



Extreme Weather, Machine Learning, and High-Performance Computing

Daniel Duffy daniel.q.duffy@nasa.gov and on Twitter @dqduffy

Michael Bowen michael.k.bowen@nasa.gov

High Performance Computing Lead for the
NASA Center for Climate Simulation (NCCS – <http://www.nccs.nasa.gov>)
Goddard Space Flight Center (GSFC)

NASA High-End Computing Program



HEC Program Office
NASA Headquarters
Dr. Tsengdar Lee
Scientific Computing Portfolio Manager
<http://www.hec.nasa.gov/>

High-End Computing Capability (HECC) Project
NASA Advanced Supercomputing (NAS)
NASA Ames
Dr. Piyush Mehrotra
<https://www.nas.nasa.gov/hecc/>

NASA Center for Climate Simulation (NCCS)
Goddard Space Flight Center (GSFC)
Dr. Daniel Duffy
<http://www.nccs.nasa.gov/>

NASA Center for Climate Simulation (NCCS)



Provides an 'integrated' -end computing environment designed to support the specialized requirements of Climate and Weather modeling.

- High-performance computing, data storage, and networking technologies
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services
- Advanced Data Analytics Platform (ADAPT)

Primary Customers (NASA Climate Science)

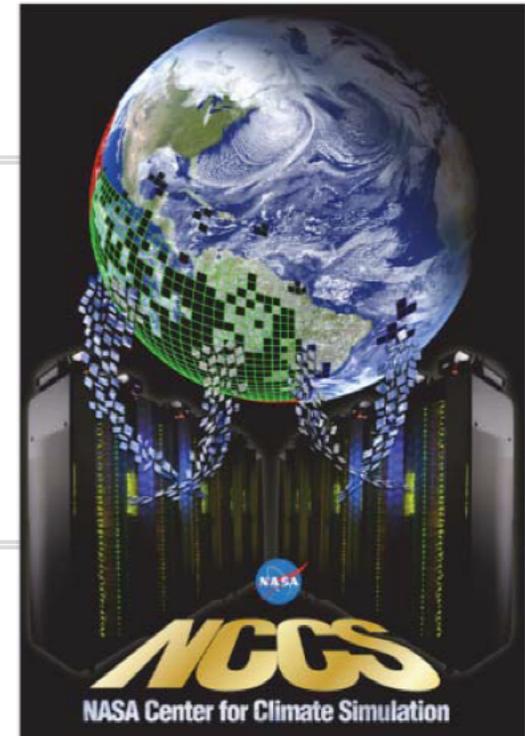
- Global Modeling and Assimilation Office (GMAO)
- Land Information Systems (LIS)
- Goddard Institute for Space Studies (GISS)
- Variety of other Research and Development (R&D)

High-Performance Science

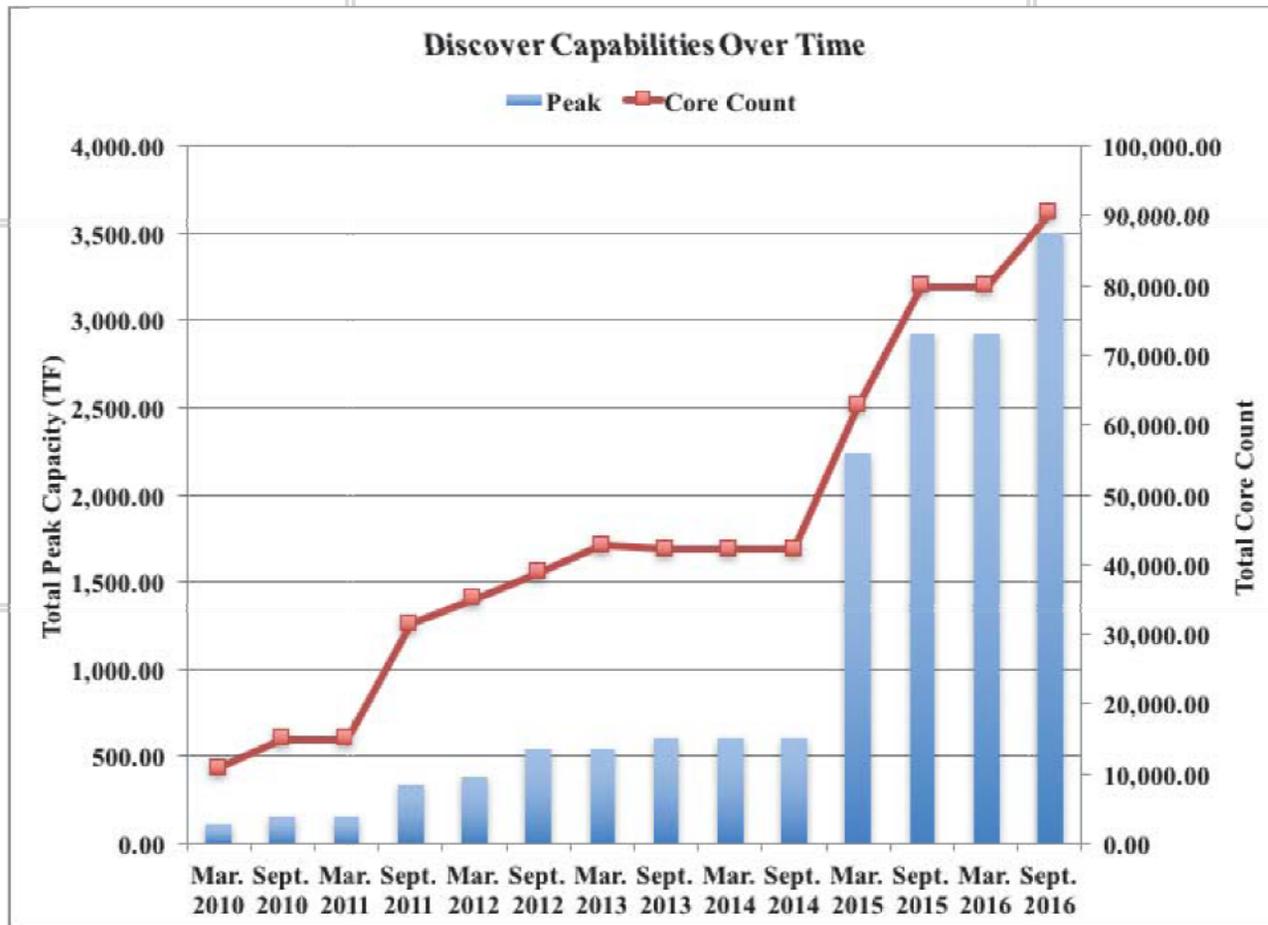
- <http://www.nccs.nasa.gov>
- Code 606.2
- Located at NASA Goddard Space Flight Center in Greenbelt, MD.

HPC WIRE 2016 Readers Choice Awards

Reader's Choice: Best Data-Intensive System (End User focused) for ADAPT!



Discover (HPC) Capacity Evolution



This includes traditional x86 processors only; does not include GPUs or Phis.

Plans in progress to expand compute capacity in FY17 and FY18 with (1) SkyLake processors, (2) small amount of deep learning systems.

Discover (HPC) Scratch Disk Evolution



Calendar	Description	Decommission	Total Usable Capacity (TB)
2012	Combination of DDN disks	None	3,960
Fall 2012	NetApp1: 1,800 by 3 TB Disk Drives; 5,400 TB RAW	None	9,360
Fall 2013	NetApp2: 1,800 by 4 TB Disk Drives; 7,200 TB RAW	None	16,560
Early 2015	DDN10: 1,680 by 6 TB Disk Drives, 10,080 TB RAW	DDNs 3, 4, 5	~26,000
Mid 2015	DDN11: 1,680 by 6 TB Disk Drives, 10,080 TB RAW	DDNs 7, 8, 9	~33,000
Mid 2016	DDN12: 1,680 by 6 TB Disk Drives, 10,080 TB RAW	None	~40,000
Early 2017	13+ PB RAW, TBD	TBD	~50,000

- Usable capacity differs from raw capacity for two reasons. First, the NCCS uses RAID6 (double parity) to protect against drive failures. This incurs a 20% overhead for the disk capacity. Second, the file system formatting is estimated to also need about 5% of the overall disk capacity. The total reduction from the RAW capacity to usable space is about 25%.



NCCCS Evolution of Major Systems

FY15

Data Portal

Mass Storage

HPC - Discover

FY16: Creation of the Advanced Data Analytics Platform (ADAPT), a High Performance Science cloud (virtual environment) designed for traditional data services, data analytics, and web services: move the data to the analysis.

FY16

ADAPT

Mass Storage

HPC - Discover

FY17: Creation of the Data Analytics Storage Service (DASS), a combined High Performance Computing and Data environment to enable emerging analytics: move the analysis to the data.

FY17

ADAPT

Mass Storage

DASS

HPC - Discover

Come to the NASA Booth to see Carrie Spear's presentation about the Data Analytics Storage Service (DASS)!

Godunov Earth Observing System (GEOS) Model NASA Global Modeling and Assimilation Office (GMAO)

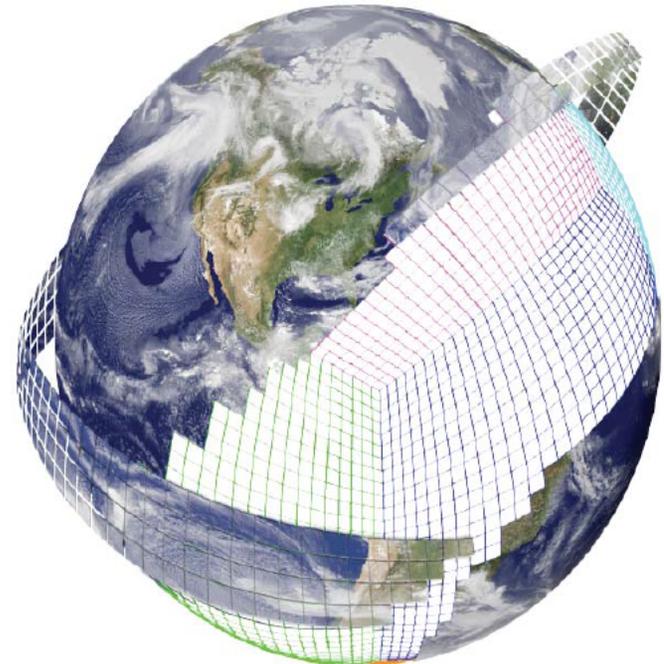


FV3 Dynamical Core uses a Cubed-Sphere which maps the Earth onto faces of a cube

- There are 6 faces of the cube and multiple vertical layers
- Total number of grid points
 - $X * Y * Z * 6$ Faces of the Cube

Current GMAO Research

- Operational research forecasts are running at 27 KM resolution using about 27 million grid points
- Target operational research forecasts at a resolution of 12 KM in the very near future
- Reanalysis (including chemistry)
- Dynamic downscaling of reanalysis and forecasts down to 6 KM
- Highest resolution research runs are at 1.5 KM global resolution



Want to see more about the GEOS model?

Come see talks at the NASA booth by Bill Putman, Matt Thompson, and Ben Auer.

Increasing the GEOS-5 Model Resolution



Target – Run approximately 100 meter global resolution research runs in 10 to 15 years
Each doubling of resolution requires 4x the grid points in the (x, y) direction; assume number of vertical layers are a constant at 132

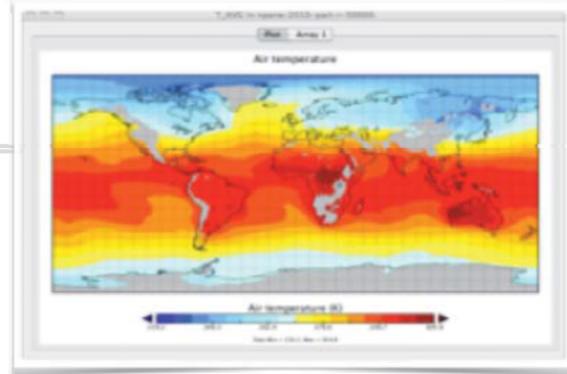
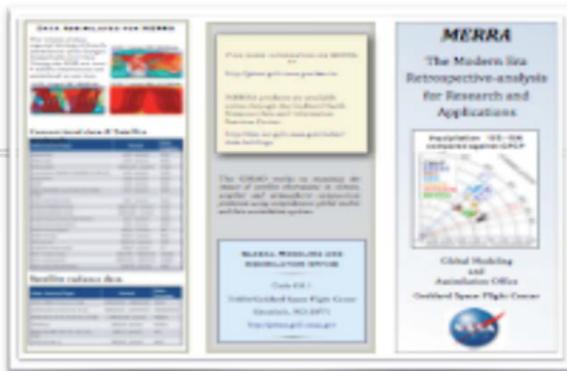
Model	X and Y Values	Grid Points	Resolution (meters)	Cores	RAM (PB)
C1440	5,760	26 x 10 ⁹	1,736	30,000	0.12
C2880	11,520	105 x 10 ⁹	868	120,000	0.48
C5760	23,040	420 x 10 ⁹	434	480,000	1.92
C11520	46,080	1,682 x 10 ⁹	217	1,920,000	7.68
C23040	92,160	6,727 x 10 ⁹	109	7,680,000	30.72

Bad News – This is only one component of the application (the atmosphere). GMAO is working on their coupled model including Atmosphere, Ocean, Waves, Ice, and More; We expect the model to require much more memory pushing us toward a higher memory to flop ratio.



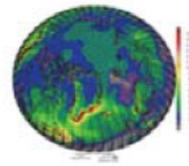
MERRA2 Data Set

MERRA2 Reanalysis



Modern Era-Retrospective Analysis for Research and Applications

- Source: Global Modeling and Assimilation Office (GMAO)
- Input: observation types (land, sea, air, space) into “frozen” numerical model. (~4 million observations/day)
- Output: a global temporally and spatially consistent synthesis of many key climate variables, including aerosols
- Spatial resolution: 1/2° latitude × 2/3° longitude × 42 vertical levels extending through the stratosphere.
- Temporal resolution: 6-hours for three-dimensional, full spatial resolution, extending from 1978–Present.
- ~ 400 T, and growing



CMIP5	MERRA	Units	ESGF MERRA published variables Description(Long Name)
rlus	rlus	W m-2	Surface Upwelling Longwave Radiation
rlut	lwtup	W m-2	TOA Outgoing Longwave Radiation
rlutcs	lwtupclr	W m-2	TOA Outgoing Clear-Sky Longwave Radiation
rsds	swgnt	W m-2	Surface Downwelling Shortwave Radiation
rsdscs	swgdncr	W m-2	Downwelling Clear-Sky Shortwave Radiation
rsdt	swtdn	W m-2	TOA Incident Shortwave Radiation
rsut	swtdn??	W m-2	TOA Outgoing Shortwave Radiation
clt	cldtot	%	Total Cloud Fraction
pr	prectot	kg m-2 s-1	Precipitation
cl	cloud	%	Cloud Area Fraction
evpsubl	evap	kg m-2 s-1	Evaporation
hfls	elux	W m-2	Surface Upward Latent Heat Flux
hfs	hflux	W m-2	Surface Upward Sensible Heat Flux
hur	rh	%	Relative Humidity
hus	qv	v	Specific Humidity
prc	precon	kg m-2 s-1	Convective Precipitation
prsn	precno	kg m-2 s-1	Snowfall Flux
nrw	tqv	kg m-2	Water Vapor Path
ps	ps	Pa	Surface Air Pressure
psl	slp	Fa	Sea Level Pressure
rls	lwgnt	W m-2	Surface Downwelling Longwave Radiation
rlsdc	lwgncr	W m-2	Surface Downwelling Clear-Sky Longwave Radiation
rsutcs	swtdn	W m-2	TOA Outgoing Clear-Sky Shortwave Radiation
ta	t	K	Air Temperature
tas	t2m	K	Near-Surface Air Temperature
taue	taux	Pa	Surface Downward Eastward Wind Stress
tauv	tauy	Pa	Surface Downward Northward Wind Stress
tro3	o3	1.00E-09	Mole Fraction of O3
ts	ts	K	Surface Temperature
ua	u	m s-1	Eastward Wind
uas	u10m	m s-1	Eastward Near-Surface Wind
va	v	m s-1	Northward Wind
vas	v10m	m s-1	Northward Near-Surface Wind
wap	omega	Pa s-1	omega (=dp/dt)
zg	h	m	Geopotential Height



**Enough already,
show some movies!**

The Questions

How can we study the evolution of extreme weather events with machine learning?

What effect might machine learning algorithms have on future HPC architectures?

For this study, we used a machine learning technique to attempt to find commonalities in how hurricanes evolve and dissipate using MERRA2 reanalysis data.



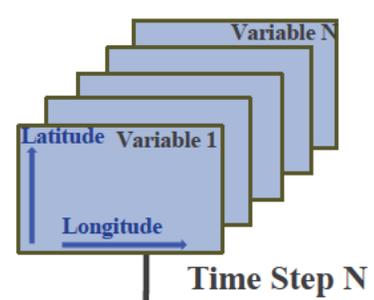
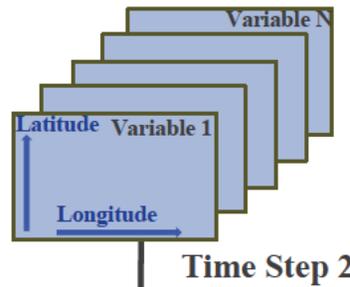
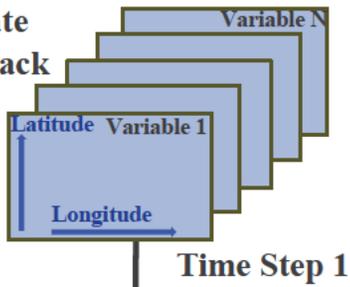
Machine Learning Algorithm



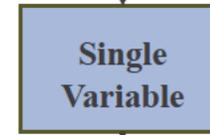
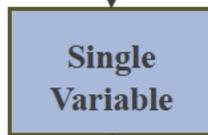
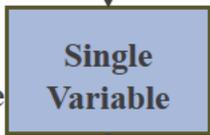
- 1. Identify hurricanes using IBTrACS database**
 - To reduce the data set to a manageable amount of data, we focused on the Western Pacific and only included the months of August and September
- 2. Extract relevant variables from the MERRA2 data set to create a hurricane stack – time series of multi-variate data for each storm**
- 3. Compute the spatial average for each variable at each time step**
- 4. Use a Symbolic Aggregate approXimation (SAX) discretization method to create a representation of the time series**
 - Each variable over time was discretized into a word (series of representative letters)
 - All variables (words) are combined together to make a paragraph for each storm
- 5. Paragraphs were compared using the SAX distance metric, similar to Euclidean distances**
- 6. The resulting distance metrics were clustered using standard clustering techniques**
 - This resulted in three clusters of hurricanes.
- 7. Compare the probability distribution functions of the variables in the three clusters.**



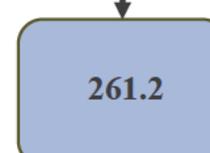
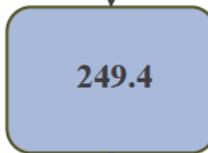
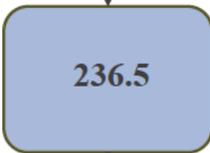
**Multi-Variate
Hurricane Stack**



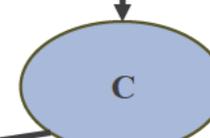
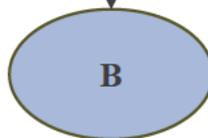
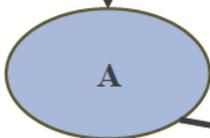
**Single-Variate
Hurricane Frame**



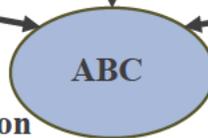
**Spatial average
of single variable**



**Symbolic
Representation**



**Word
Representation**

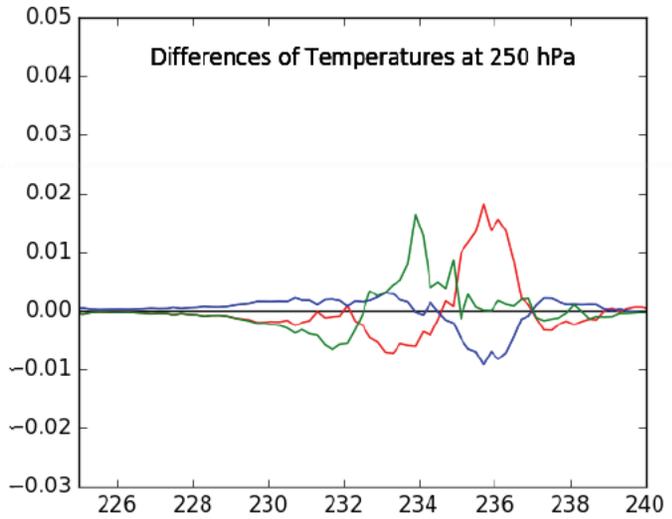
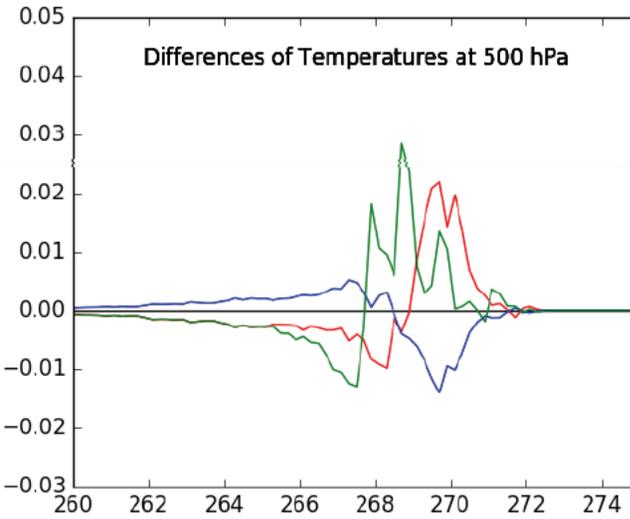
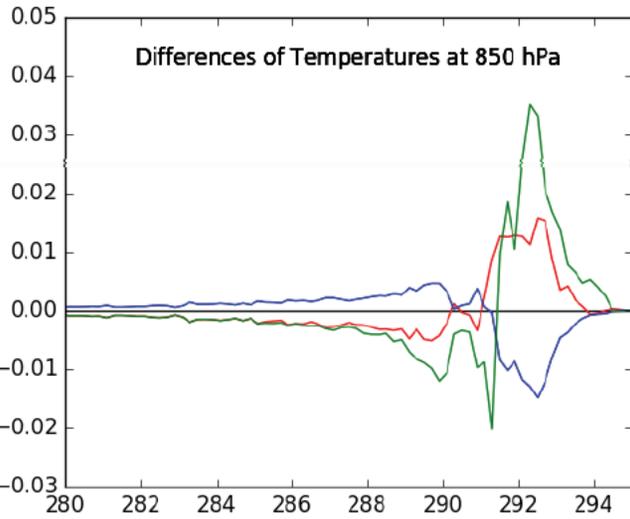
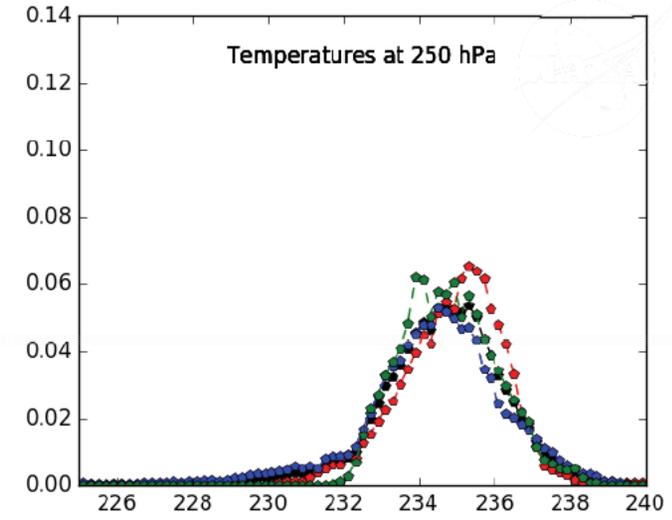
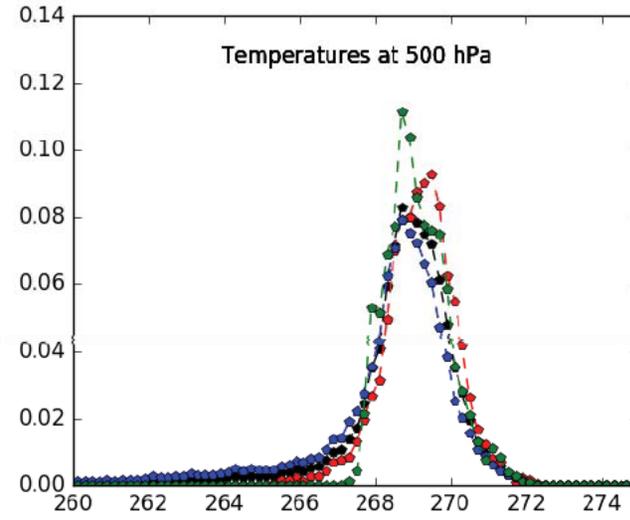
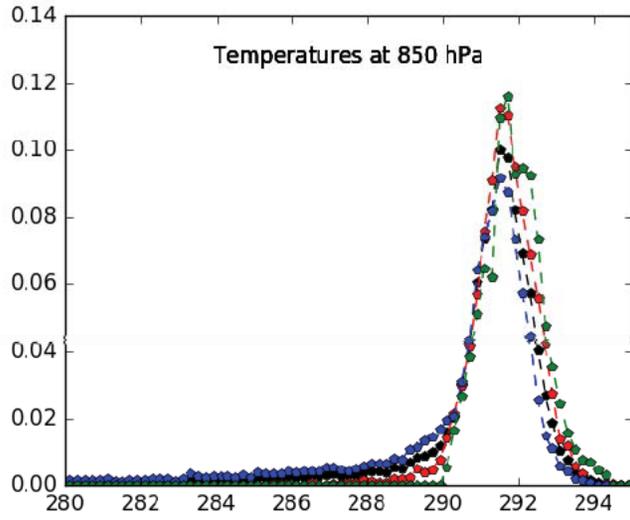


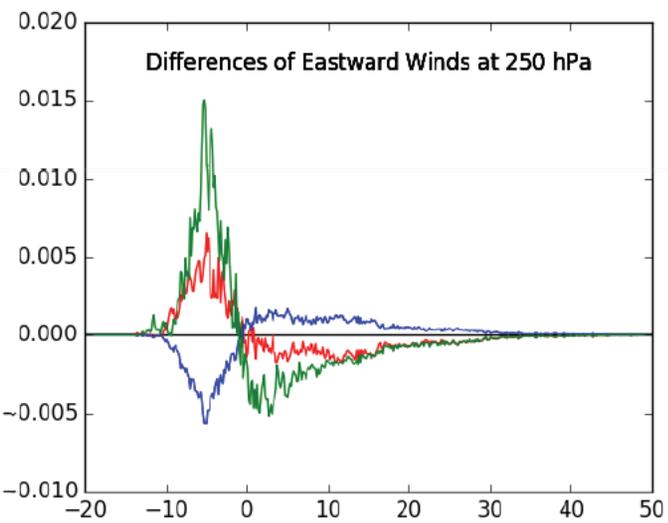
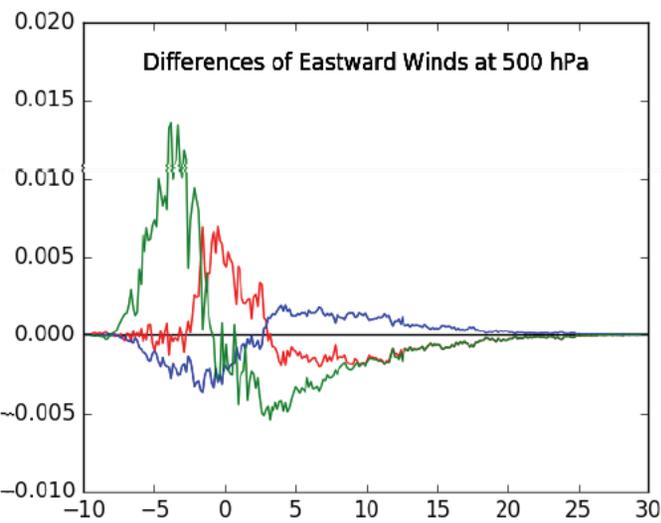
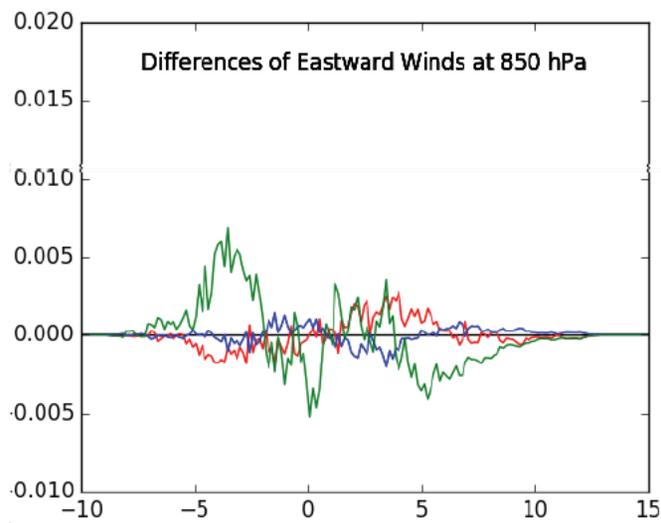
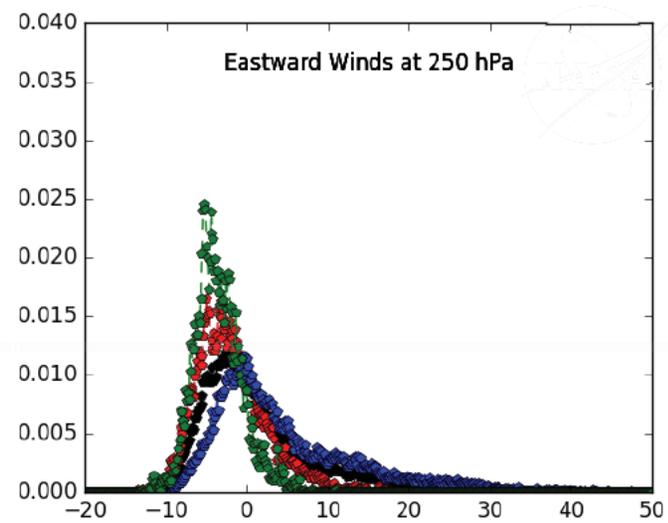
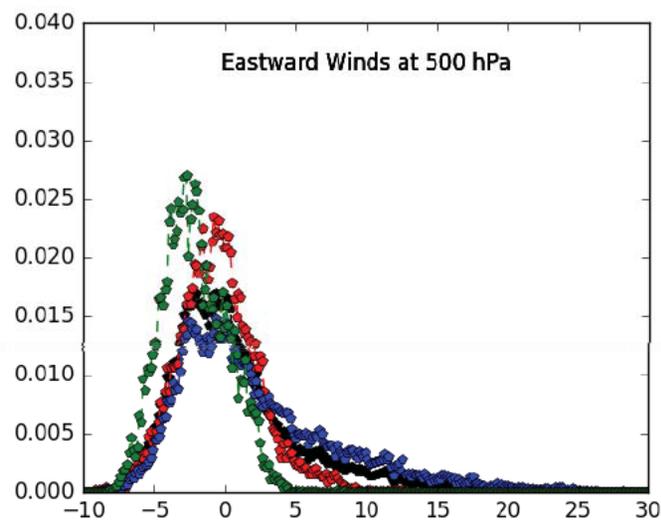
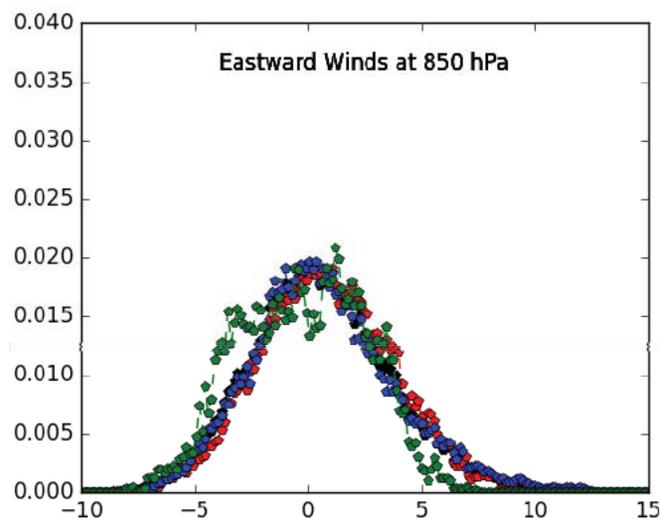
**Multiple variables together
make a paragraph.**

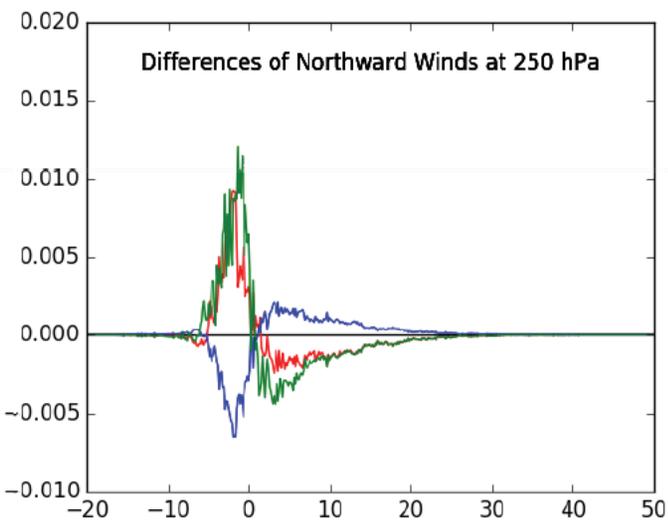
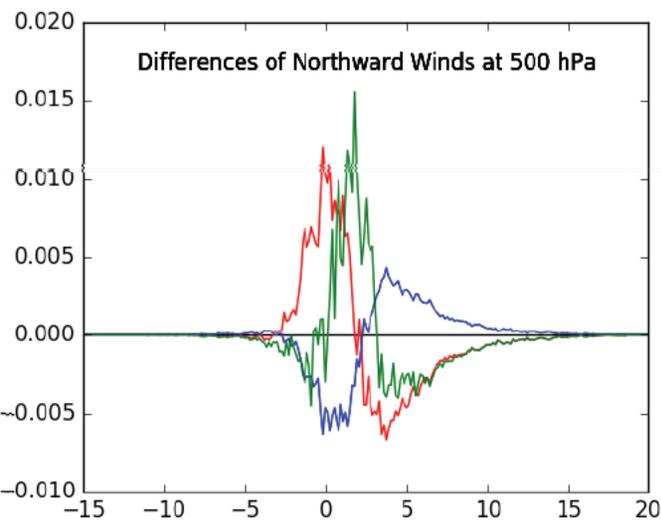
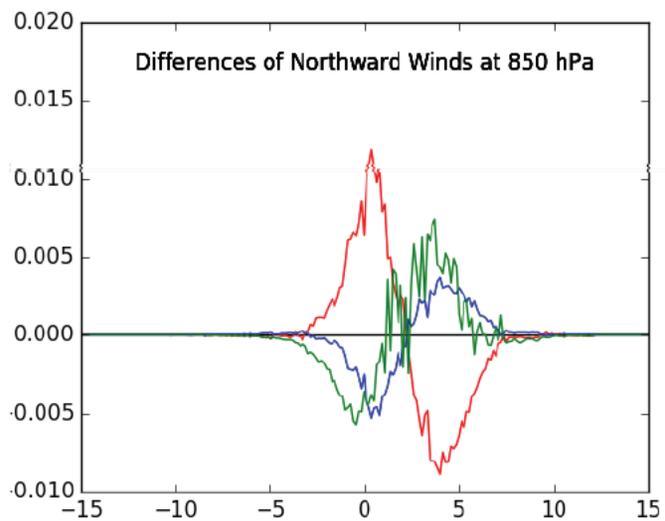
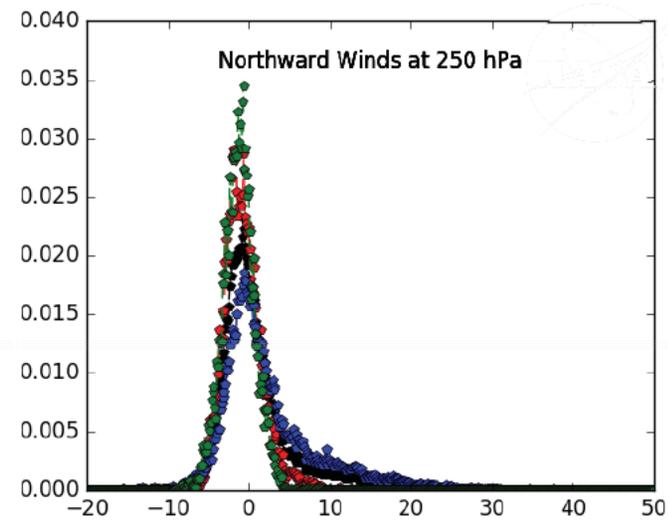
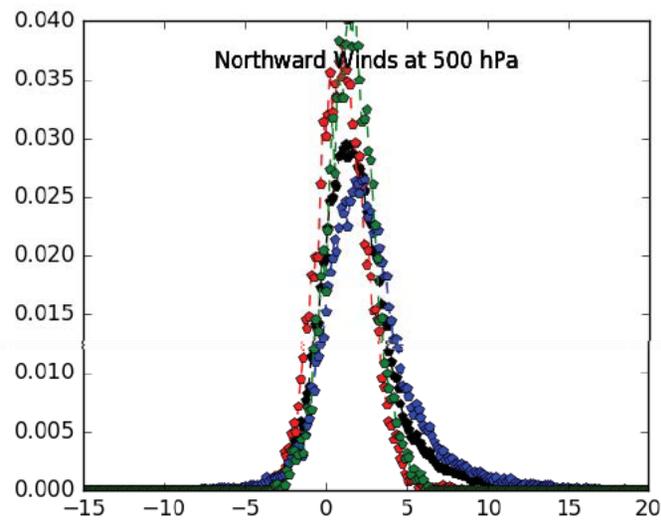
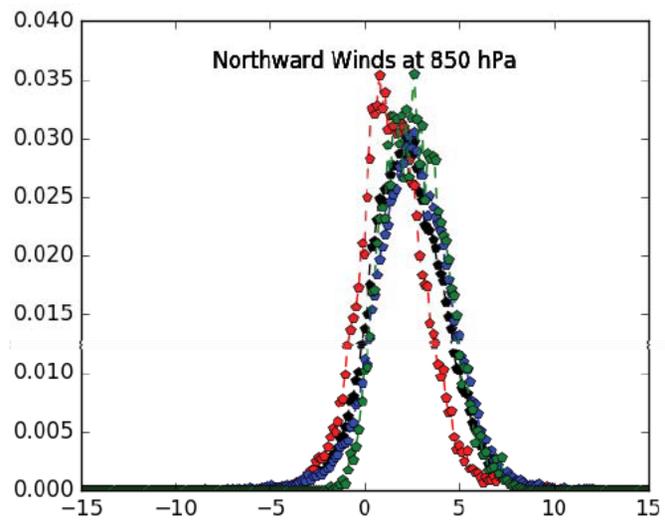
Variables of Interest

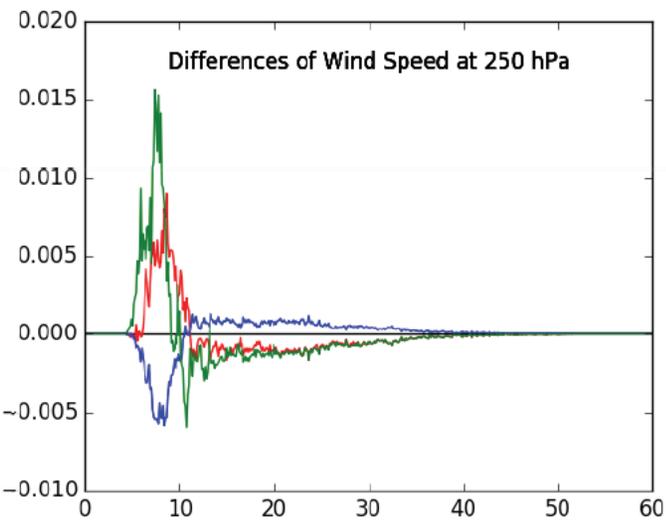
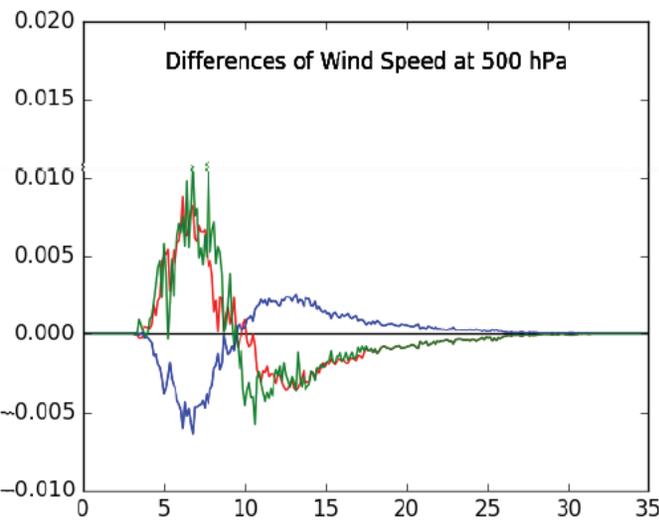
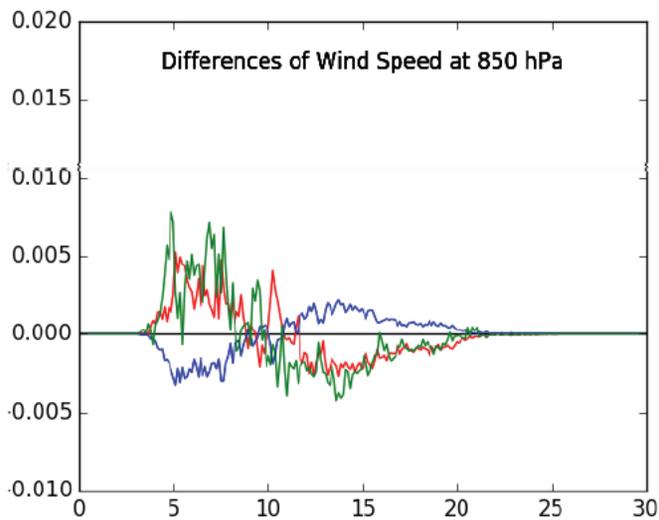
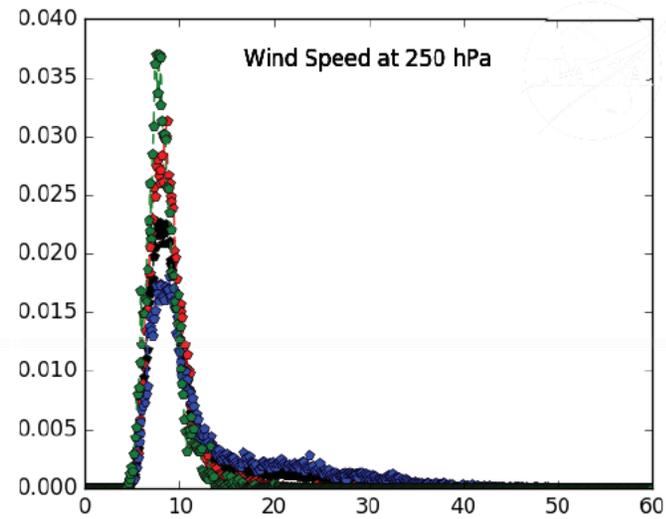
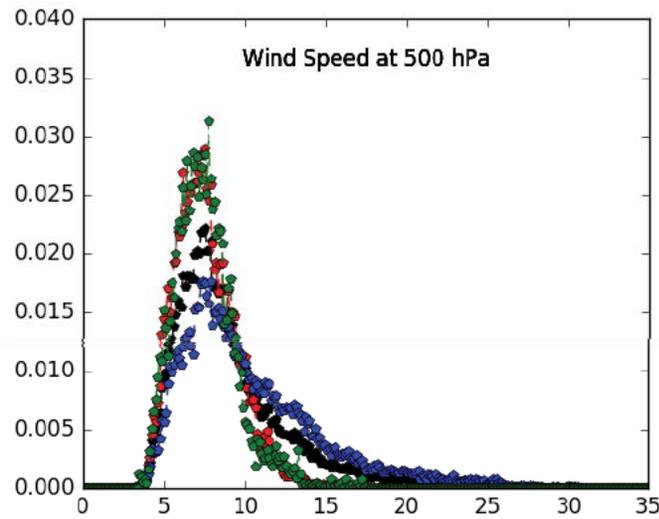
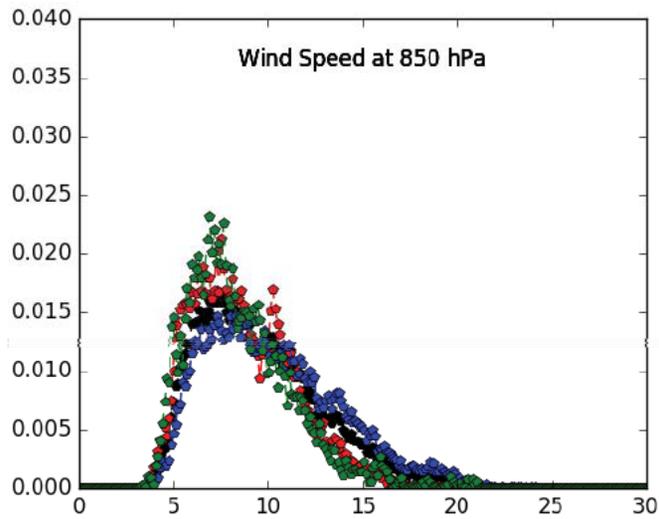


Variable Name	Description	Unit
T850	Air temperature at 850 hPa	Kelvin
T500	Air temperature at 500 hPa	Kelvin
T250	Air temperature at 250 hPa	Kelvin
U850	Eastward wind speed at 850 hPa	meters/second
U500	Eastward wind speed at 500 hPa	meters/second
U250	Eastward wind speed at 250 hPa	meters/second
V850	Northward wind speed at 850 hPa	meters/second
V500	Northward wind speed at 500 hPa	meters/second
V250	Northward wind speed at 250 hPa	meters/second









What Next?



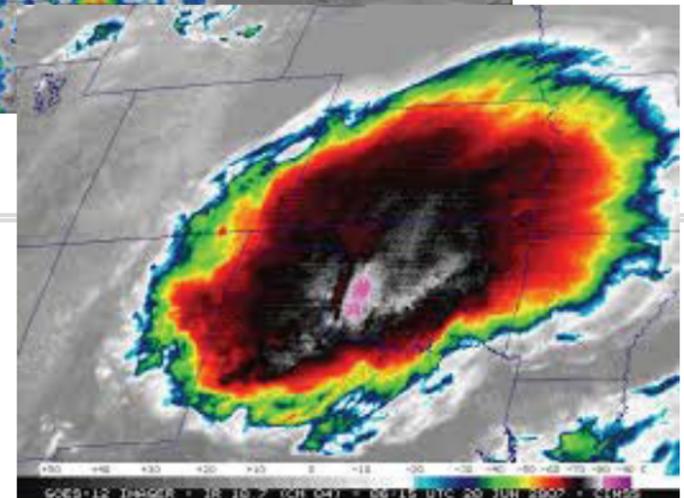
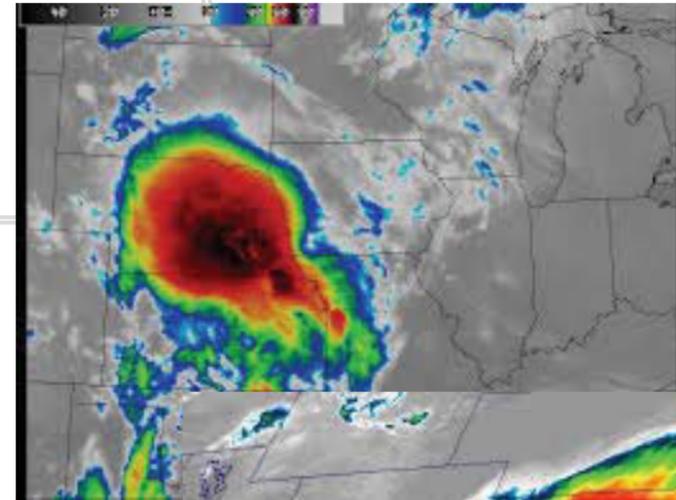
Looking at a method to apply a dynamic time-warping algorithm to the data to mitigate the effects of different lengths of storms on the clusters

Expand to include the full hurricane seasons in both the Atlantic and Pacific

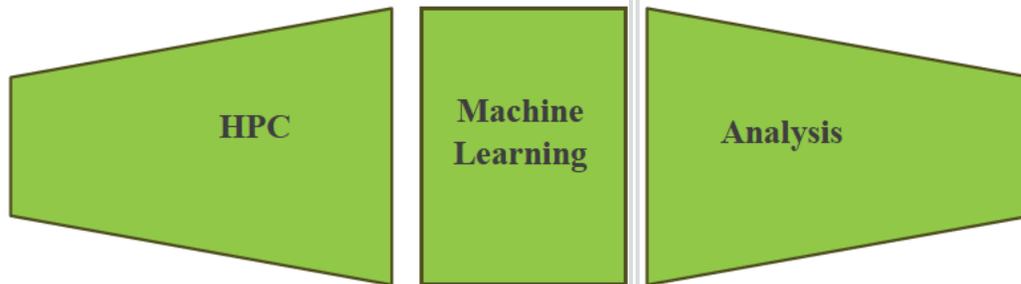
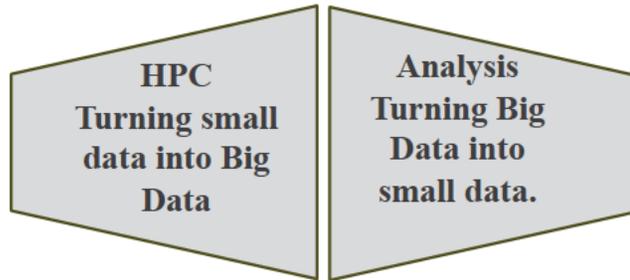
Include additional variables into the study

Applying this methodology to Mesoscale Convective Systems (MCS)

Ultimately find a method to look for teleconnections



Evolving HPC Environments



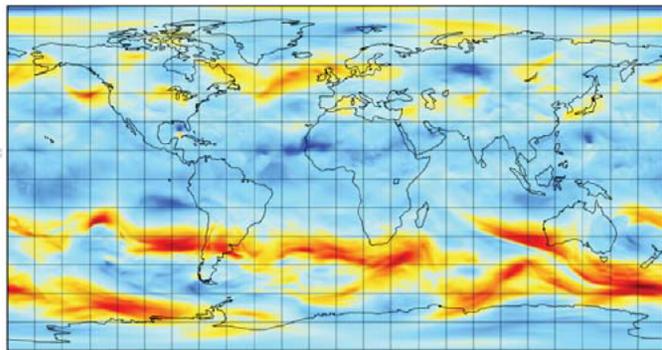
Machine Learning Requirements

- Both tightly coupled and loosely coupled applications
- Large amounts of data needed (high bandwidth)
- Potential to create large amounts of data prior to final analysis

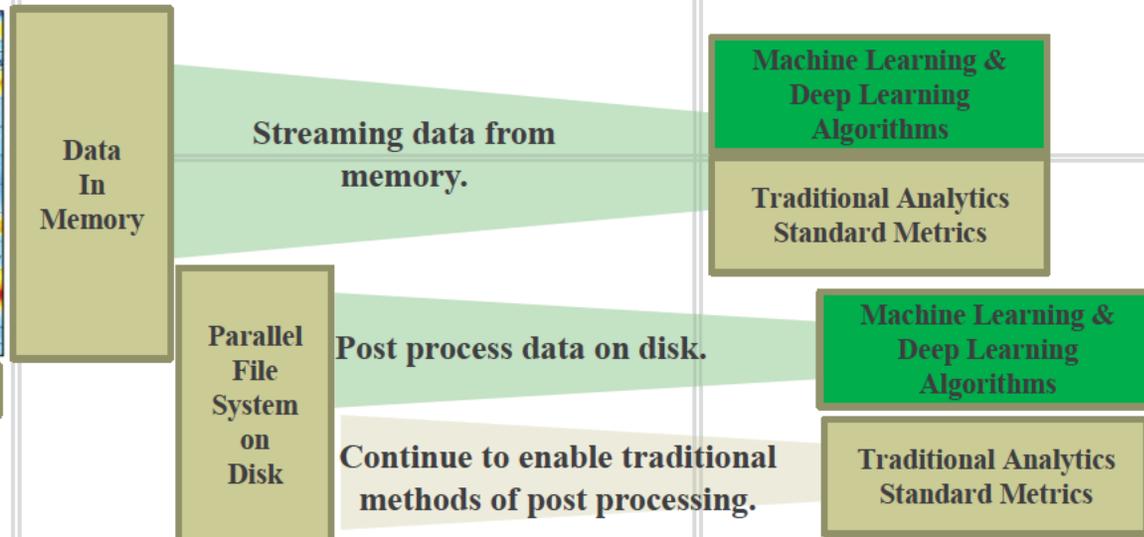
What does this mean for HPC Environments?

- Adaptive environments are needed to accommodate emerging software environments
- Compute and data need to be co-located with software stacks that enable emerging machine learning and deep learning techniques
- Systems need to be heterogeneous (beyond just the processors): combination of systems with different compute, memory, and storage capabilities are needed

Future of Machine Learning, Data Analytics and HPC

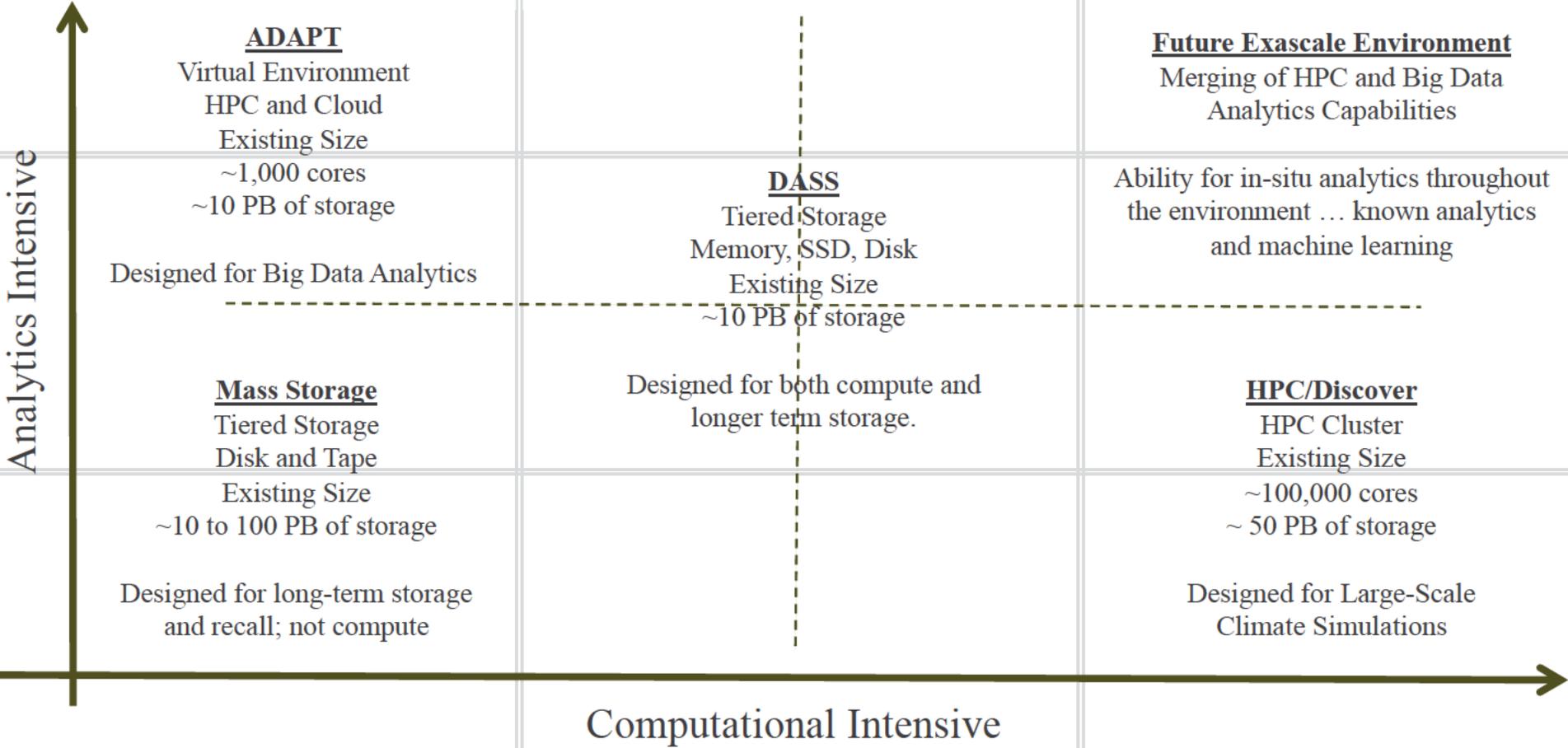


Climate/Weather Models (HPC)

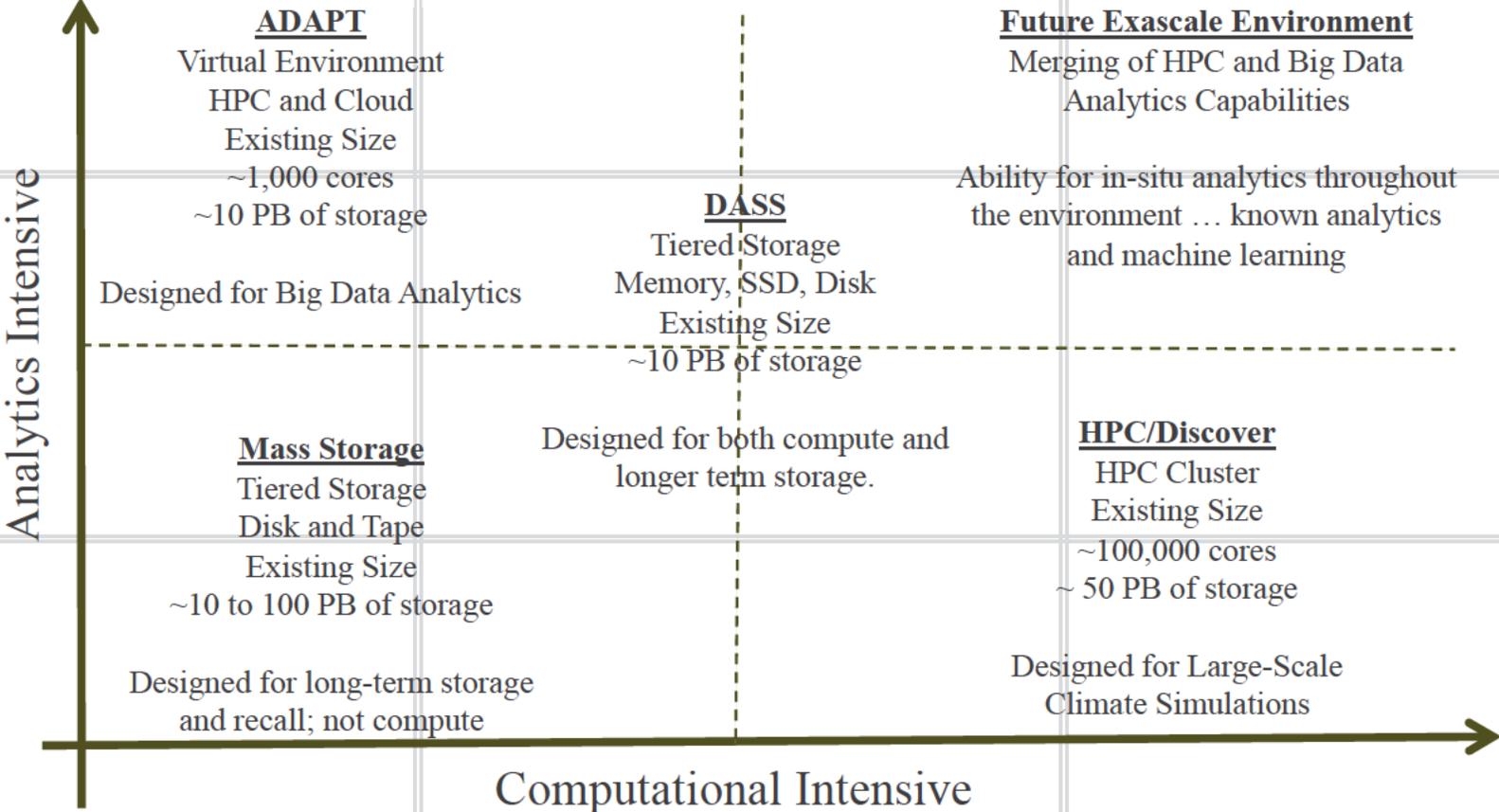


- Future HPC systems must be able to efficiently transform information into knowledge using both traditional analytics and emerging *machine learning* techniques.
- Requires the ability to be able to index data in memory and/or on disk and enable analytics to be performed on the data where it resides – even in memory
- All without having to modify the data

Looking toward Exascale



Looking toward Exascale

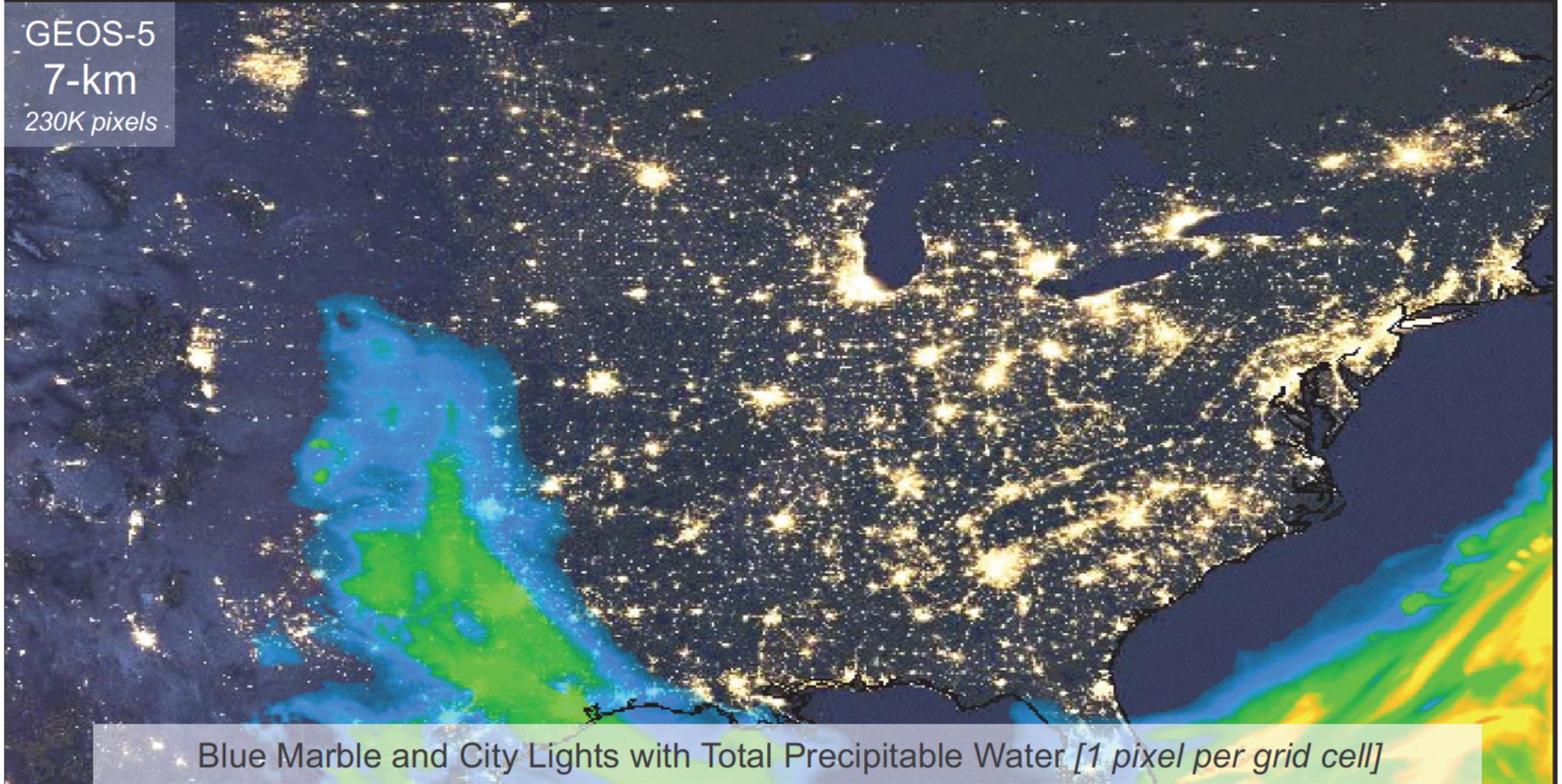


MERRA
50-km
4K pixels

Blue Marble and City Lights with Total Precipitable Water [1 pixel per grid cell]

Courtesy of Bill Putman, GSFC GMAO.

GEOS-5
7-km
230K pixels



Courtesy of Bill Putman, GSFC GMAO.

GEOS-5
3.5-km
920K pixels



Blue Marble and City Lights with Total Precipitable Water [1 pixel per grid cell]

Courtesy of Bill Putman, GSFC GMAO.

6 KM GEOS-5 Outgoing Longwave Radiation (OLR) (Global Modeling and Assimilation Office)

Outgoing Longwave Radiation [W m⁻²]

