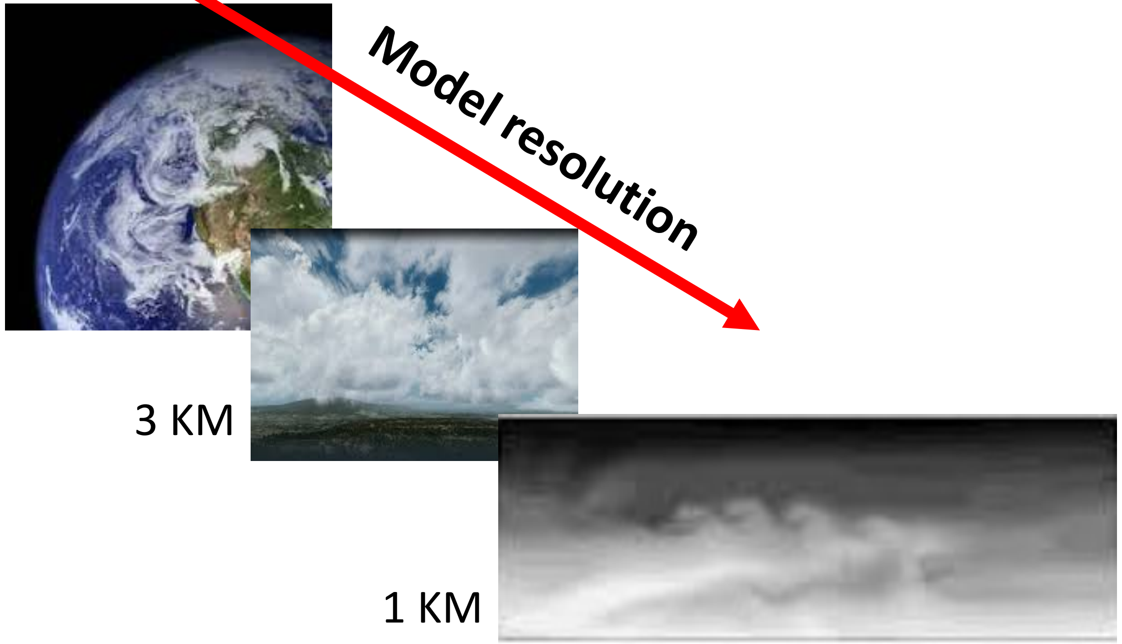
Ensembles

Model complexity

Global Weather System Components



Model resolution



13 KM

3 KM

1 KM

Computational Challenges

- Processors are not getting faster
- ESPC HPC Working Group: 2016 -
  - NOAA, NASA, DoE, DoD Navy, NCAR
  - Discuss HPC challenges, limitations for weather & climate applications
  - Position paper describing concerns

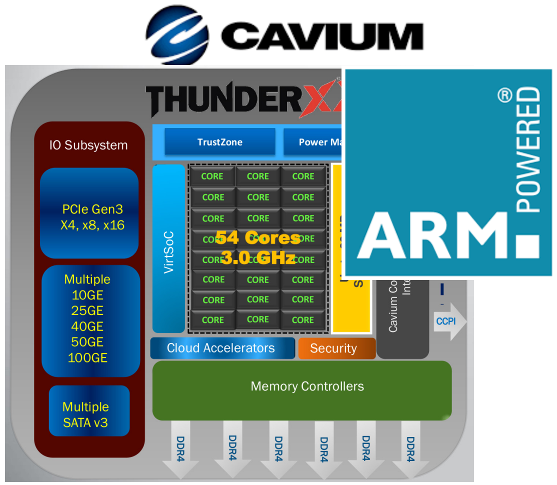
"**HPC architectures are developing in the wrong direction** for state-heavy, low computational intensity (CI) Earth system applications."

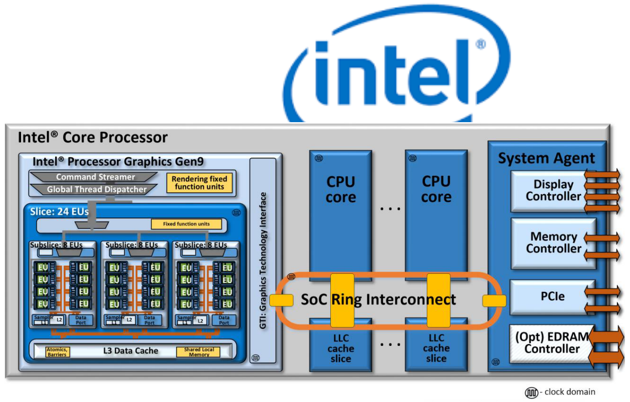
"**NWP applications average less than 2% of peak performance**, constrained by their ability to perform sufficient calculations for each expensive access to memory."

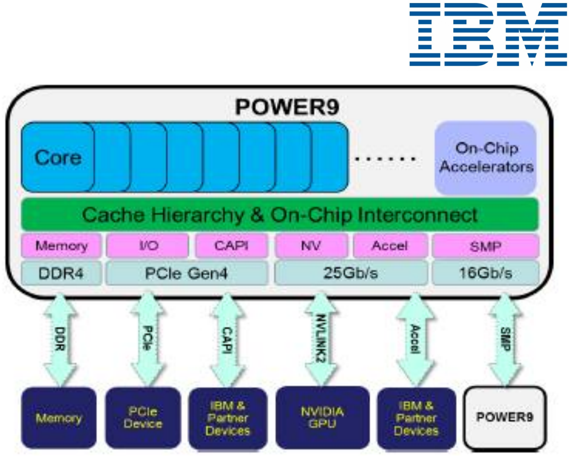
Carman, et al. "Position Paper on High Performance Computing Needs in Earth System Prediction." National Earth System Prediction Capability (ESPC) program. April 2017. https://doi.org/10.7289/V5862DH3

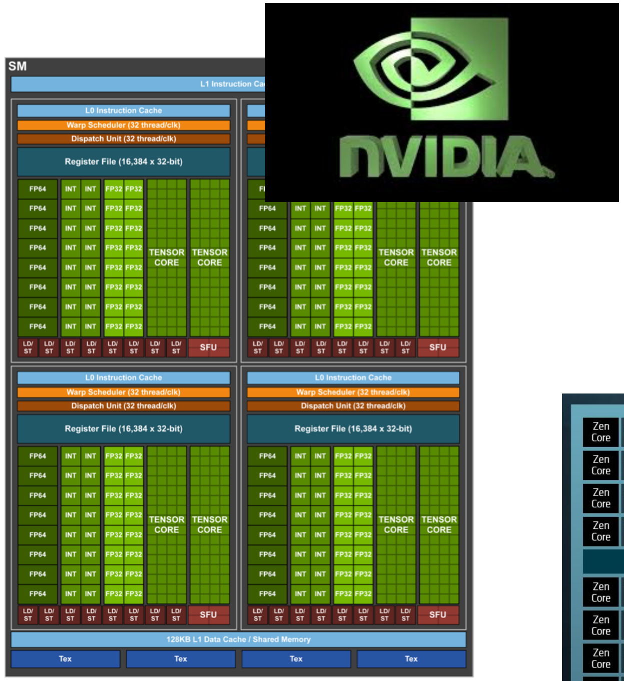
Processor Technologies

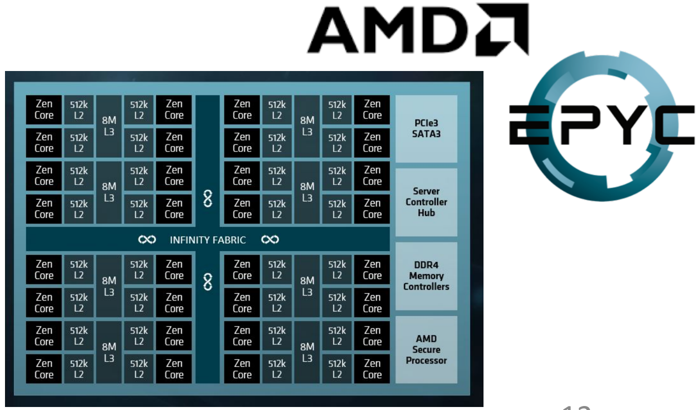
- CPU, GPU TPU, FPGA, ARM
- Diversity
  - Processor
    - Clock speed, energy
    - 10’s to 1000’s of cores
    - Single, double, half precision
  - Memory
    - Size, speed, type
- Burden on compilers, standards
  - Portability
  - Performance
  - Interoperability





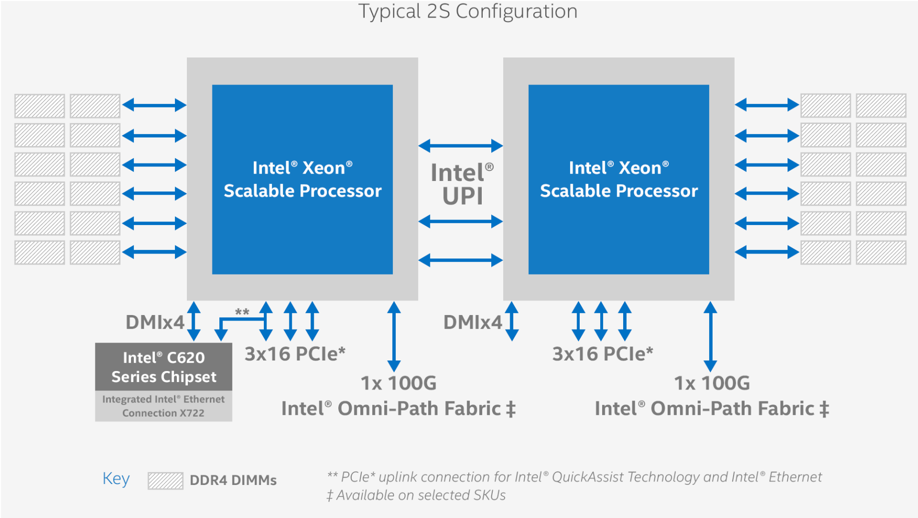




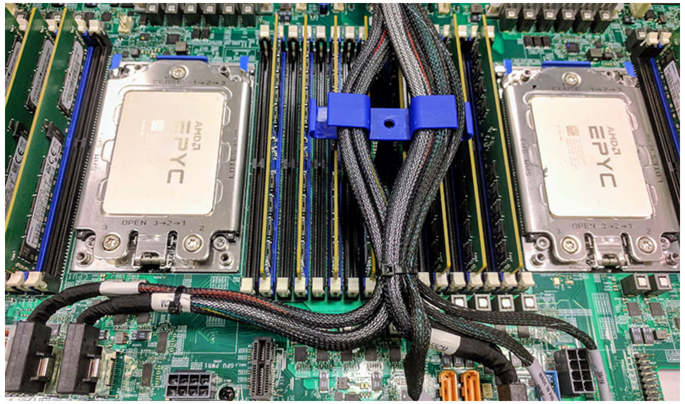


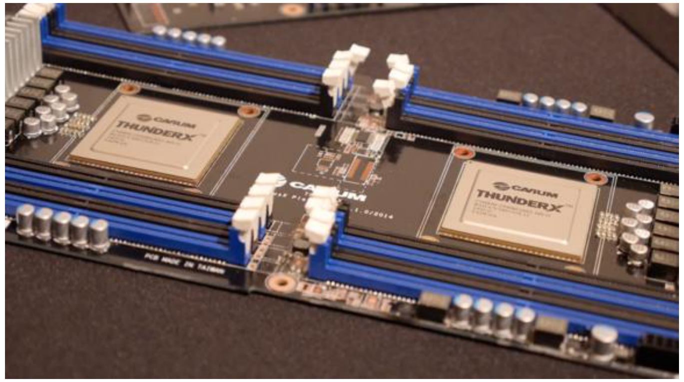
Node Technologies

- Increasing diversity
  - Number of sockets, processors
    - Tens to thousands of cores
  - Memory
    - Speed, bandwidth
  - Communications
    - Intra-socket
    - Intra-node (PCIe, NVLINK)
    - File system
- Many vendors, choices
  - Performance, energy, cost



Intel Skylake dual-socket

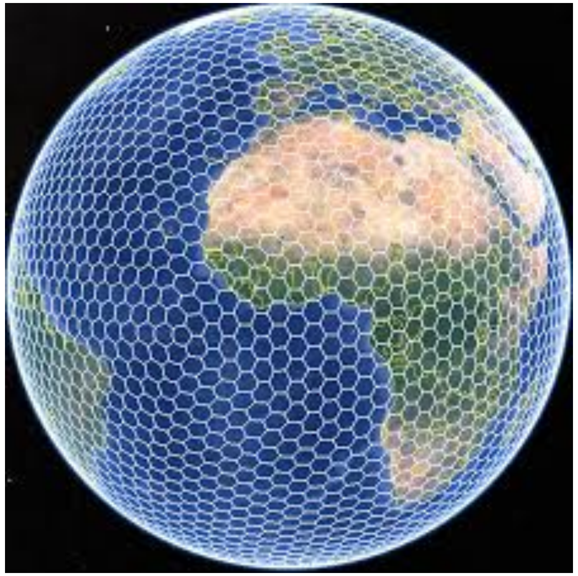

Super-micro dual socket EPYC


Cray dual-socket ARM

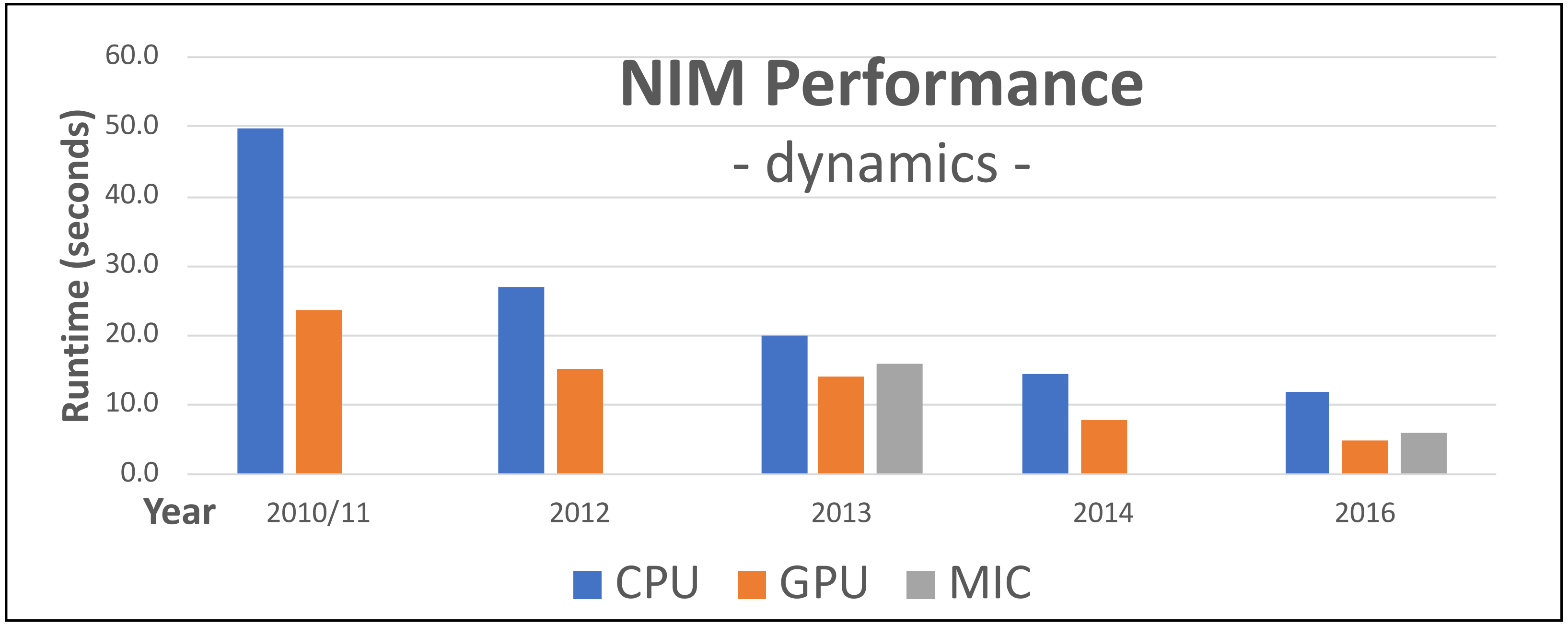
Application Performance – Single Node

Research model developed by NOAA ESRL/GSD (2010—2016)

- **Directive-based (OpenACC, OpenMP, SMS), performance portable**
- GPU is 2-3 times faster than CPU (Fermi to Pascal generation GPUs)



Uniform Icosahedral Grid



M.Govett, et. al., Parallelization and Performance of the NIM Weather Model on CPU, GPU and MIC Processors, BAMS, October 2017

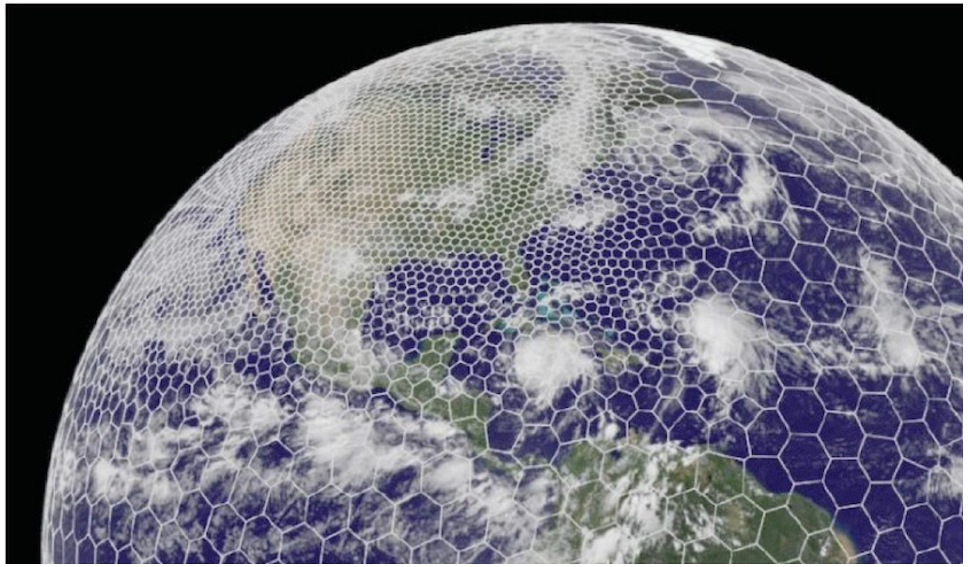
Application Performance – Single Node

MPAS
Model for Prediction Across Scales

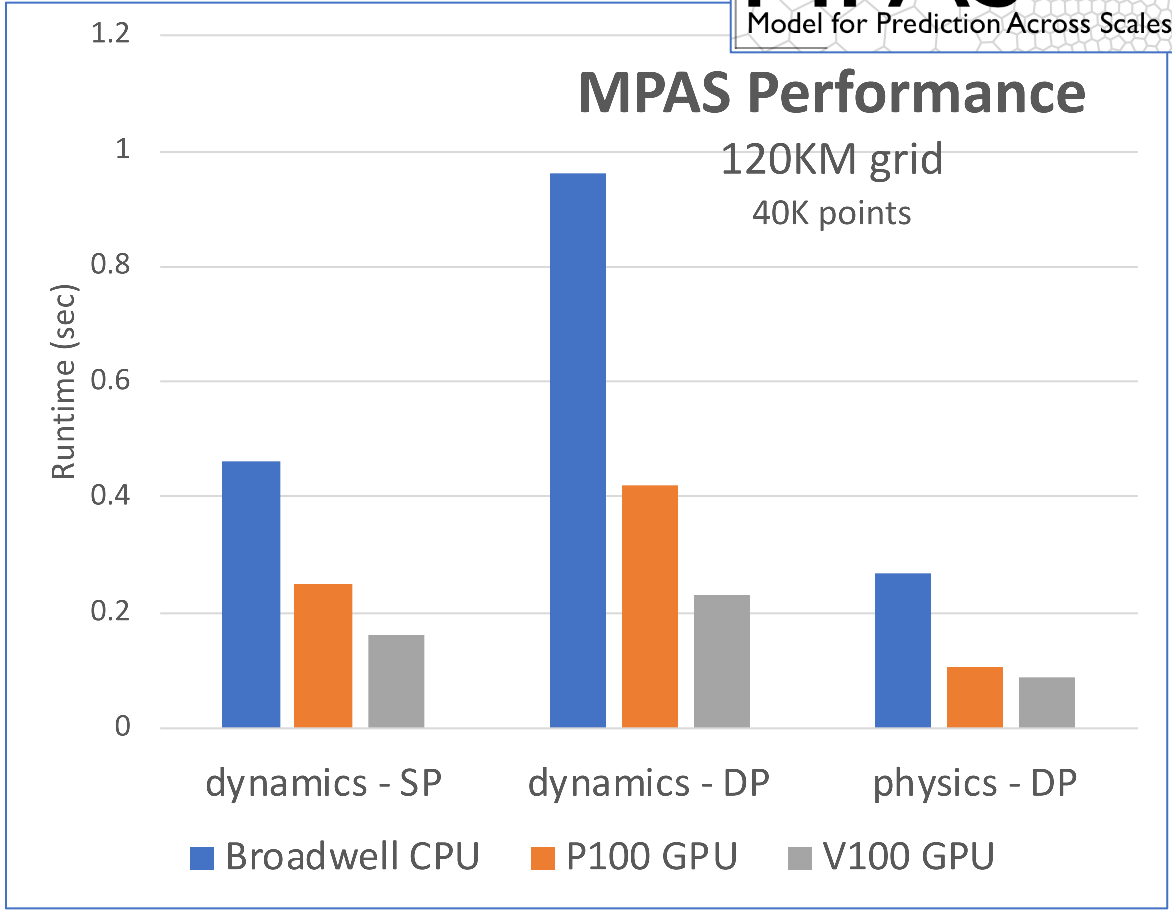
MPAS model developed at NCAR

adopted by IBM Weather Company

- GPU is 3X faster than CPU (Volta versus Broadwell)
- Directive-based, performance portable



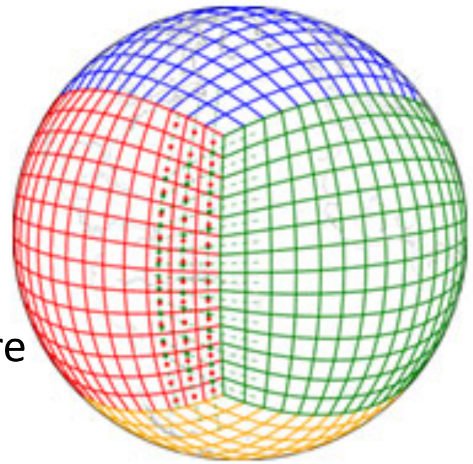
Non-uniform Icosahedral Grid



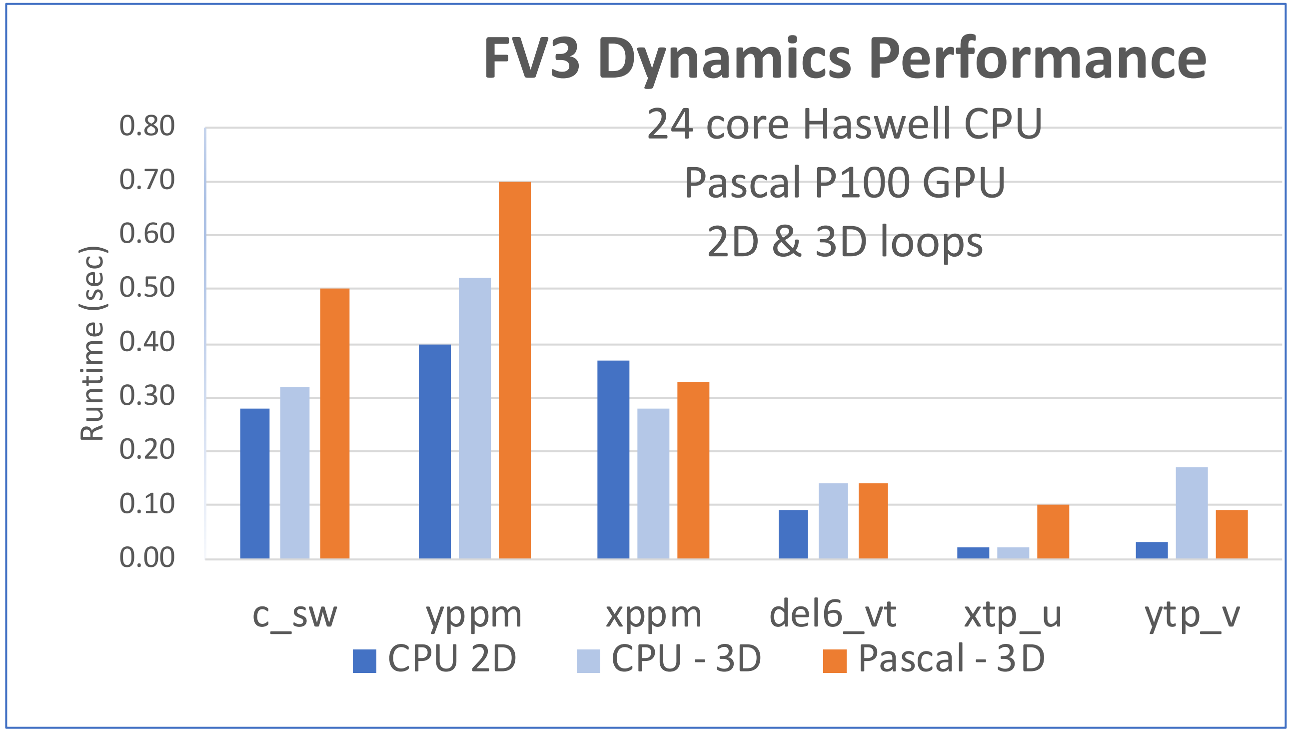
R.Loft, Sept 2018, ECMWF HPC Workshop

FV3GFS Performance – Single Node

- Finite-Volume Cube-Sphere Model selected by NOAA NWS
  - Designed for CPU
  - Efficient use of cache memory
- Slower on GPU
  - Code changes slowed down CPU
  - Not performance portable
- Inefficiencies
  - Limited parallelism
  - Non-uniform cube-sphere grid
  - Pervasive edge & corner calculations
- Ongoing efforts to address GPU performance challenges



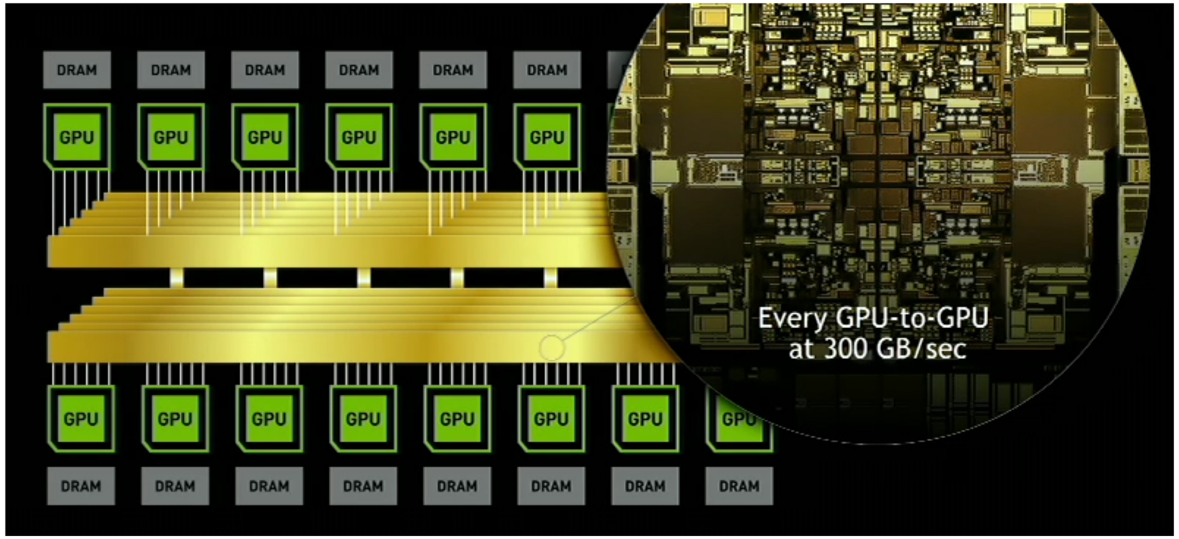
Cube-sphere grid



M. Govett, June 2018, PASC Symposium

Advanced Node Technologies

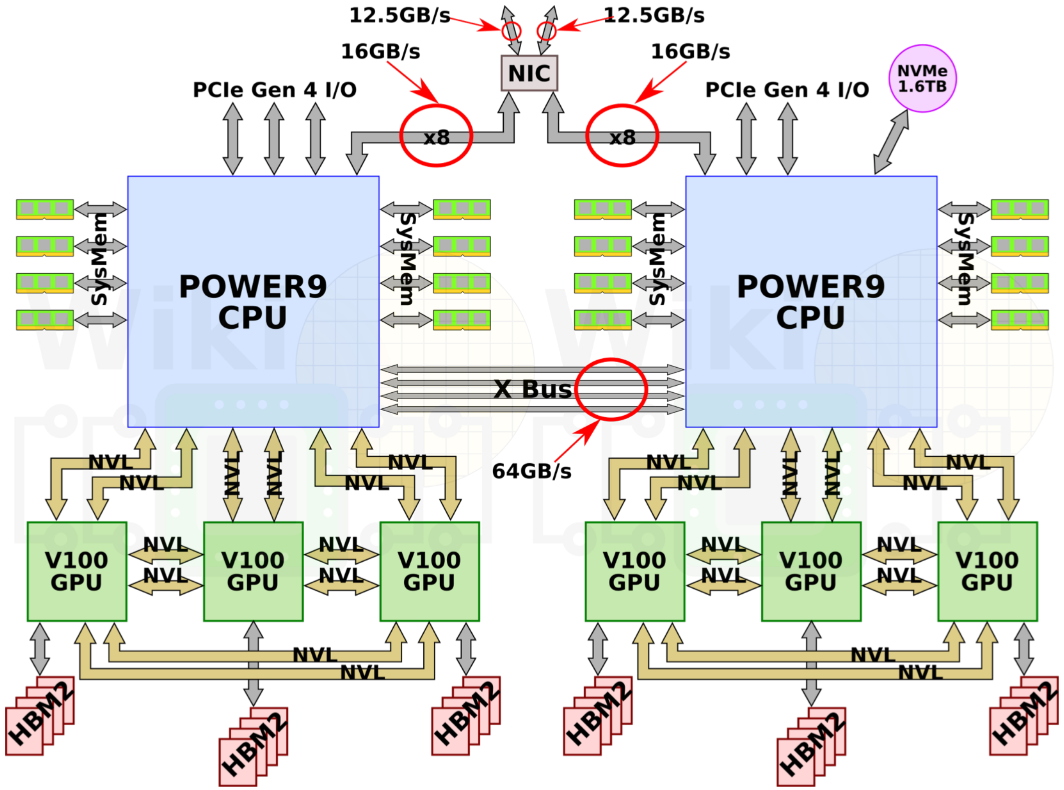
- Diversity
  - Performance, cost, power
- Complexity



NVIDIA DGX-2: 16 Tesla V100 GPUs, (81K GPU, 10K Tensor cores).

- 1.5 TB DDR4 RAM, 500 GB HBM2, 10kW power
- 300 GB/s NVLINK
- PCIe Gen3, 8x EDR IB / 100 Gigabit Ethernet

ORNL Summit Node



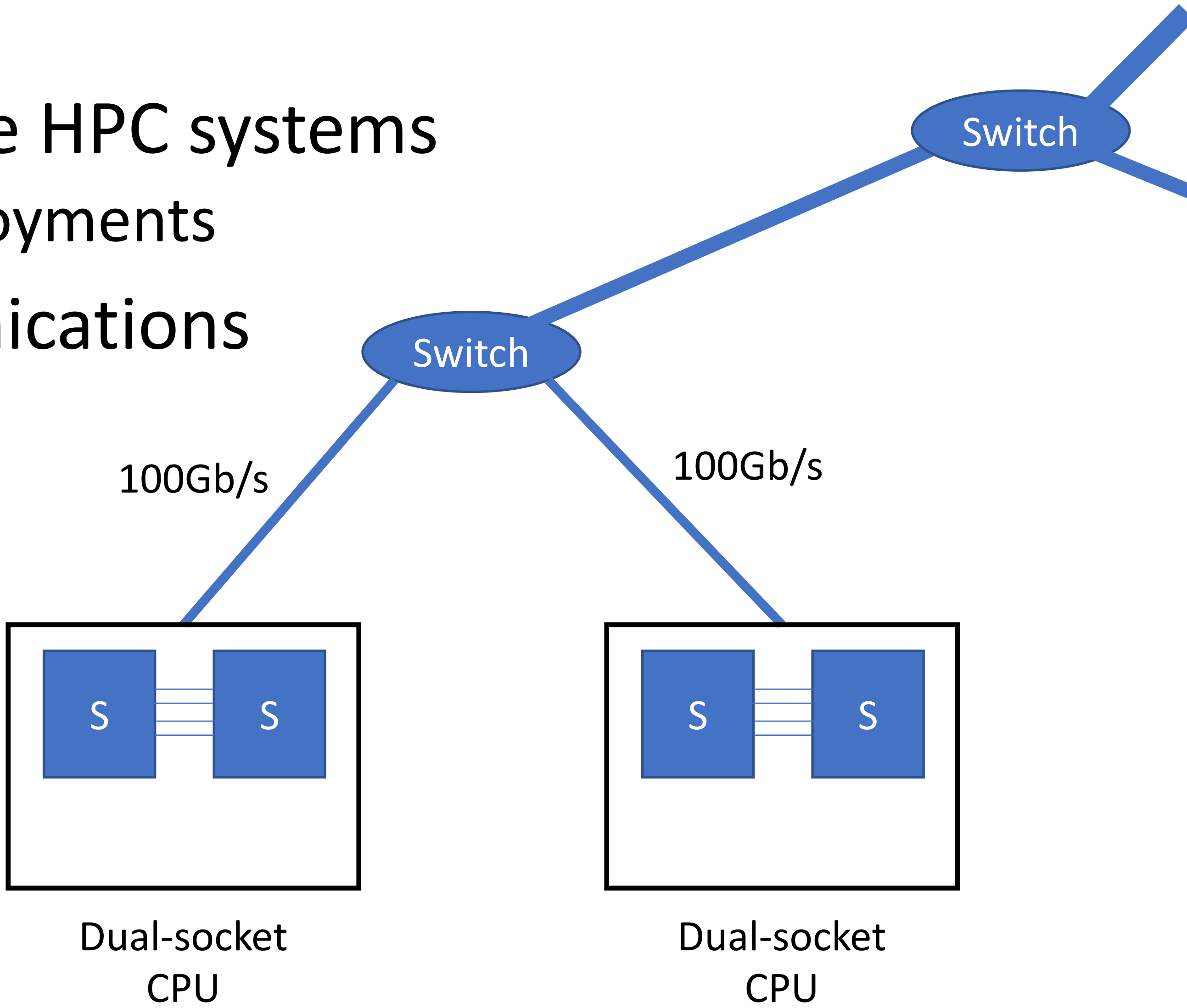
DOE Summit node:

- IBM Power9 CPU, 6 V100 GPUs, 30K GPU cores
- 512 GB DDR4 RAM, 96 GB HBM2
- NVLINK, 50GB/s bandwidth per link
- PCIe Gen 4 (16GB/s) for inter-node, I/O

Summit System: 4600 nodes, 27K GPUs

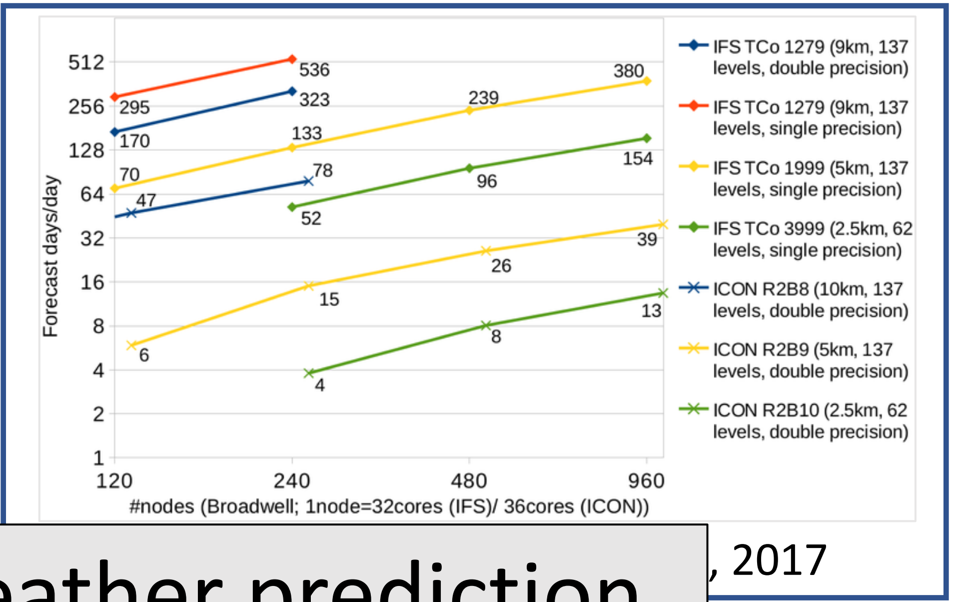
System Inter-connect Technologies

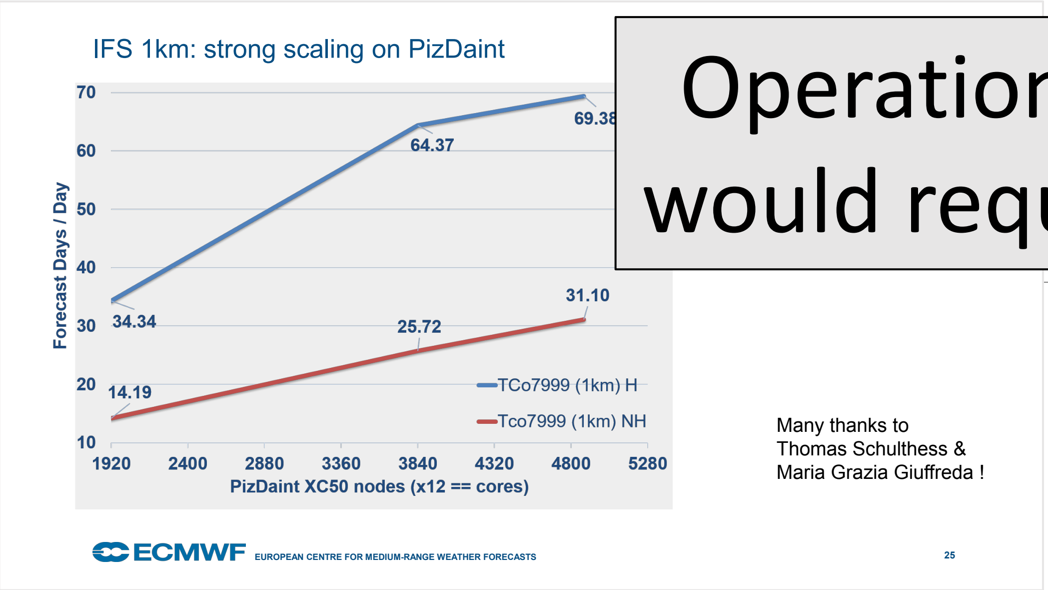
- Interconnect required for large HPC systems
  - Weakness in large system deployments
- Applications use MPI communications
  - Pack message buffer
  - Inter-process communications
  - Unpack message buffer
- Scalability a big challenge for application performance



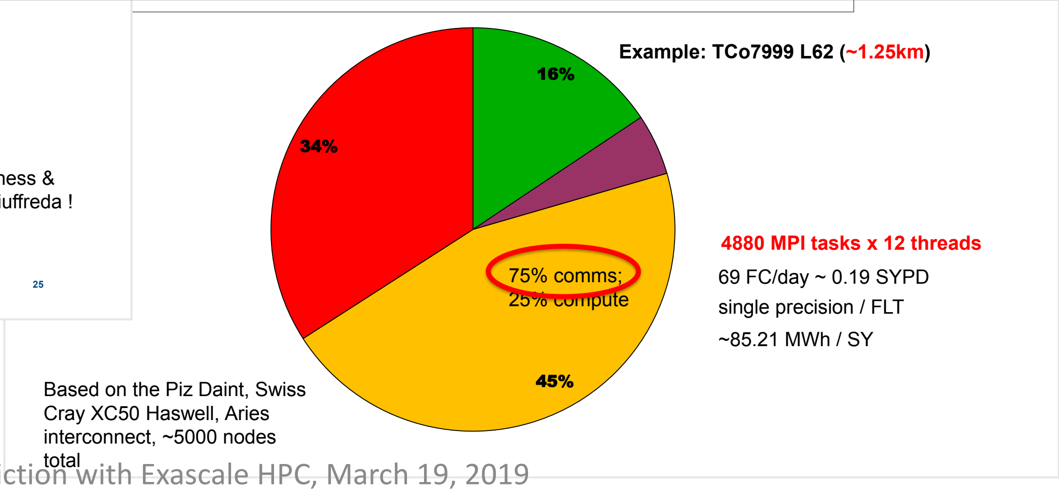
Application Scalability

- ECMWF Scalability Programme (2014 -)
  - ESCAPE, NextGenIO, ESiWACE, ESCAPE-2
  - Scaling, I/O, compilers, algorithms
- Targeting 1-3 KM resolution for global models





Operational weather prediction would require 200-240 days / day



Nils Wedi, ESCAPE Project Presentation
ECMWF HPC Workshop, Sep 2018

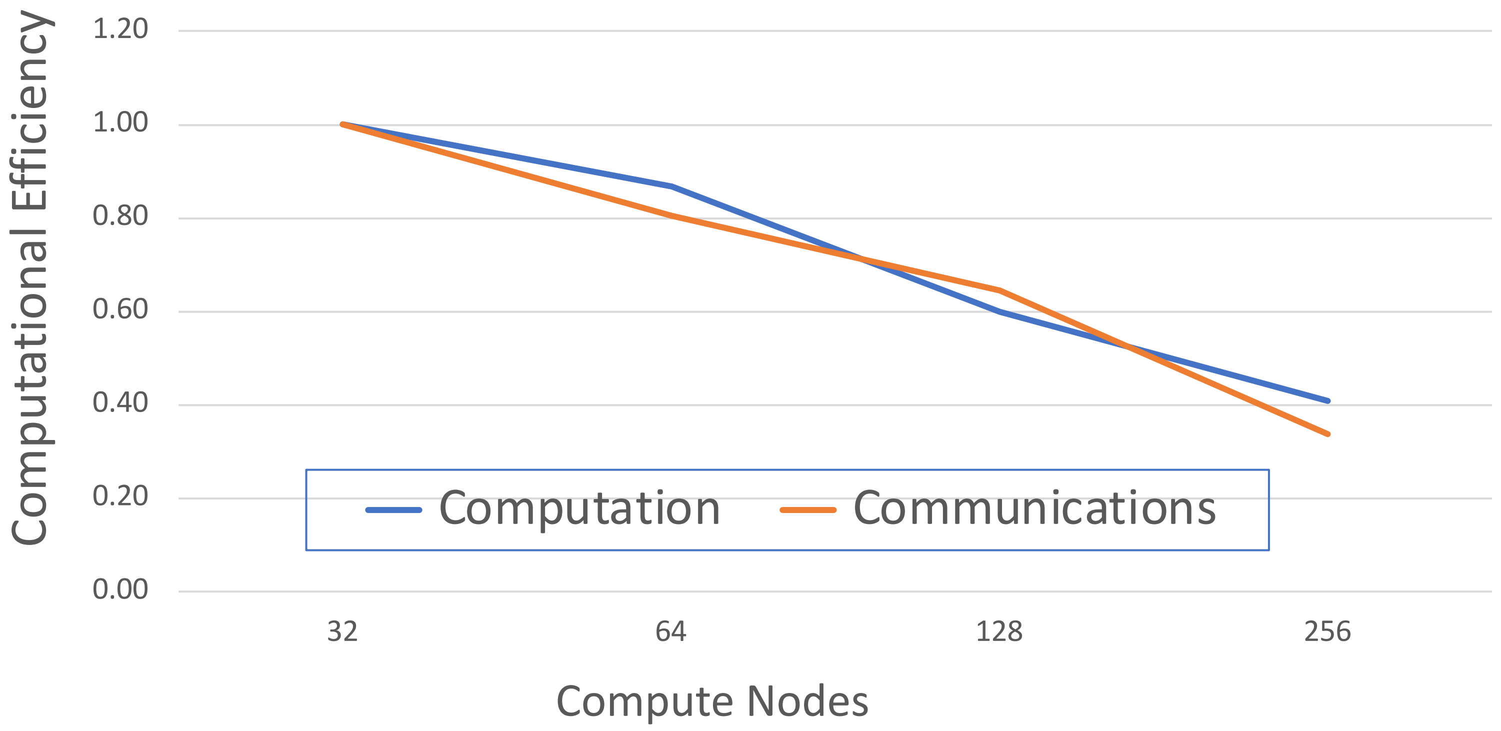
Scaling Factors

- Computation
  - Parallelism
  - Algorithms
  - Model grid
- Communications
  - Frequency
  - Data volume
  - Overlapping

FV3GFS Strong Scaling Efficiency

Physics + dynamics

14 KM resolution



Time to Solution by the Numbers

FV3GFS Performance

3 KM resolution, 5 day forecast

Weak Scaling

Operational requirement: 5 days in 2250 seconds (10 days in ~1.25 hours)

	Actual Performance		Estimated Performance	
Resolution	**28 KM**	**13 KM**	**6.50 KM**	**3.25KM**
Time Step	225 sec	112.5 sec	56 sec	28 sec
CPU Nodes	64	256	1024	4096
CPU cores	**1536**	**6144**	**24576**	**98304**
Total Time	1094	1916	3357	5880
Dynamics	560	792	1120	1584
Communications	440	710	1146	1851

Runtimes in seconds for a 5 day forecast, *NOAA theia system with 24 cores Haswell nodes*

Time to Solution by the Numbers

FV3GFS Performance

3 KM resolution, 5 day forecast

Strong Scaling

Operational Requirement: 5 day in 2250 seconds (10 days in 1.25 hours)

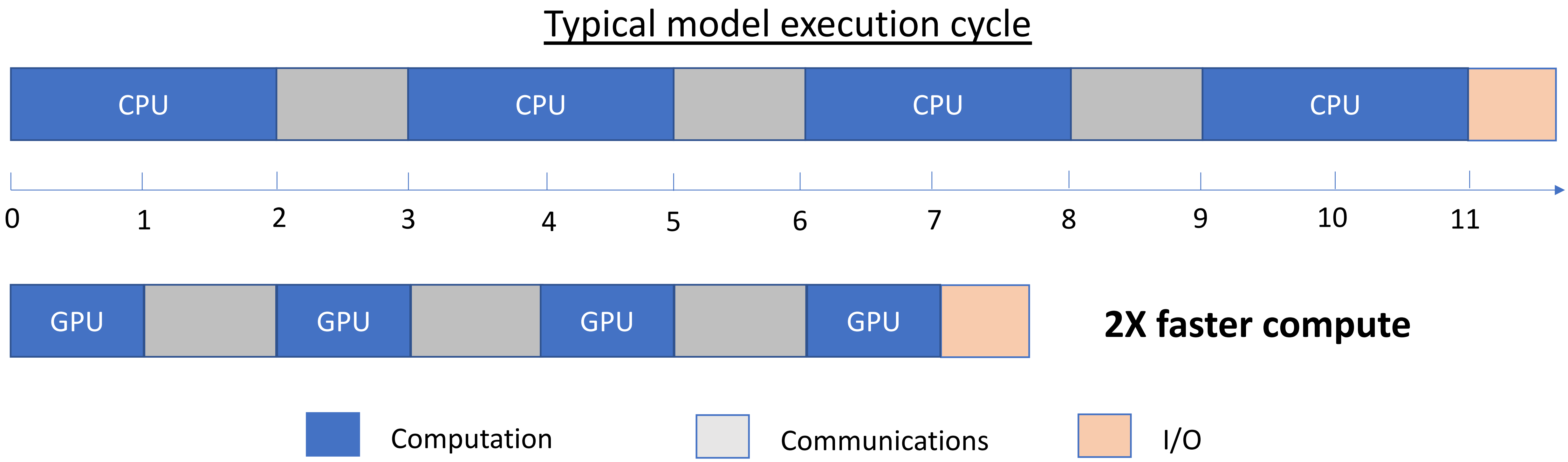
Tile Size / MPI	48 x 48	24 x 48	24 x 24
CPU Cores	98,304	196,608	393,216
Total Time	**5880**	**3962**	**2095**
Dynamics	1584	1275	643
Communications	1851	1390	801

Estimated performance, *NOAA theia system: 27,000 cores, 24 Haswell cores / node*

- 393,216 cores = 16,384 CPU nodes
- 30% of runtime is for inter-process communications

Performance and Scalability

CPU and GPU



2X faster compute does not mean 2X faster

This example is only 1.6X faster

Data Challenges

Data is only useful if it can be used

Observations

Assimilation

Prediction

Output

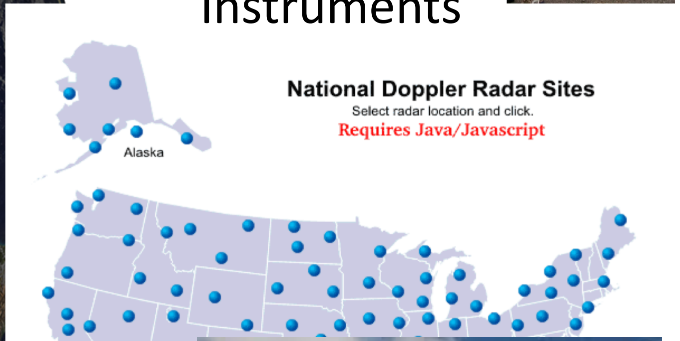
Distribution

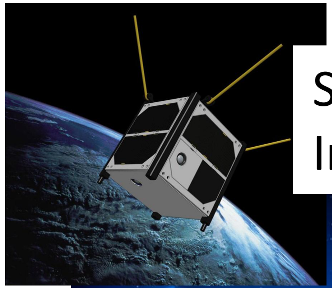
Dissemination

Observations

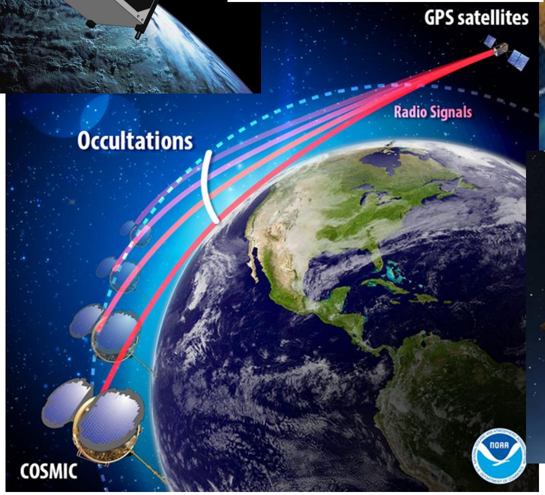
- More data than we can use
  - GOES, JPSS, cubesats, nano
  - Radar, balloons, ships, planes
  - Autos, cell, sensors, …
- Tremendous potential

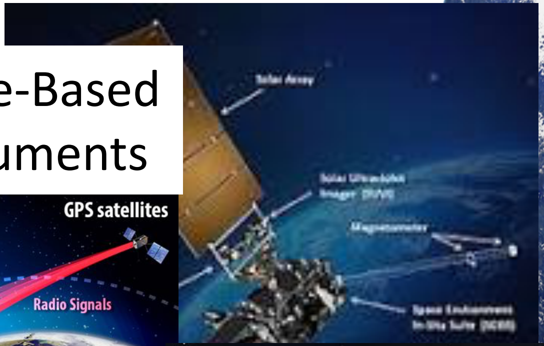
Ground-Based Instruments



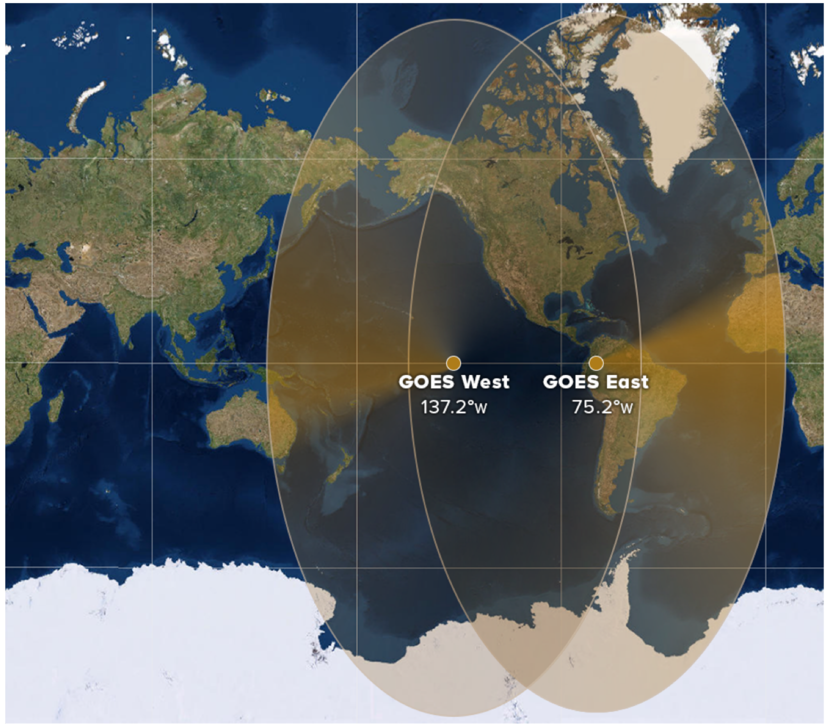


Space-Based Instruments



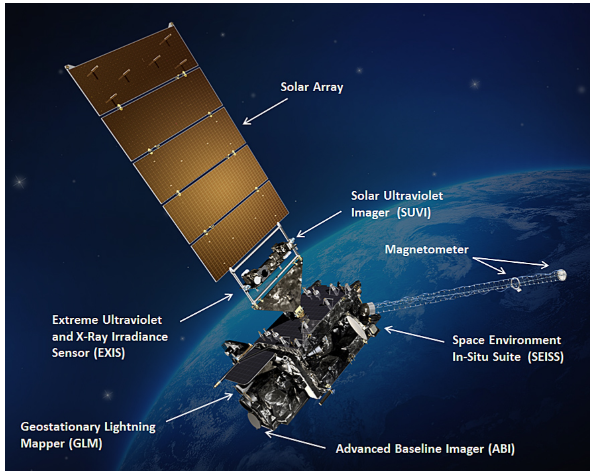


Geostationary Operational Environmental Satellite (GOES)



Only 1% is being used

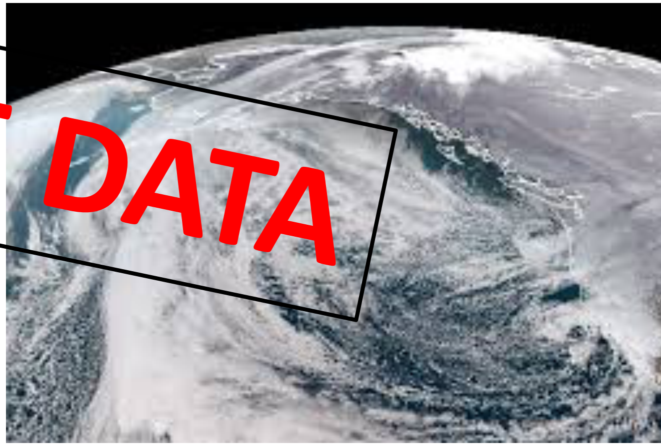
- 2012-2017: GOES-13, GOES-14, GOES-15
  - Scans every 3 hours, 10 bit precision
  - 4 spectral bands @ 4KM
  - 1 visible band @ 1KM



- 2017 - ~2027: GOES-16, GOES-17
  - Scan every 15 minutes, 14 bit precision
  - 14 spectral bands @ 2KM resolution
  - 2 visible bands @ 0.5KM resolution
  - High-res nest every 30-60 seconds

100x MORE DATA

water vapor image



Model Output: 14KM to 3KM resolution

- Each 3D variable: pressure, temperature, moisture, winds, ….

Resolution (KM)	Vertical Levels	Number of Grid Cells (Millions)	Total Cells (Billions)	Increase in Cells	Per field storage (SP)
14 (1x)	64 (1x)	3.5 (1)	0.25	1x	1 GB
3.5 (4x)	128 (2x)	56.6 (16)	5.4	21x	21 GB

- Model output:

14KM - 10 model fields, 6 hourly output, 10 day forecast 400 GB per run

3KM - 10 model fields, 3 hourly output, 10 day forecast 21.8 TB (52X)
3KM - 10 model fields, hourly output, 2 day forecast 12.0 TB (26X)

Distribution

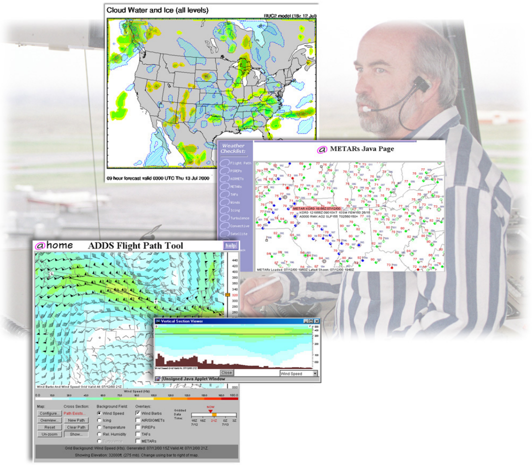
- Diverse user requirements
  - Global, regional, local, observations products
- NWS AWIPS
  - NOAA network is saturated
  - **Everyone gets same data**

AWIPS Workstation

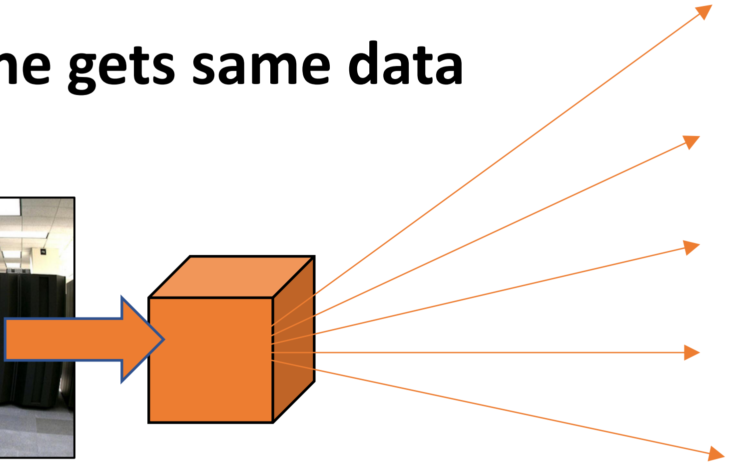
NWS office

NWS Forecast Offices
Hurricane Prediction Center
Storm Prediction Center
National Water Center
Aviation Weather Center
Fire Weather Centers

State, Local, Public
- Floods, fire, winds, hail, …



FAA Air Traffic Control



data center

model output

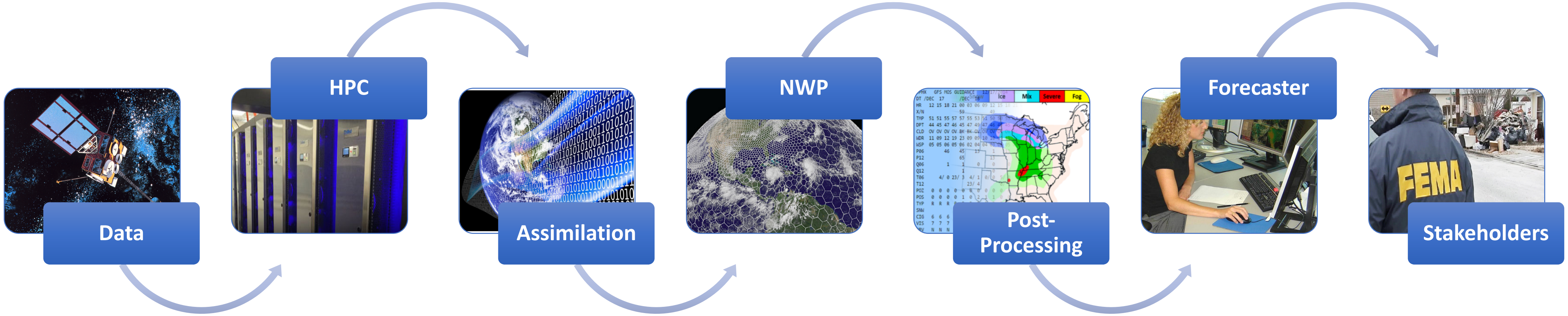
users

State of Play for NWP (2019)

- Scientific advances increasingly constrained by computing, data
- HPC
  - No expected increase in processing speed
  - Limited increases in memory speed
  - Parallelism & scalability limitations
  - Operational time-to-solution constraints
- Data
  - Too much data to process
  - Too many observations to use
  - Too large to distribute

15 KM

3 KM



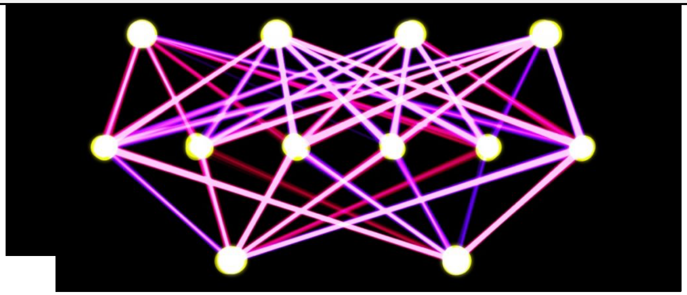
Advancing Weather Prediction
in the next decade

Where do we go from here?

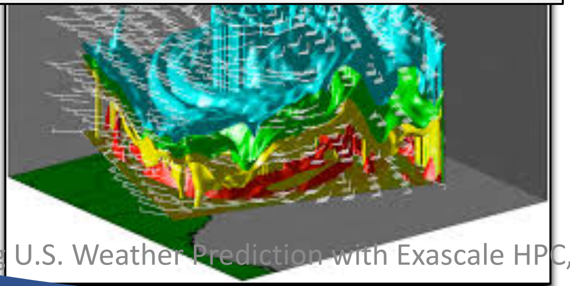
Technology Convergence

SuperComputing

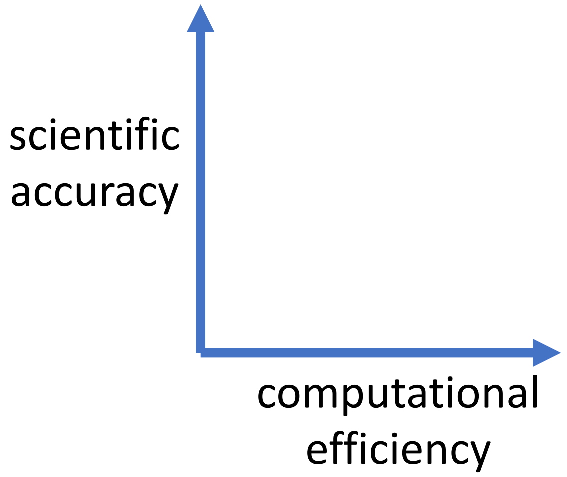
Machine Learning



Big Data



#1 Improve Model Performance



Weather Prediction Models

- dynamics -

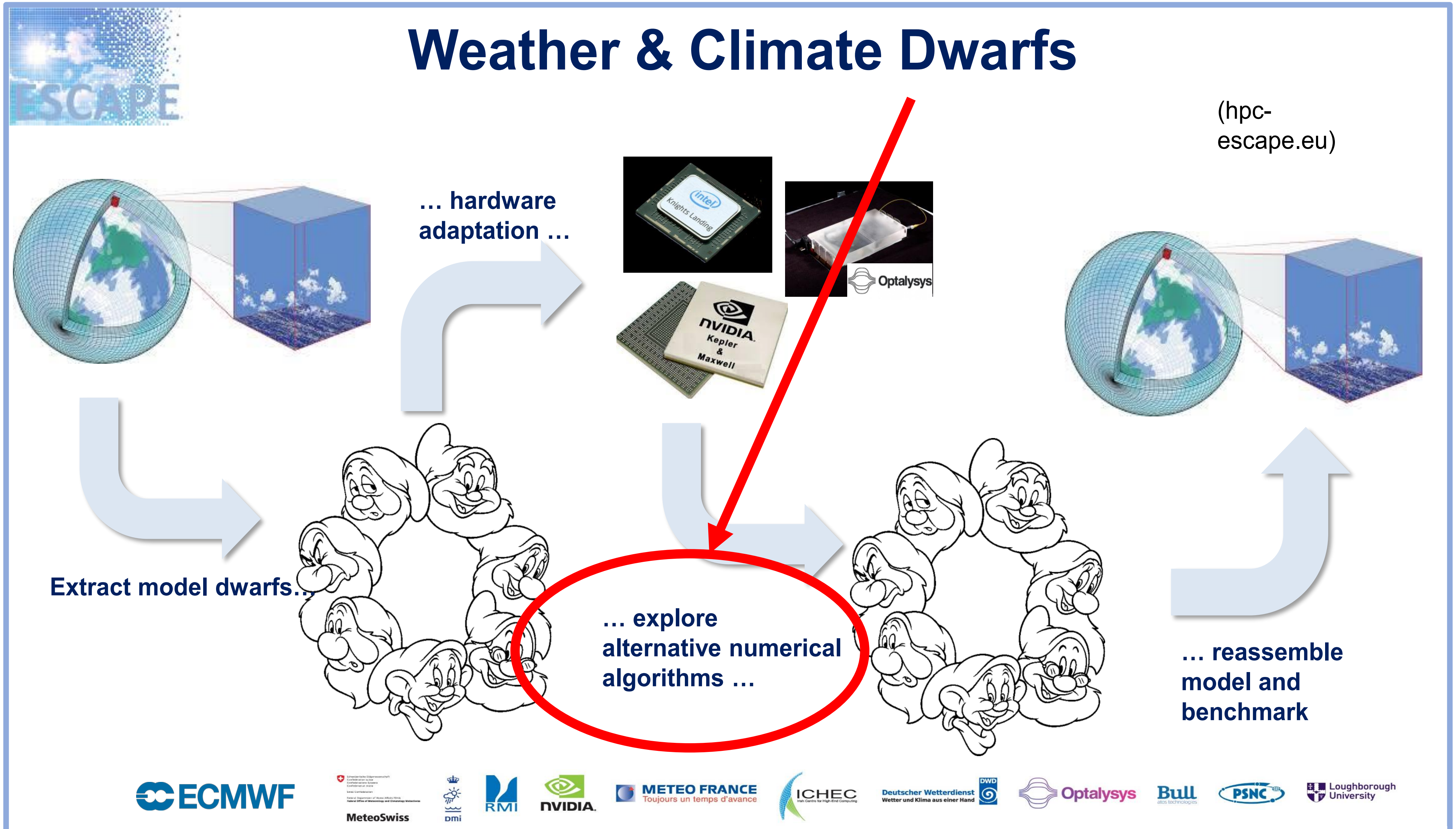
- What are the best models, approaches?
  - algorithms, grids, time-step, physics, etc
  - computational efficiency, scalability

Model Type	Horizontal Grid	Time-Step	Staggering	Models
Finite-volume	Cube-sphere	SISL	A-grid, C-grid, D-grid	FV3GFS
Finite-volume	Icosahedral	HEVI	A-grid	NICAM
Finite-volume	Icosahedral	HEVI	C-grid	MPAS, ICON
Finite-element	Cube-sphere	SISL	C-grid	LFRiC
Spectral-element	Cube-sphere	HEVI	No staggering	NUMA, Neptune, KIM
Spectral	Polar	HEVI	No staggering	IFS, GFS

G.Mengaldo, et.al.,Current and Emerging Time-integration Strategies in Global Numerical Weather and Climate Prediction, https://doi.org/10.1007/s11831-018-9261-8(0123456789().,-vo

Weather & Climate Dwarfs

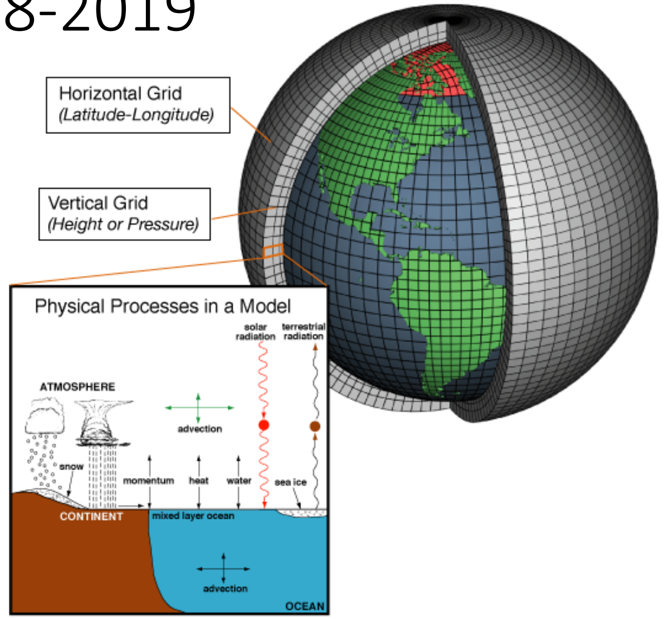
(hpc-escape.eu)



P. Bauer, ECMWF ESCAPE Project Briefing, 2017

Dwarf Development with GeoFLOW

Duane Rosenberg, Bryan Flynt, NOAA ESRL, 2018-2019



- GeoFLOW is an application framework to simplify dwarf development in order to evaluate **computational efficiency vs scientific accuracy** of various approaches
- C++ objects to define communications, grid, discretization & time-stepping operators
- Evaluate for 1-3KM global models on CPU, GPU, ARM, …

Icosahedral Finite Volume (IFV)	Spectral Element (CG, DG)
• Low order/low accuracy	• High order/high accuracy
• 2D, 3D control volumes	• 2D, 3D elements
• Icosahedral grid	• Cube-sphere grid
• Deep communication	• Shallow communication
• staggered (Arakawa) centering	• Un-staggered centering
• Explicit time step	• Explicit & semi-implicit time step

Focus Areas

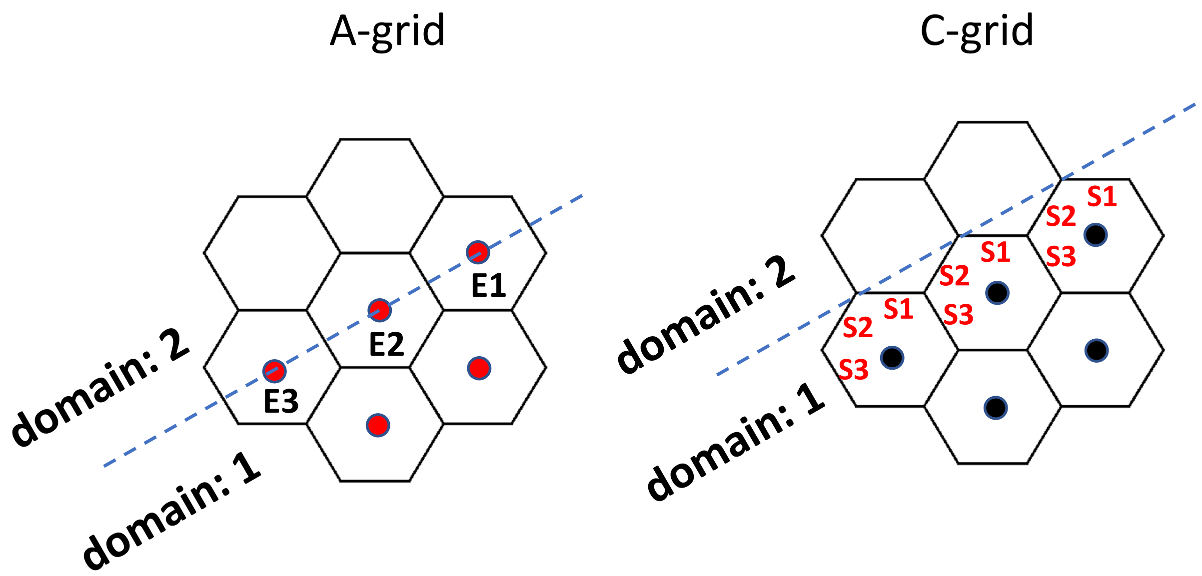
Advection
+ Convection
+ Radiation
+ …

Shallow Water Dwarf: A-grid versus C-grid staggering

Yonggang Yu, Ning Wang, Jacques Middlecoff, NOAA ESRL, 2018-2019

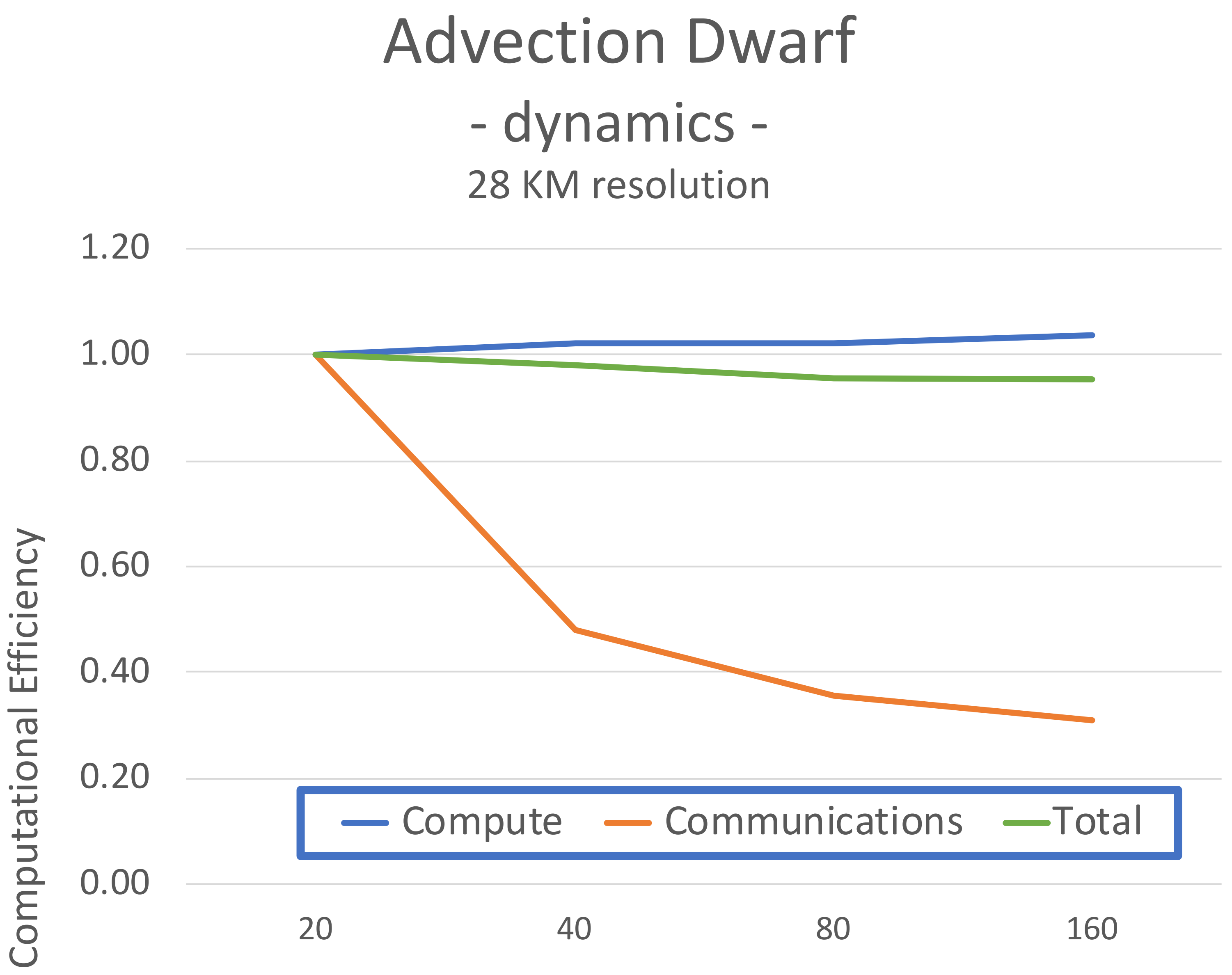
Evaluate performance, scaling and scientific accuracy

- Develop shallow water model for A-grid and C-grid with identical design, grid construction, optimizations, …
- Replicate published dynamical core idealized test results for A-grid (NICAM), C-grid (MPAS)
- OpenMP, OpenACC, MPI parallelization
- Performance & scaling comparison for 3 KM resolution or finer scales
  - NOAA system with 800 Pascal GPUs
- Published results expected soon



Scaling Patterns

- Computation
  - Good parallelism
  - Icosahedral grid
  - Efficient algorithm
- Communications
  - Minimal frequency
  - Low data volume
  - Some overlapping

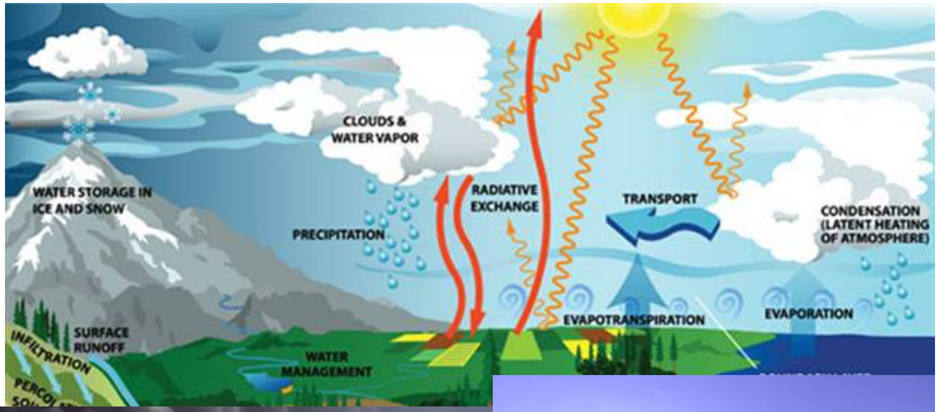


Weather Prediction Models

- physics -

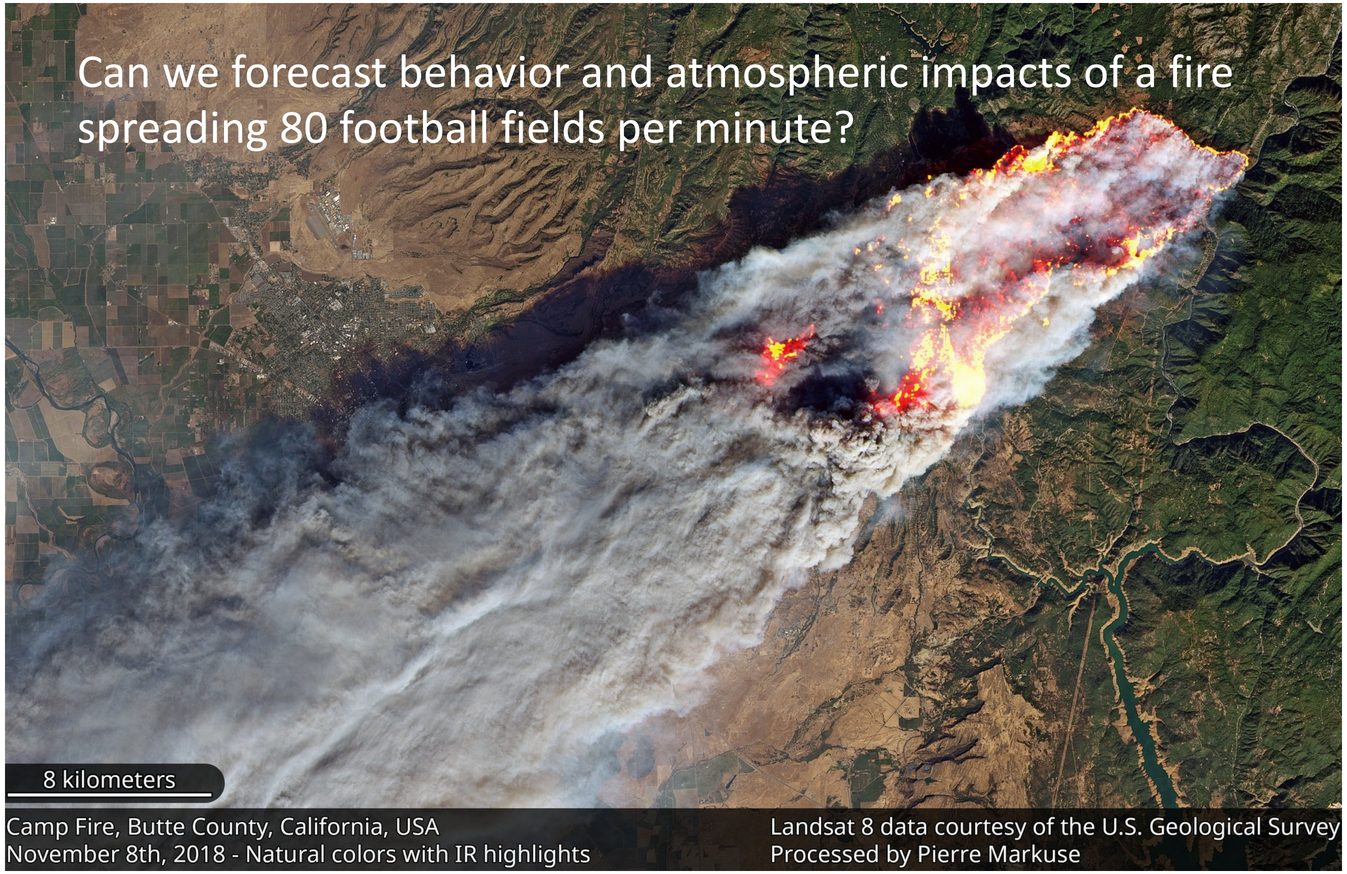
- Convection, radiation micro-physics, surface & boundary layers, gravity & orographic wave drag
- Computationally expensive, complex interactions, limited parallelism
- **Good potential for ML / DL**
  - Significantly faster than original code
  - Extensive training required for non-linear formulations

  - Krasnopolsky, V., A neural net emulator for microphysics schemes, 2017
  - O'Gorman, P., Using machine learning to parameterize moist convection, 2018

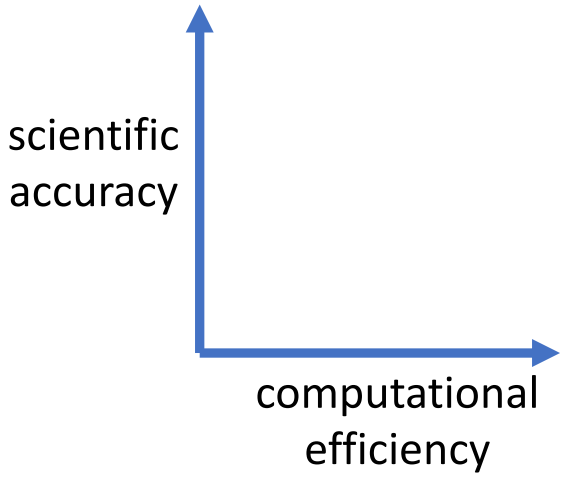


Weather Prediction Models
- chemistry -

- Simple to complex interactions
  - Fire weather
  - Air quality
- Computationally very expensive
  - 5X more than dynamics, physics
- Candidates for ML / DL
  - R.Ahmadov, J.Stewart, NOAA ESRL, Deriving relationships between weather and fire intensity from satellite data. *planned work*



#2 Improve Data Assimilation Performance



Assimilation

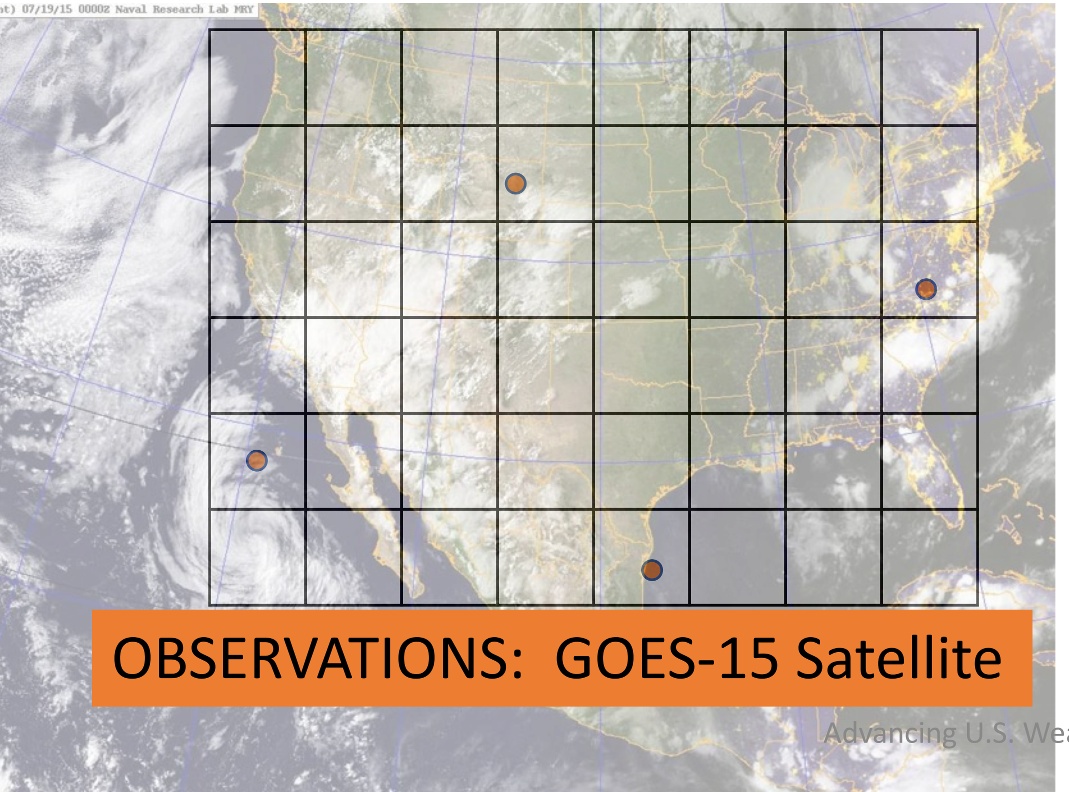
- Improve initial state of the forecast model
  - Variational, ensemble, hybrid approaches
- Complex, computationally expensive

<u>Calculations</u>

- Estimate model error, observation error
- Interpolate model to observation
- Adjust nearby grid points, other model fields (winds, temp, …)

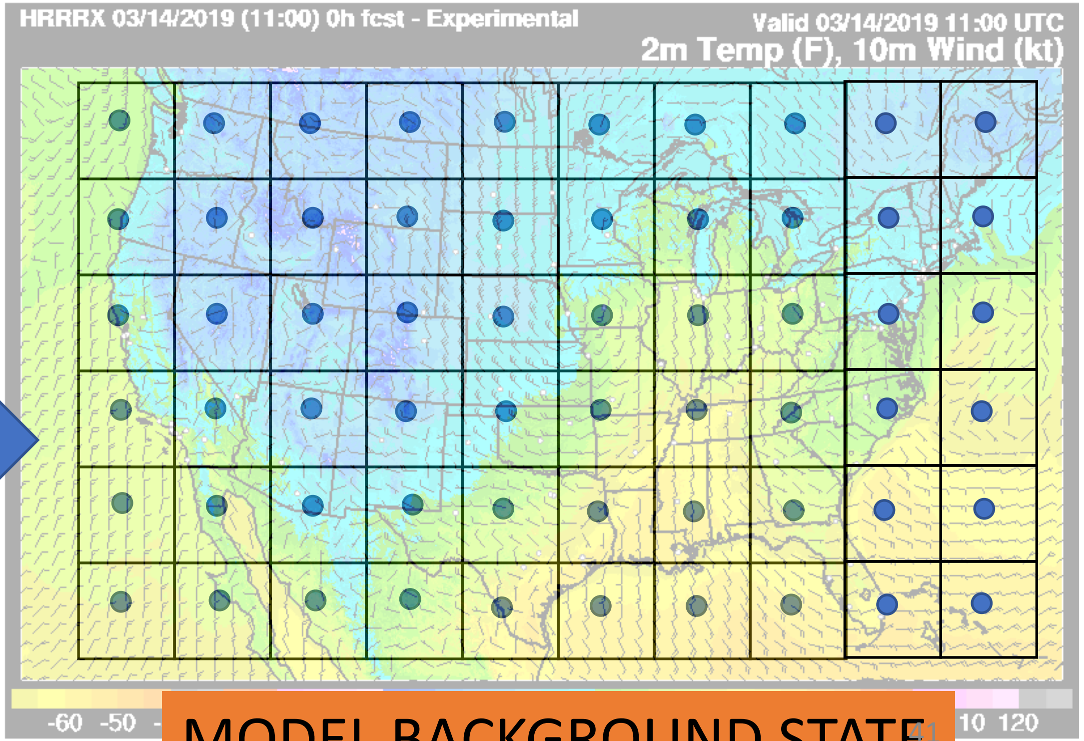
HRRR: 3 KM resolution, 2M temperature

GOES-15: 4 KM resolution IR, 1 KM visible
Assimilation can handle every 100th point



OBSERVATIONS: GOES-15 Satellite

update



MODEL BACKGROUND STATE

What Is JEDI?

A Next-Generation Unified Data Assimilation System

FV3 (GSI+GOES) (NOAA/NASA)

MPAS (NCAR)

NEPTUNE (NRL)

LFRic (UKMO)

MOM6 (JCSDA/NOAA)

...

JEDI

Radiosondes

Radiance (AMSU-A)

Aircraft

Aerosols (AOD)

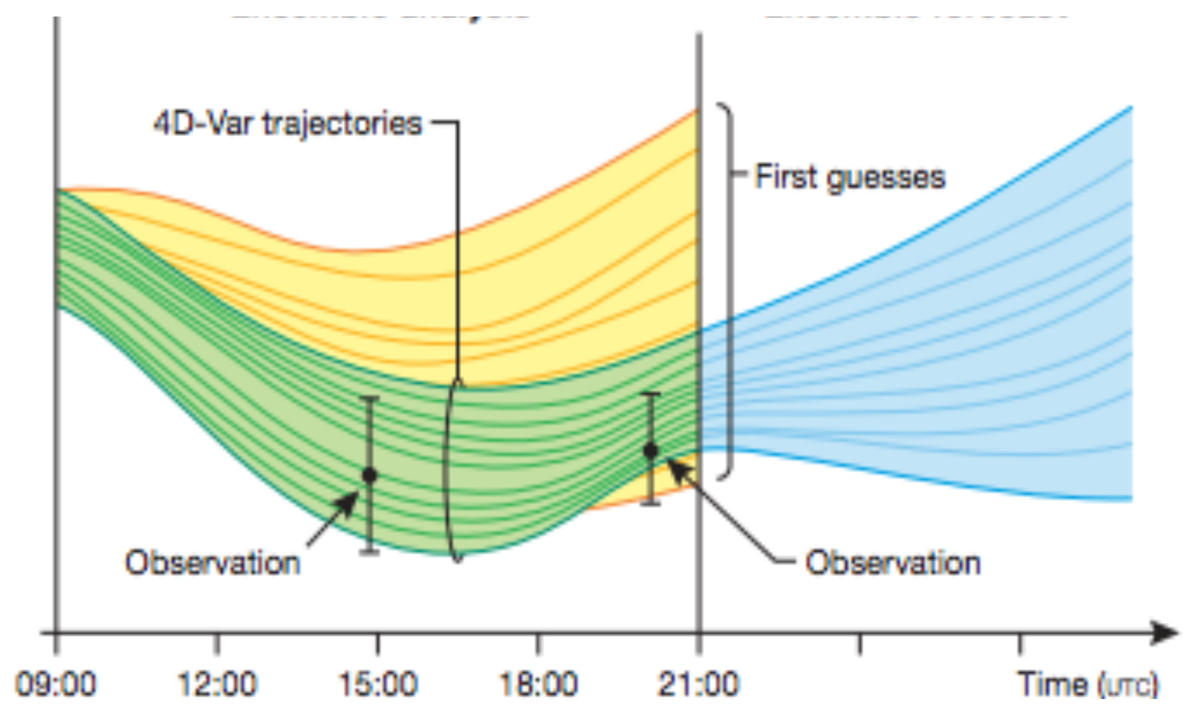
Sea Ice (fraction, thickness)

...

Slide, courtesy of JEDI project team, 2019

Data Assimilation Computational Issues

- 3D Ensemble Based Assimilation
  - 10-100 members, low resolution
  - **I/O, computational limitations**
- 4D Variational Assimilation
  - More accurate than ensemble methods
  - **~3X slower than 3DVAR methods**
- Investigating techniques to improve performance



Ensemble analysis and forecast cycle
Image courtesy of M.Bonavita (ECMWF)

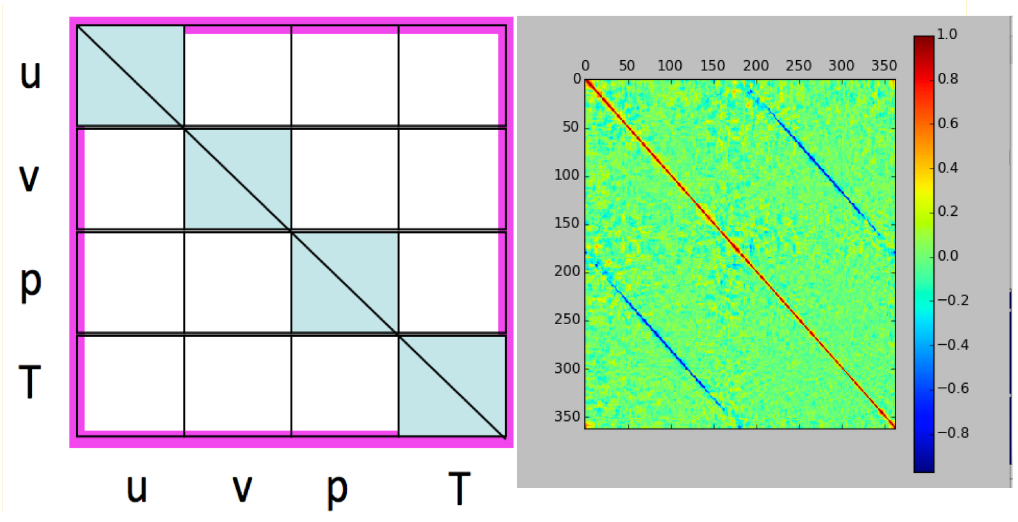
Advanced Data Assimilation Research

Isidora Jankov, Lidia Trailovic, Chris Harrop, NOAA ESRL/GSD, 2018-2019

The focus is on improving accuracy while maintaining/improving performance of DA systems

- JEDI activities
  - Shallow Water (SW) model with its Adjoint and Tangent Linear has been added to JEDI 4DVar suite
  - Testing of variety of features within JEDI framework
- Background Error Covariance (B) work
  - Improving accuracy by adjusting the B matrix localization

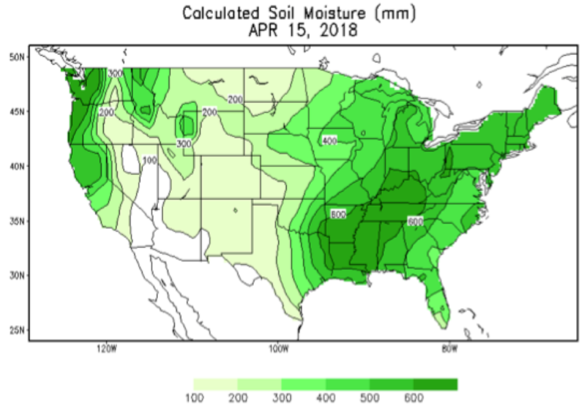
$$B \approx \frac{1}{N_e - 1} \sum_{k=1}^{N_e} \left(x_k - \bar{x}\right)\left(x_k - \bar{x}\right)^T$$



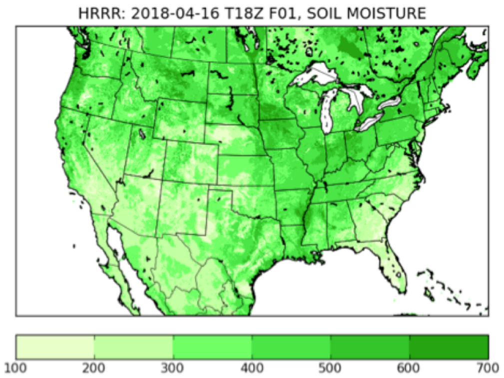
Use of Machine Learning for Improved Initial Soil Moisture State in RAP/HRRR

Isidora Jankov, Jebb Stewart, Lidia Trailovic, NOAA ESRL/GSD, 2018-2019

CPC



HRRR



- soil moisture field from CPC and HRRR for April 15, 2018
- similar features in the two data sets
- over Southeast U.S., CPC has higher values with a spatial pattern not present in HRRR
- potential room for improvement in HRRR representation of soil moisture.

Improvement of RAP/HRRR initial soil state field by using ML will be performed in two steps:

1) improve correlation between observed surface variables and soil state (currently used correlation is empirical and based on limited number of case studies)
2) 2) “nudge” the estimated soil moisture state by utilizing 10.3 micron channel from GOES-16/17 for the CONUS with a spatial resolution of 2 km and temporal resolution of 5 minutes

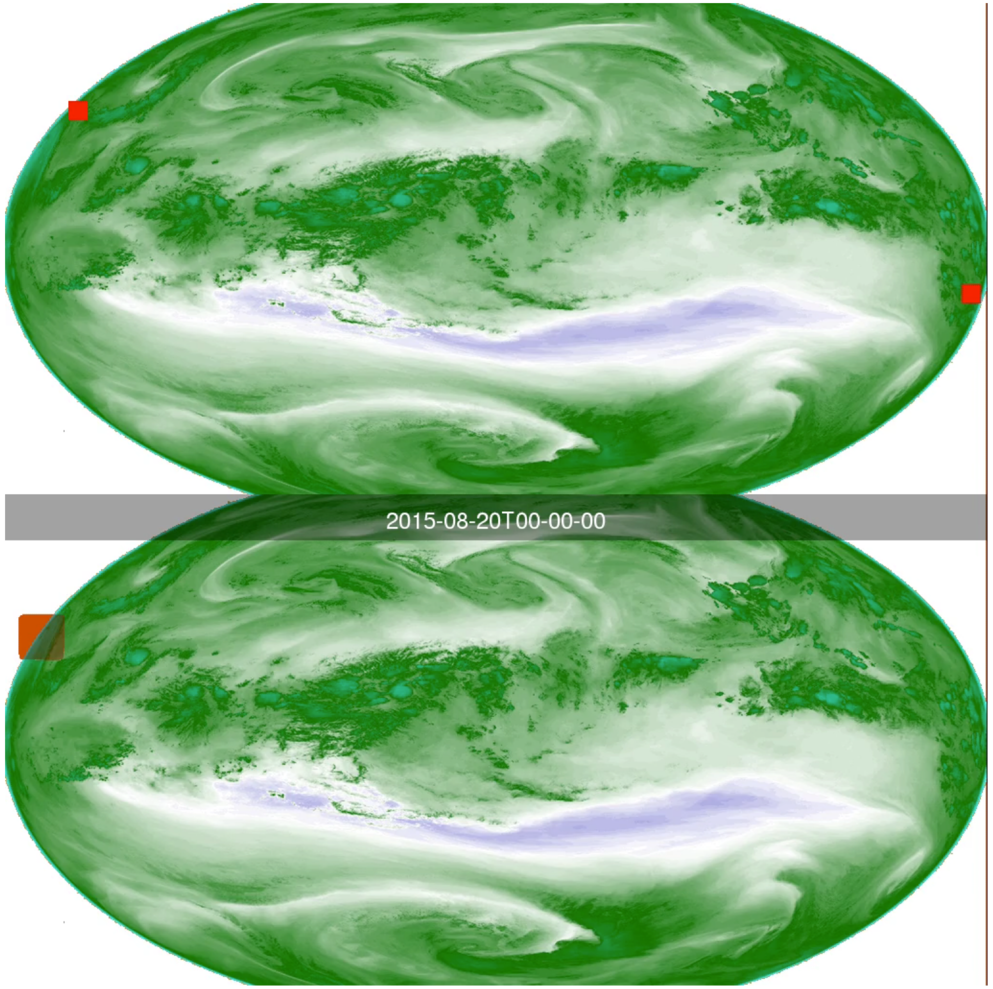
The effort will facilitate:

- more general use of the high-resolution GOES-16/17 ABI data set in data assimilation
- expansion of ML use in areas of Numerical Weather Prediction (NWP) and data assimilation.

Feature Detection – Typhoons

Christina Bonfanti, Jebb Stewart, NOAA ESRL/GSD, 2018-2019

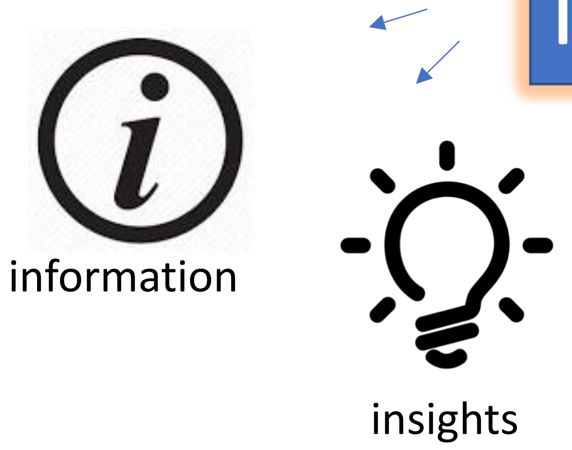
- Identify typhoons from satellite data
  - Accurate Identification
  - Early detection – prior to formation
- Training - 6 years of data
  - Model output, satellite
  - 11.5 hours (CPU), 3 minutes (GPU)
  - 5 weeks (CPU), 3 hours (GPU)
- Inference
  - 1 second (CPU), 40 milliseconds (GPU)

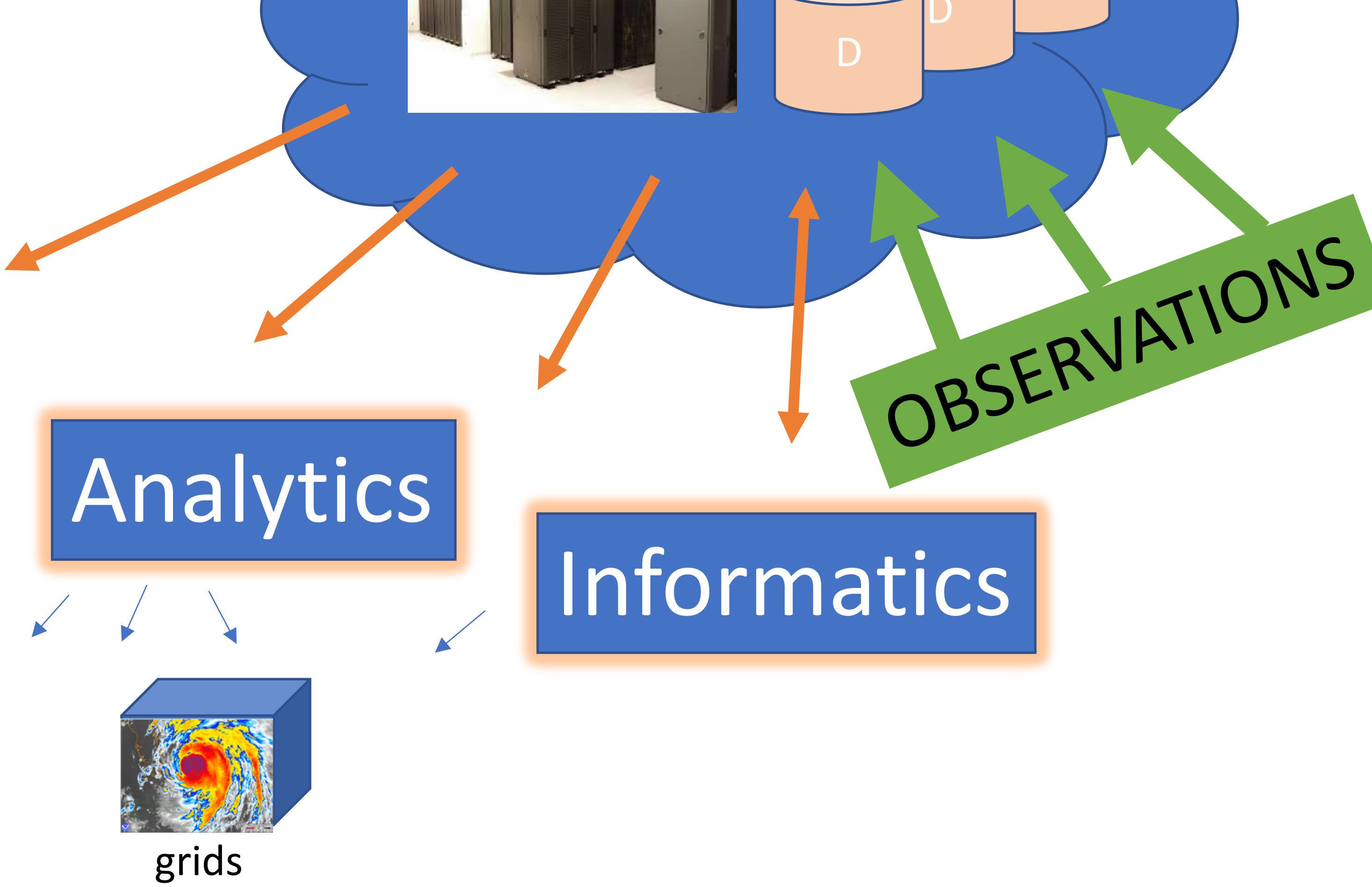


#3 Getting Data to End-Users

Big Data Handling

- Data is too big to move
  - Co-locate HPC & data
  - On-demand access
  - ML/DL driven analytics

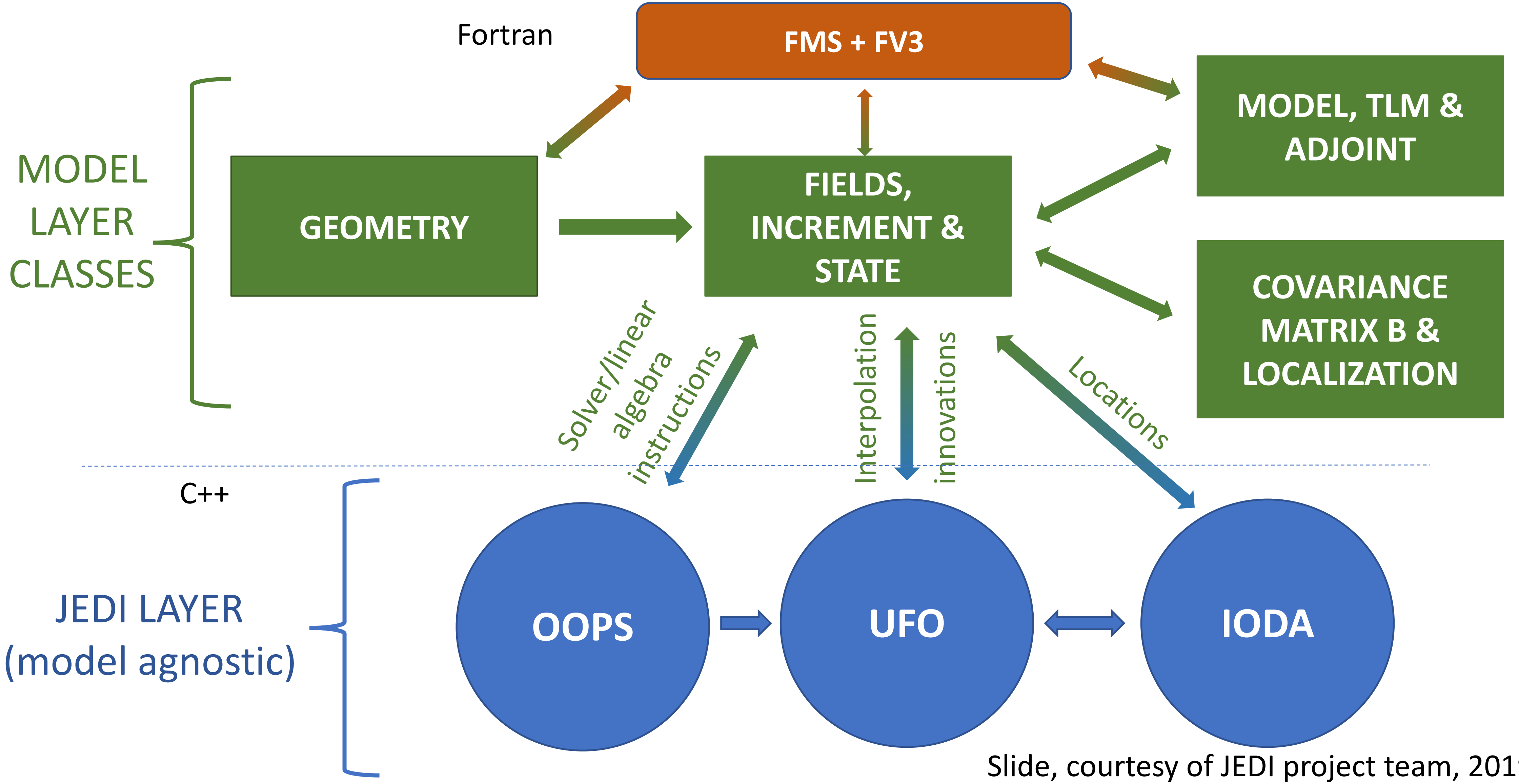




#4 Improve Software Architecture and Development Process

Design
Develop
Test
Commit
Refactor

JEDI System Software Architecture



Slide, courtesy of JEDI project team, 2019

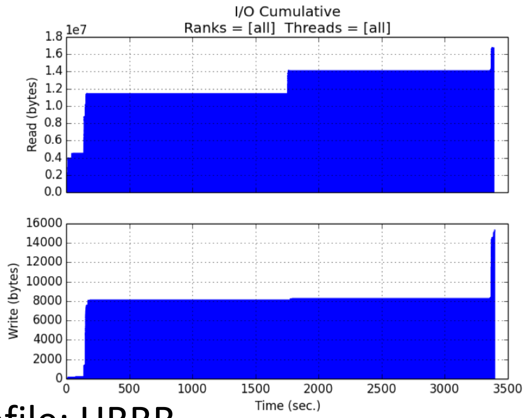
Conclusion

- Described challenges in current prediction system
  - Computer hardware, applications, data volume, software
- Tremendous opportunity with convergence of HPC, Big Data and AI
  - NVIDIA GPUs are a key technology
- I/O challenges, sensor networks, distributed assimilation not discussed
- Early in investigation of AI applied to weather prediction
  - David Hall, NVIDIA, "Deep Learning for Improved Utilization of Satellite Data in Weather Forecasting", Tuesday 10:00 – 11:00
  - Sid Boukabara, NOAA, "Exploring using Artificial Intelligence for Remote Sensing, NWP and Situational Awareness", ITSC-XXI Conference, November 2017
  - Jebb Stewart, NOAA, Organizing committee, NOAA AI Conference, April 2019

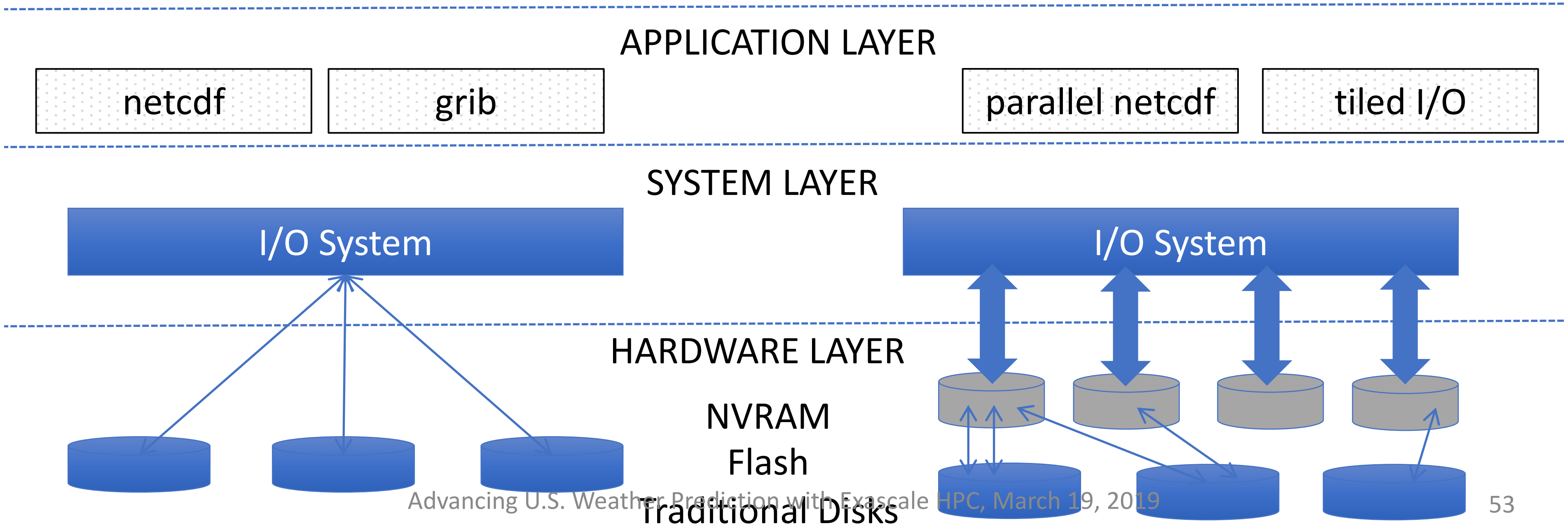
Additional Slides

I/O Dwarf

- Configurable application to mimic model, DA I/O patterns
  - Realistic projections for exascale
    - 3KM global, 50 - 100 ensembles, hourly output
- Test & tune on our HPC systems
- Share with vendors
- Use for HPC procurements



I/O Profile: HRRR



I/O - Impact of NVRAM on Data Access

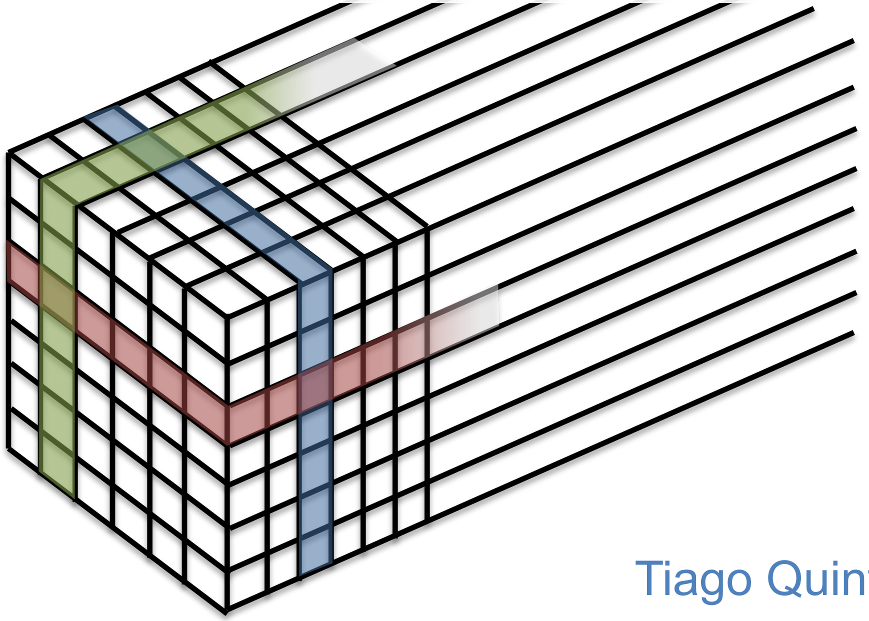
Byte Addressable Hypercubes

- Longitude (3600)
- Latitude (1800)
- Atmospheric levels, Physical parameters (~200)
- Time steps (~100)
- Probabilistic pertubations (50)

@ double precision

- 9km **48 TiB**
- 5km **192 TiB**
- 1.25km **1.82 PiB**

Clients want to do **different** analytics across **multiple** axis



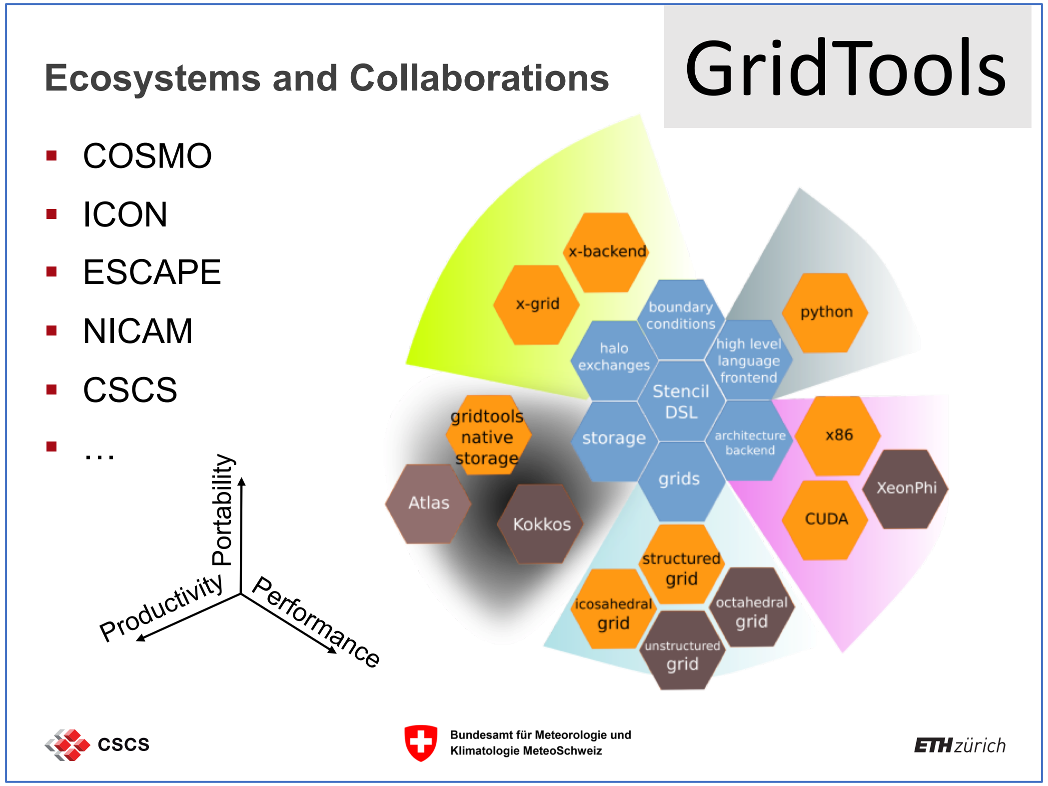
Tiago Quintino

Not included: *historical observations, multiple models, etc...*

ECMWF archives ~150TB / Day
Growing exponentially …

Portability

- Directives
  - OpenACC
  - OpenMP
- Libraries
  - MPI, netCDF
- Tools
  - GridTools (CSCS)
  - PSyclone (Ukmet)
  - ATLAS (ECMWF)

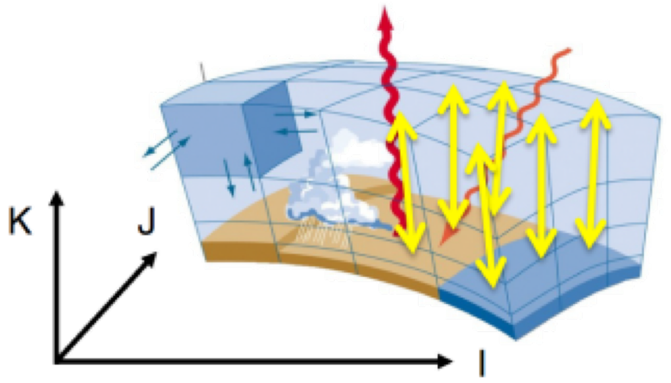


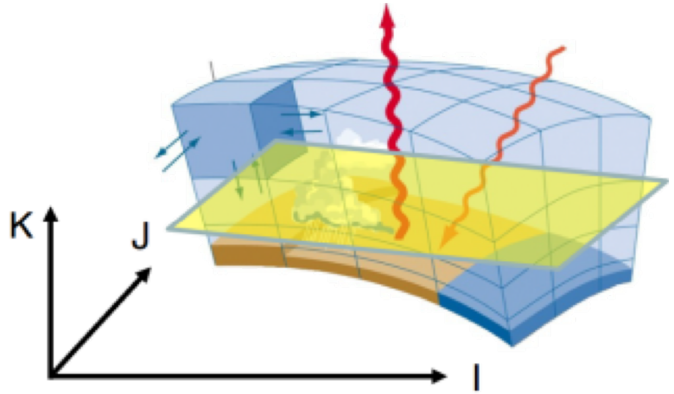
Slide courtesy of Oliver Fuhrer, CSCS

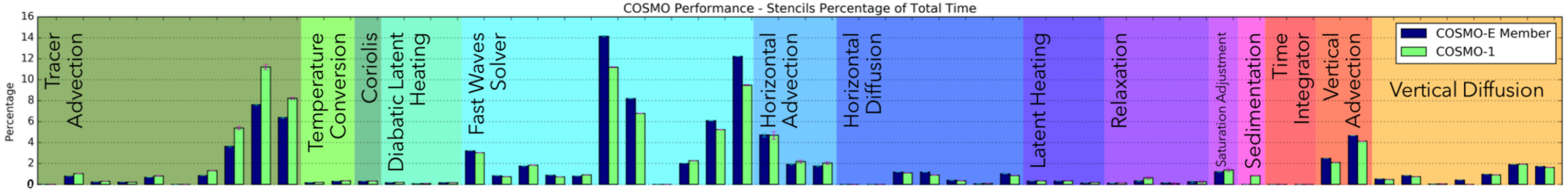
Grid Tools

Algorithmic Motifs

- Regular and Structured grids
  - **Algorithmic 3D stencils** (almost)
  - Parallelism on the first 2 dimensions
  - Dependencies on the third
    - Parallel, Forward, Backward
  - Reductions
  - **General boundary conditions**
  - **Halo-update**
  - Combination of BC and Comm







cscs

Bundesamt für Meteorologie und Klimatologie MeteoSchweiz

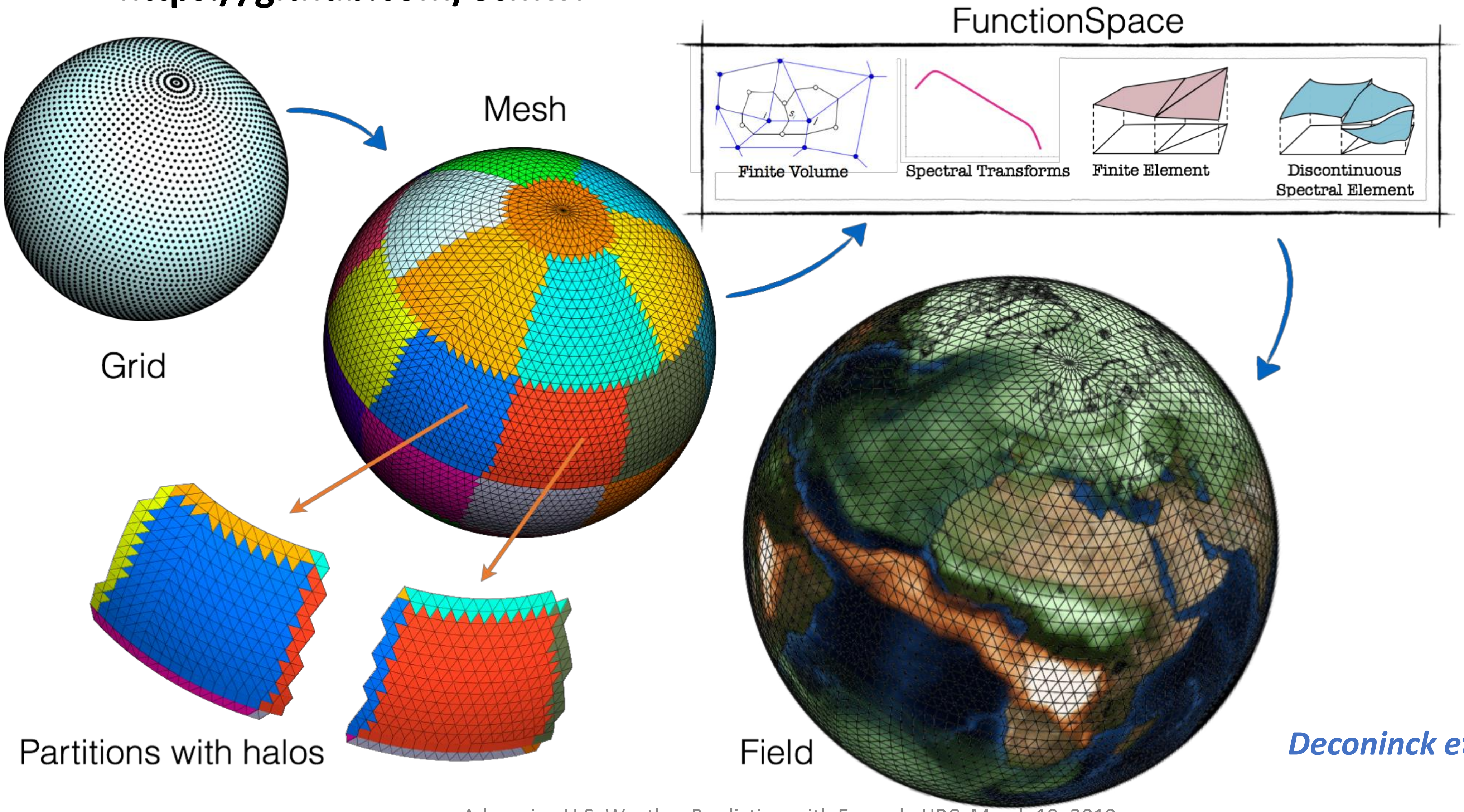
3

ETH zürich

Slide courtesy of Oliver Fuhrer, CSCS

Atlas: a library for NWP and climate modelling

https://github.com/ecmwf



Deconinck et al. 2017