CESM Usage Metrics and Machine Learning

Lolita Mannik

Regis University

National Center for Atmospheric Research (NCAR)

Summer Internships in Parallel Computational Science (SIParCS)

December 18, 2019

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Goal

Demonstrate what we can do with CESM performance metadata

- ❑ Track versions over time
- ❑ Track performance over time
- ❑ Predict performance

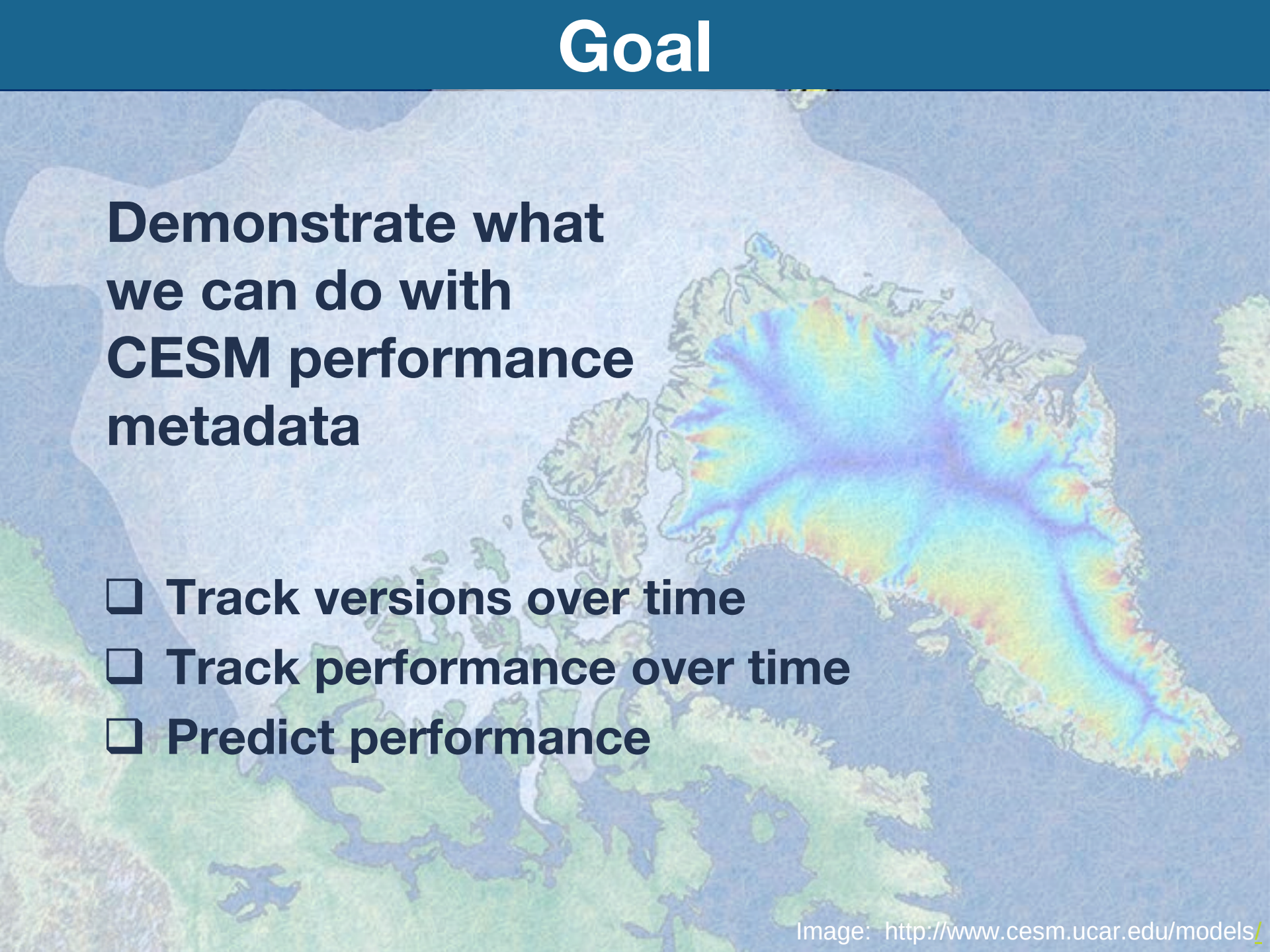


Image: http://www.cesm.ucar.edu/models/

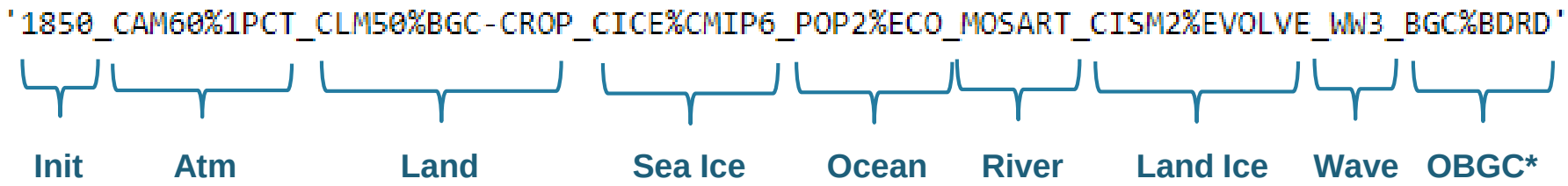
Method


```
---------------- TIMING PROFILE ---------------------
  Case        : b.e21.BHIST.f09_g17.CMIP6-historical.001
  LID         : 2979765.chadmin1.181015-050236
  Machine     : cheyenne
  Caseroot    : /gpfs/fs1/work/cmip6/cases/b.e21.BHIST.f09_g17.CMIP6-historical.001
  Timeroot    : /gpfs/fs1/work/cmip6/cases/b.e21.BHIST.f09_g17.CMIP6-historical.001/Tools
  User        : cmip6
  Curr Date   : Mon Oct 15 10:01:22 2018
  grid        : a%0.9x1.25_l%0.9x1.25_oi%gx1v7_r%r05_g%gland4_w%ww3a_m%gx1v7
  compset     : HIST_CAM60_CLM50%BGC-CROP_CICE_POP2%ECO%ABIO-DIC_MOSART_CISM2%NOEVOLVE_WW3_BGC%BDRD
  run_type    : hybrid, continue_run = TRUE (inittype = FALSE)
  stop_option : nyears, stop_n = 5
  run_length  : 1825 days (1825.0 for ocean)

  component       comp_pes    root_pe   tasks  x threads instances (stride)
  ---------        ------     -------   ------   ------  ---------  ------
  cpl = cpl        3456        0        1152   x 3       1      (1     )
  atm = cam        3456        0        1152   x 3       1      (1     )
  lnd = clm        2592        0        864    x 3       1      (1     )
  ice = cice       864         864      288    x 3       1      (1     )
  ocn = pop        768         1152     256    x 3       1      (1     )
```

Data Prep: Parsing

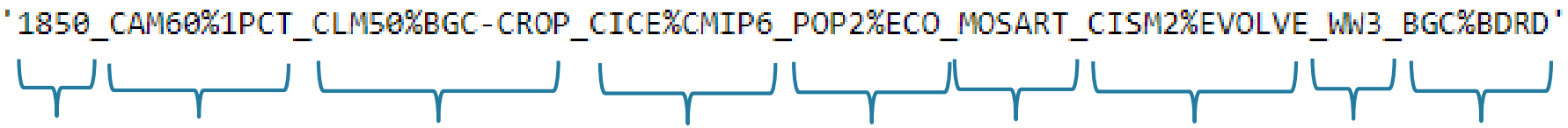
Component string = compset

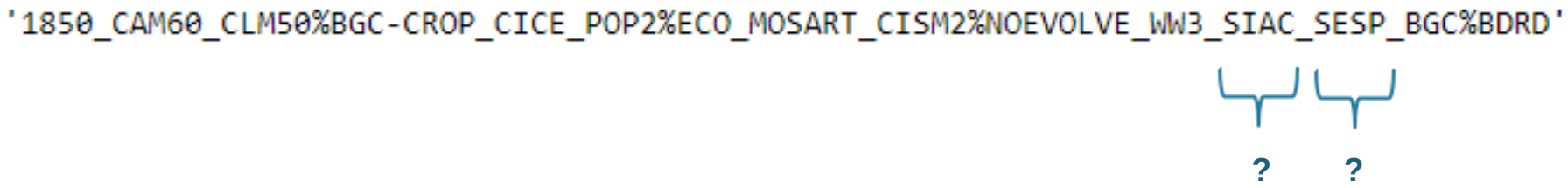


OBGC = Ocean Bio-geo-chemistry

Data Prep: Parsing

Component string = compset



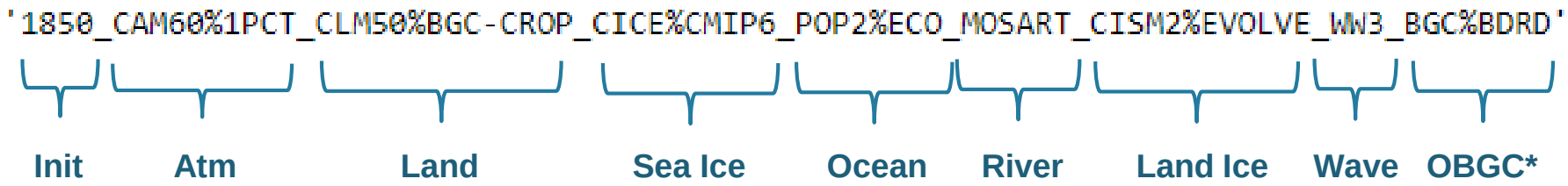


Problems: Manual inspection
Components not in the same order

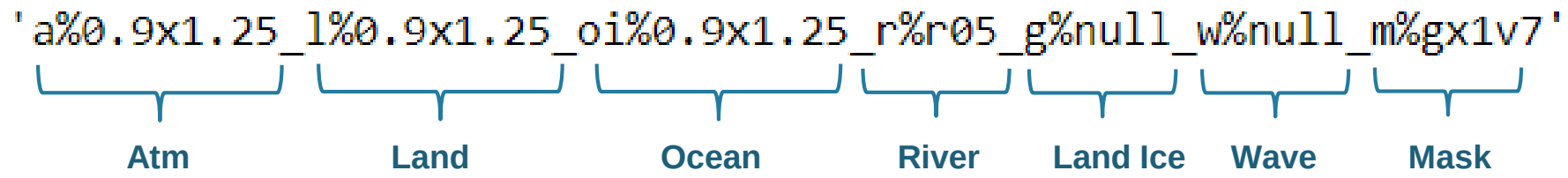
OBGC = Ocean Bio-geo-chemistry

Data Prep: Parsing

Component string = compset

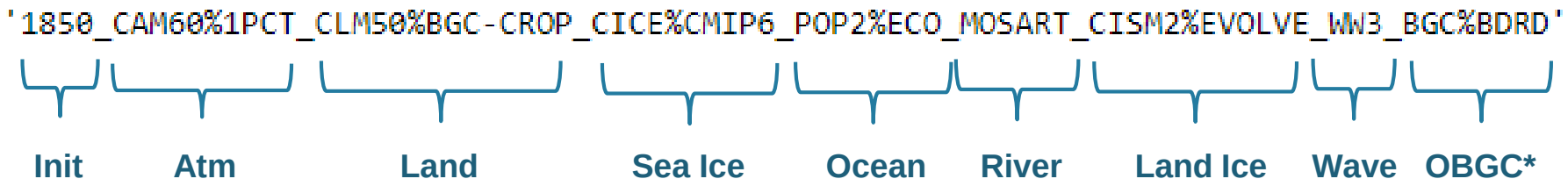


Grid string has prefixes

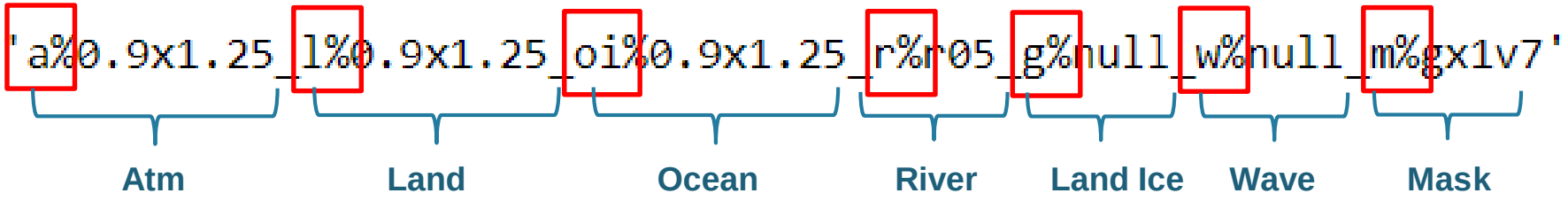


Data Prep: Parsing

Component string = compset



Grid string has prefixes



Data Prep: Parsing

Component string = compset

'1850_CAM60%1PCT_CLM50%BGC-CROP_CICE%CMIP6_POP2%ECO_MOSART_CISM2%EVOLVE_WW3_BGC%BDRD'

Init | Atm | Land | Sea Ice | Ocean | River | Land Ice | Wave | OBGC*

Grid string has prefixes

'a%0.9x1.25_l%0.9x1.25_oi%0.9x1.25_r%r05_g%null_w%null_m%gx1v7'

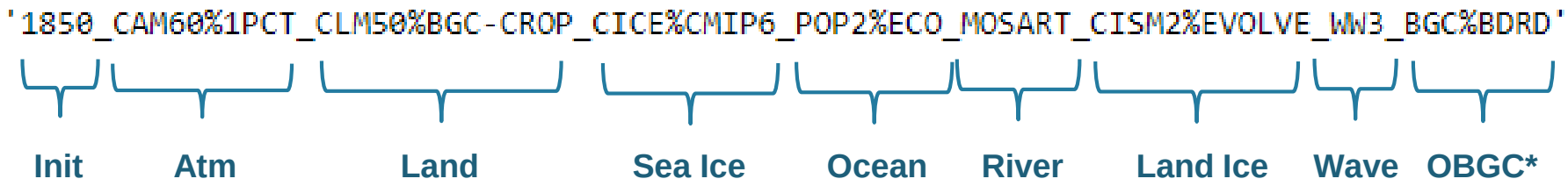
Atm | Land | Ocean | River | Land Ice | Wave | Mask

Random location:

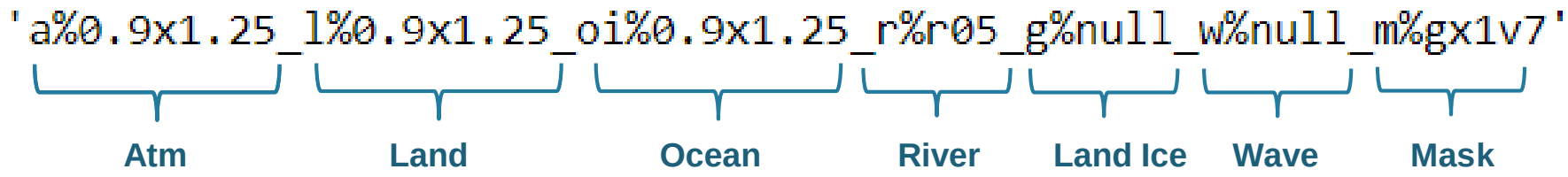
'a%1x1_vancouverCAN_l%1x1_vancouverCAN_oi%null_r%null_g%null_w%null_m%null'

Data Prep: Parsing

Component string = compset



Grid string has prefixes



Random location:

'a%1x1_vancouverCAN_l%1x1_vancouverCAN_oi%null_r%null_g%null_w%null_m%null'

Data Prep: Parsing

Component string = compset

'1850_CAM60%1PCT_CLM50%BGC-CROP_CICE%CMIP6_POP2%ECO_MOSART_CISM2%EVOLVE_WW3_BGC%BDRD'

Init | Atm | Land | Sea Ice | Ocean | River | Land Ice | Wave | OBGC*

Grid string has prefixes

'a%0.9x1.25_l%0.9x1.25_oi%0.9x1.25_r%r05_g%null_w%null_m%gx1v7'

Atm | Land | Ocean | River | Land Ice | Wave | Mask

Random location:

'a%1x1_vancouverCAN_l%1x1_vancouverCAN_oi%null_r%null_g%null_w%null_m%null'

'a%ne0np4colorado.ne30x16'

Analysis: CMIP Totals

416 Days

948
Unique Cases

21,785
Simulated Years

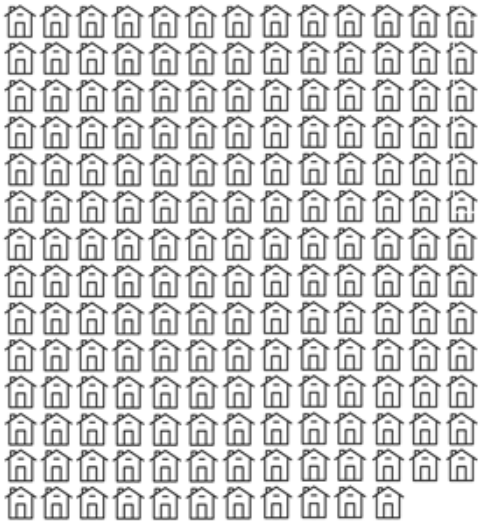
137,112,802
CPU Hours

Power Equivalence

137,112,802
CPU Hours

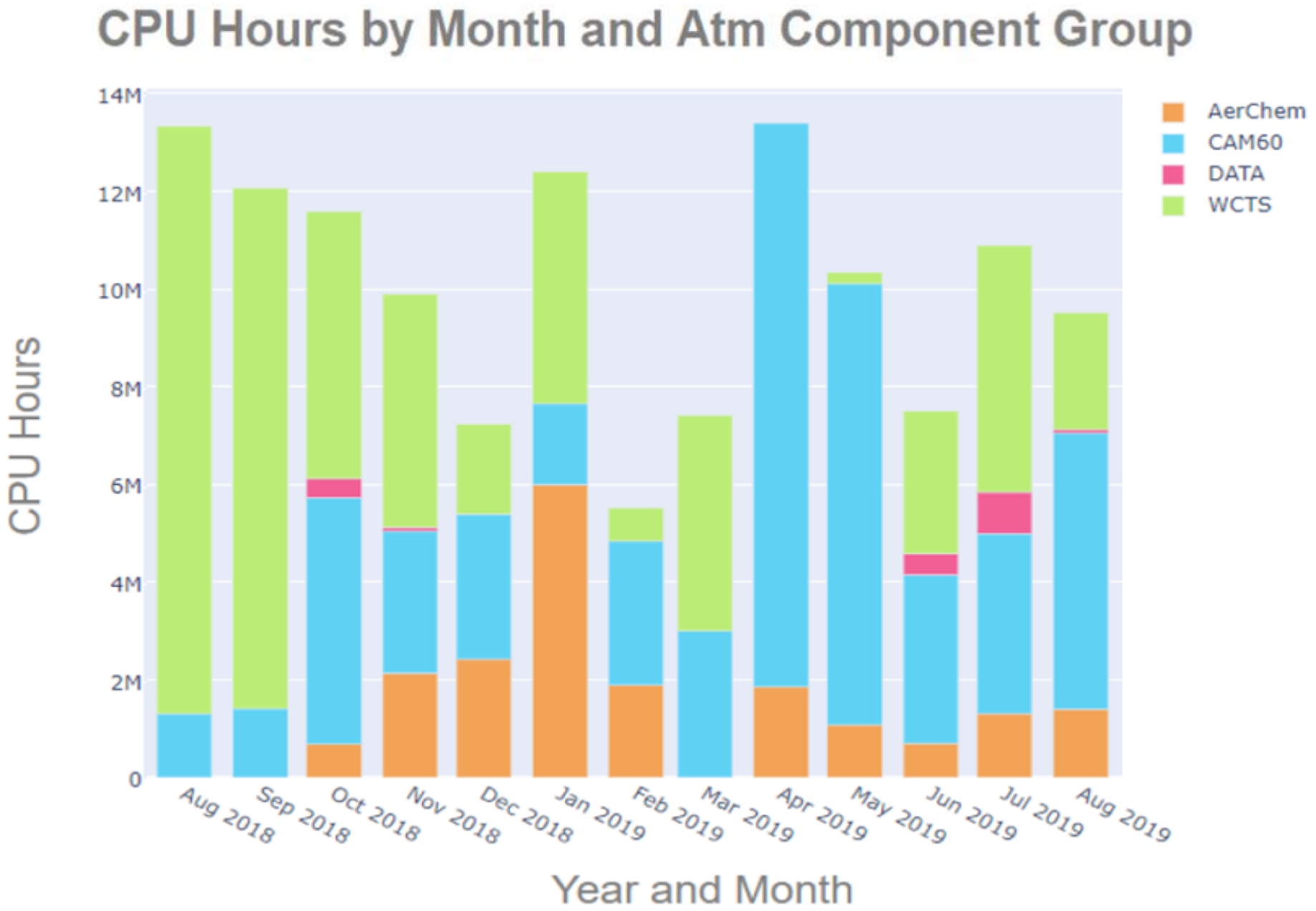
218 trips around the equator in a Nissan Leaf

or



Annual power for 180 Colorado homes

Analysis: Monthly Totals - CMIP



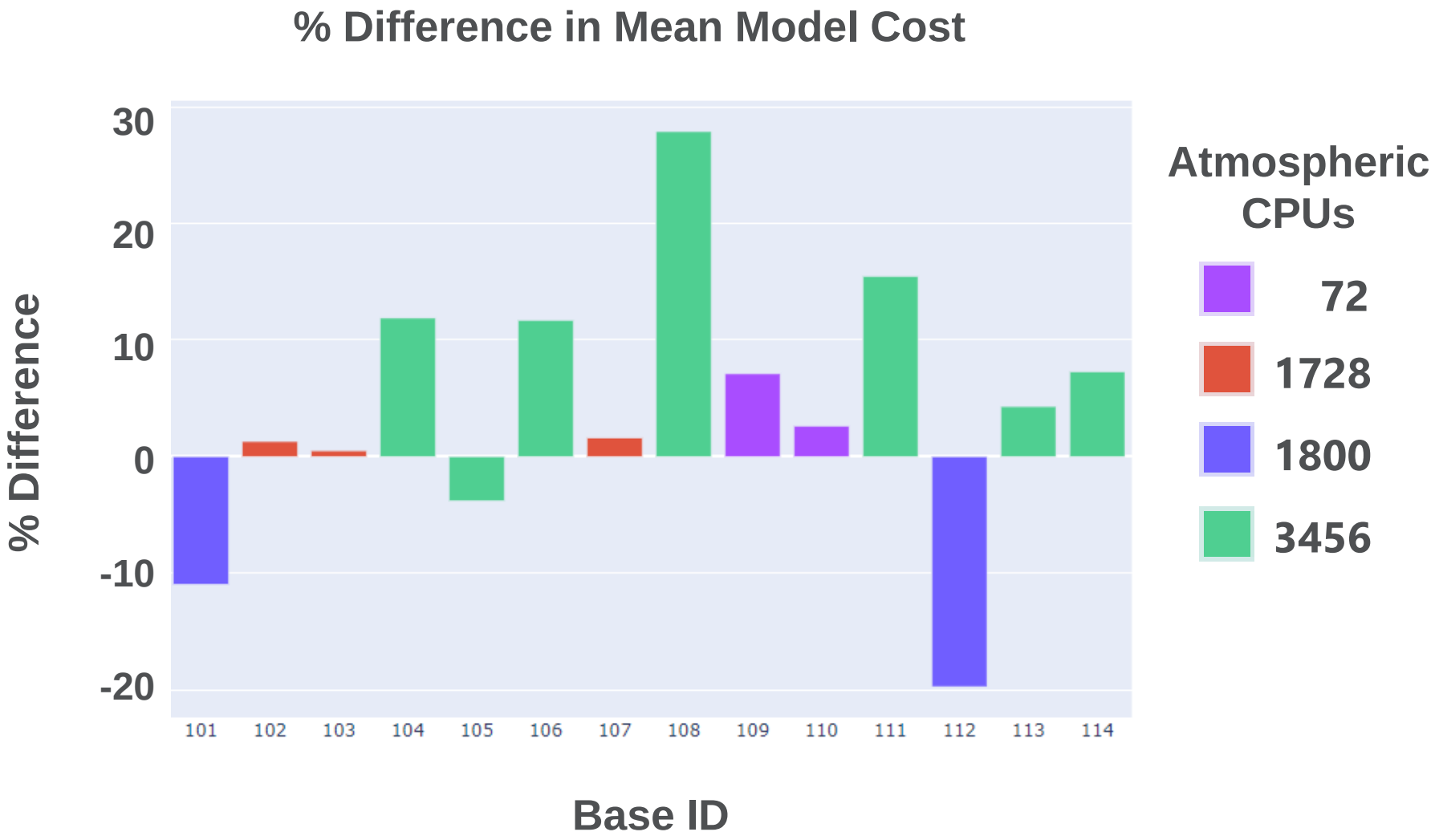
Analysis: System Upgrade

- **Cheyenne Supercomputer: 145,152 processors**
- **Upgrade: June 25-July 5, 2019**
- **Install SUSE Linux Enterprise Server Service Pack 4 to update security and support**

Subset by ensemble (like cases)
(1206 data points, 4271 sim years, 14 bases)

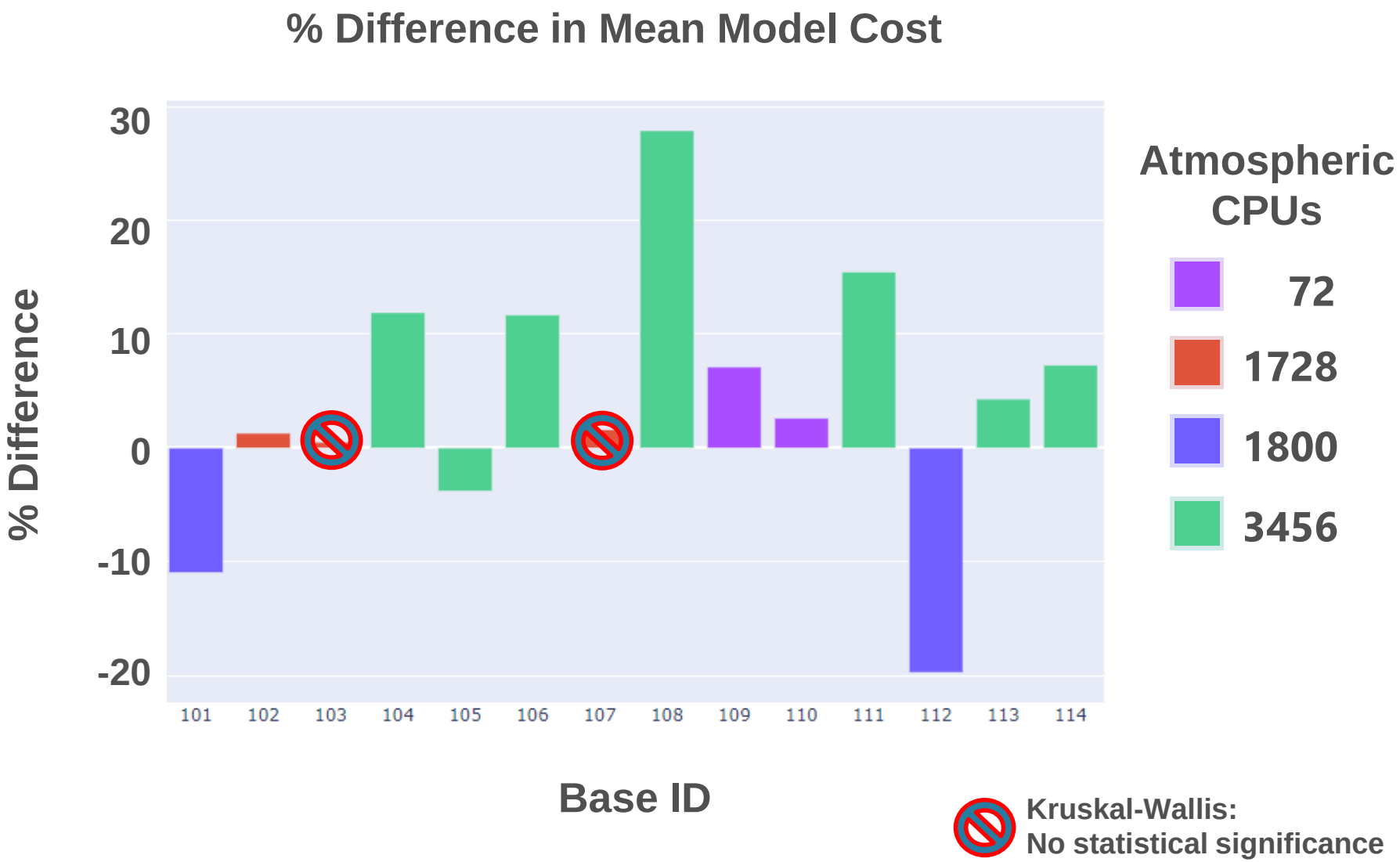
Analysis: System Upgrade

Ensembles that span the upgrade



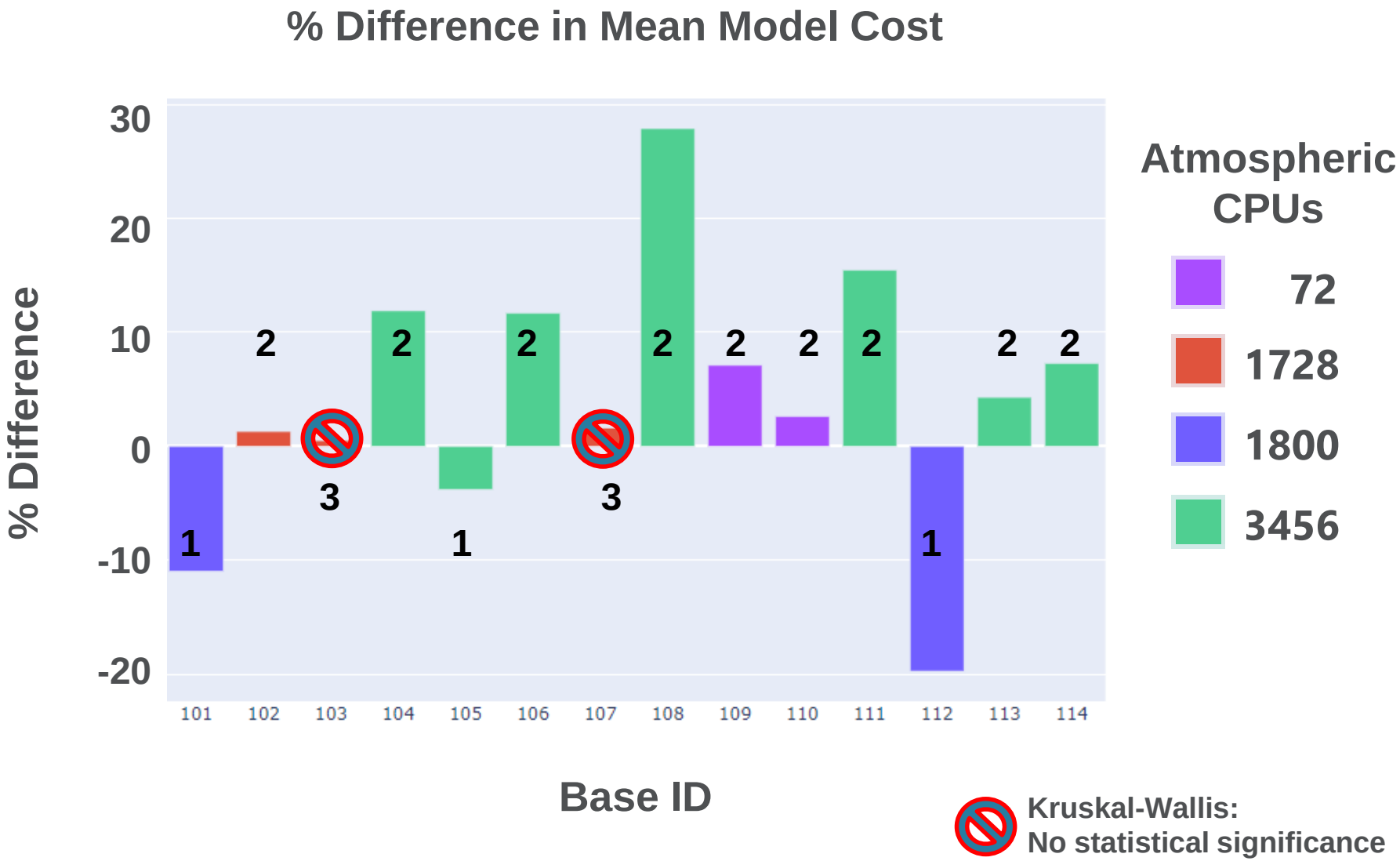
Analysis: System Upgrade

Ensembles that span the upgrade



Analysis: System Upgrade

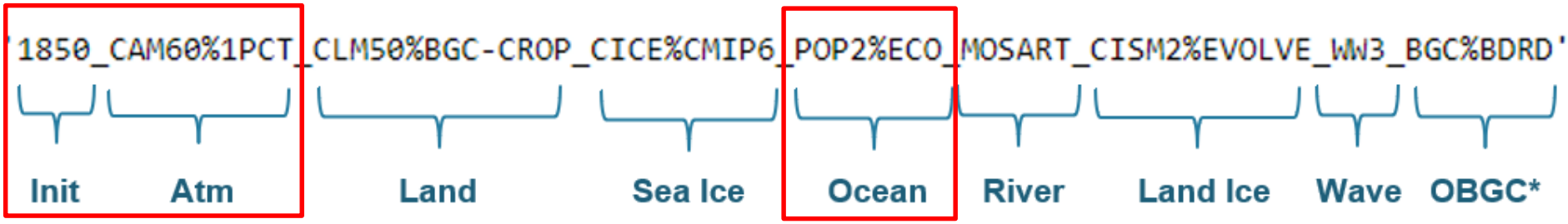
Ensembles that span the upgrade



Analysis: System Upgrade

Machine Learning

Logistic Regression
Random Forest

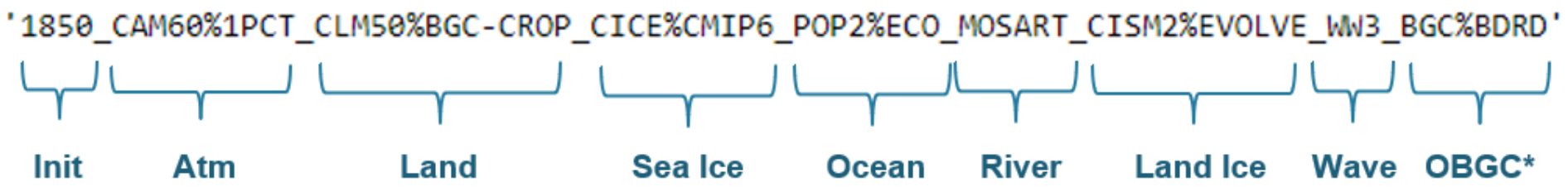


compset_init + compset_atm + compset_ocn + comp_pes_atm + RandNum ~ Performance (1, 2, or 3)

Analysis: System Upgrade

Machine Learning

Logistic Regression
Random Forest

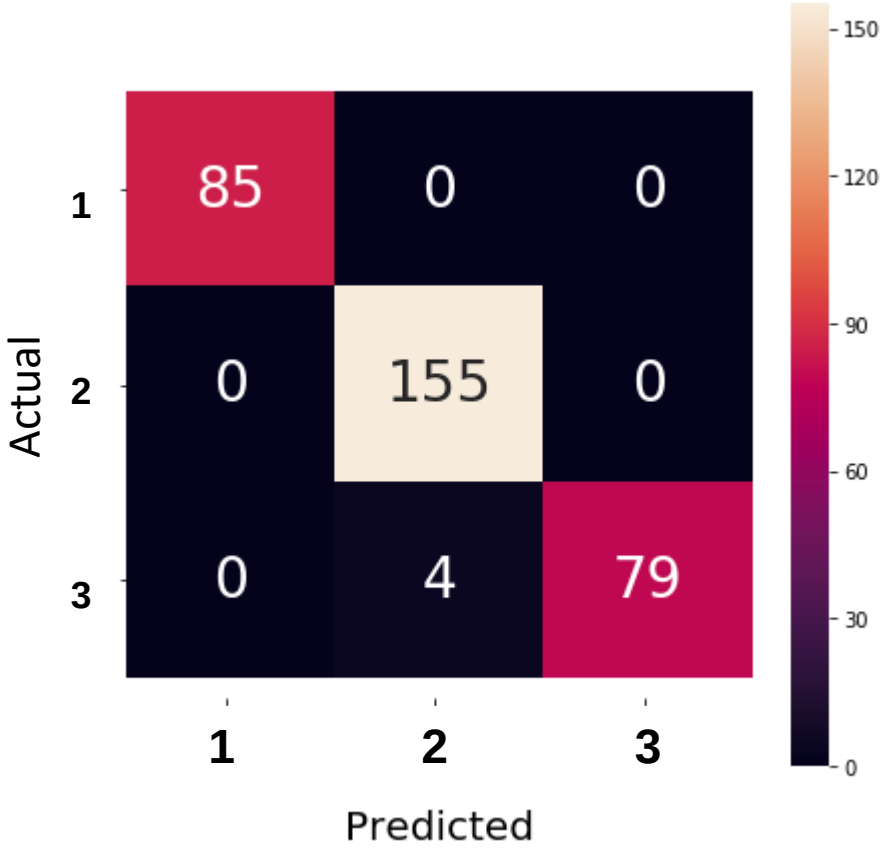


compset_init + compset_atm + compset_ocn
+ comp_pes_atm + RandNum ~ Performance (1, 2, or 3)

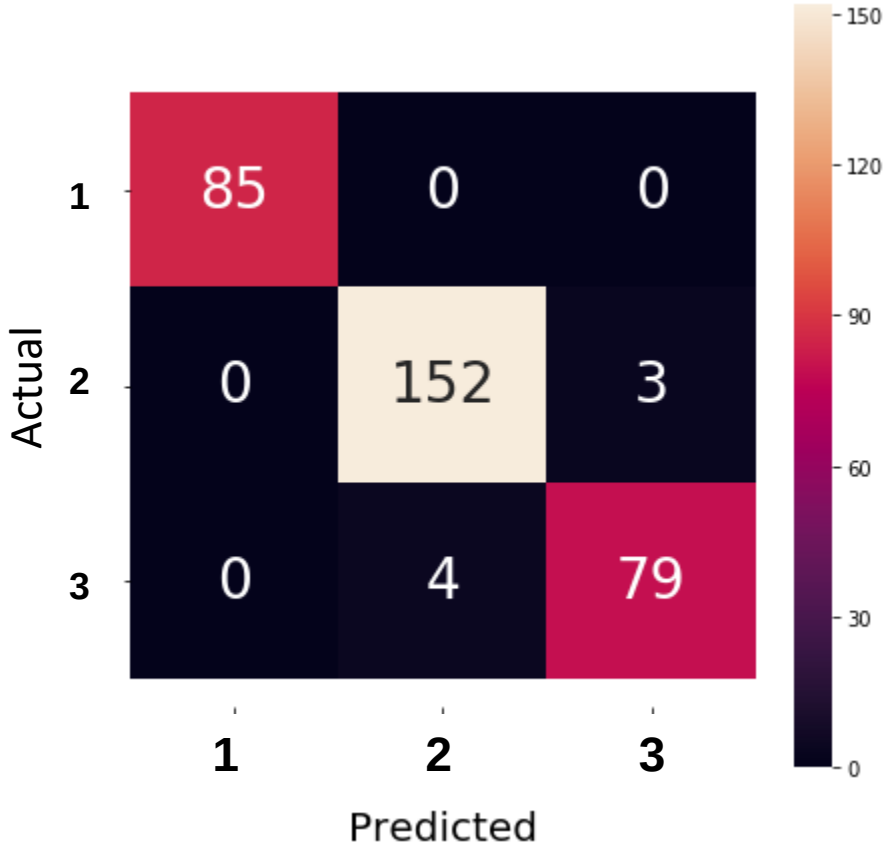
Analysis: System Upgrade

Machine Learning

Logistic Regression

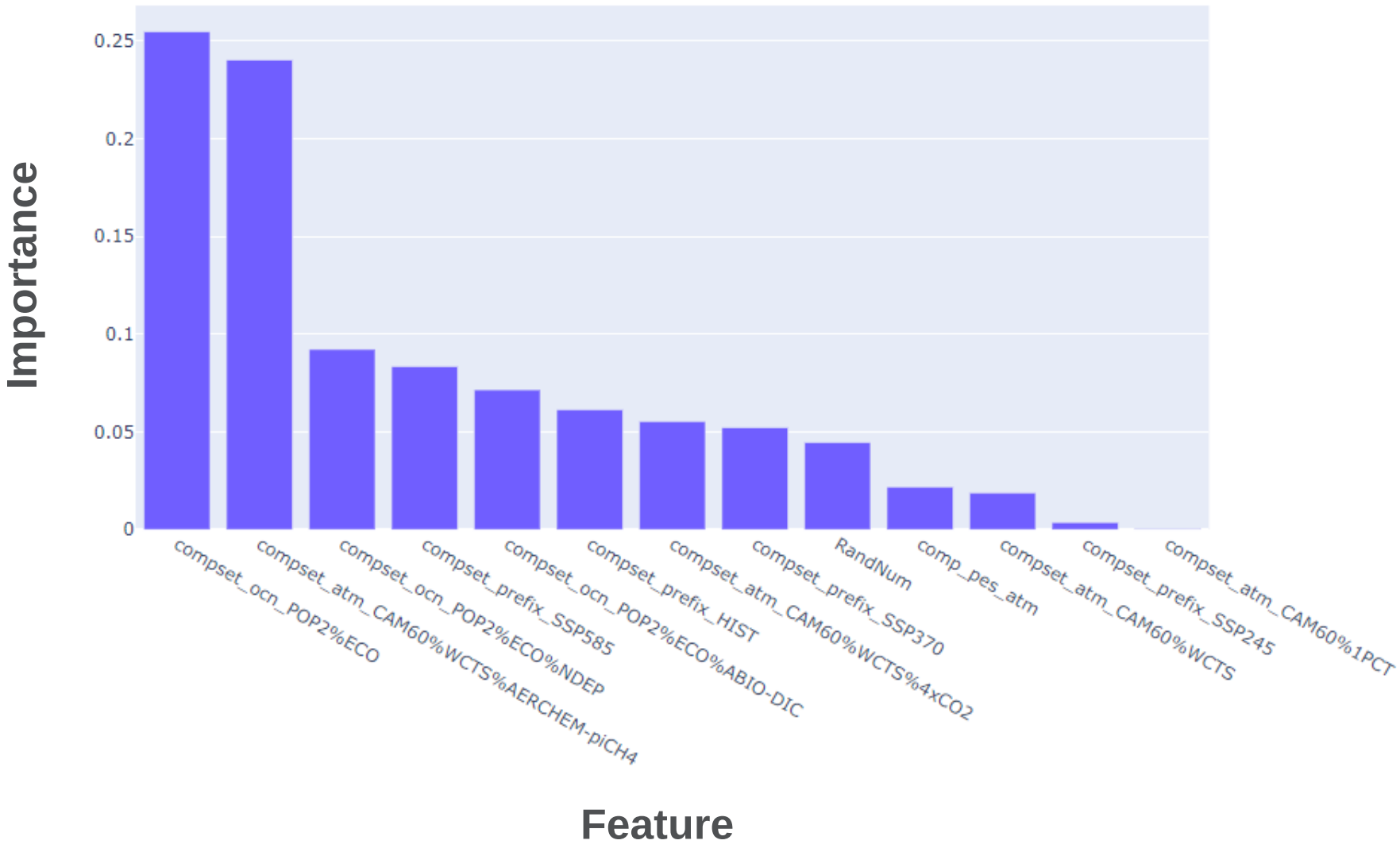


Random Forest



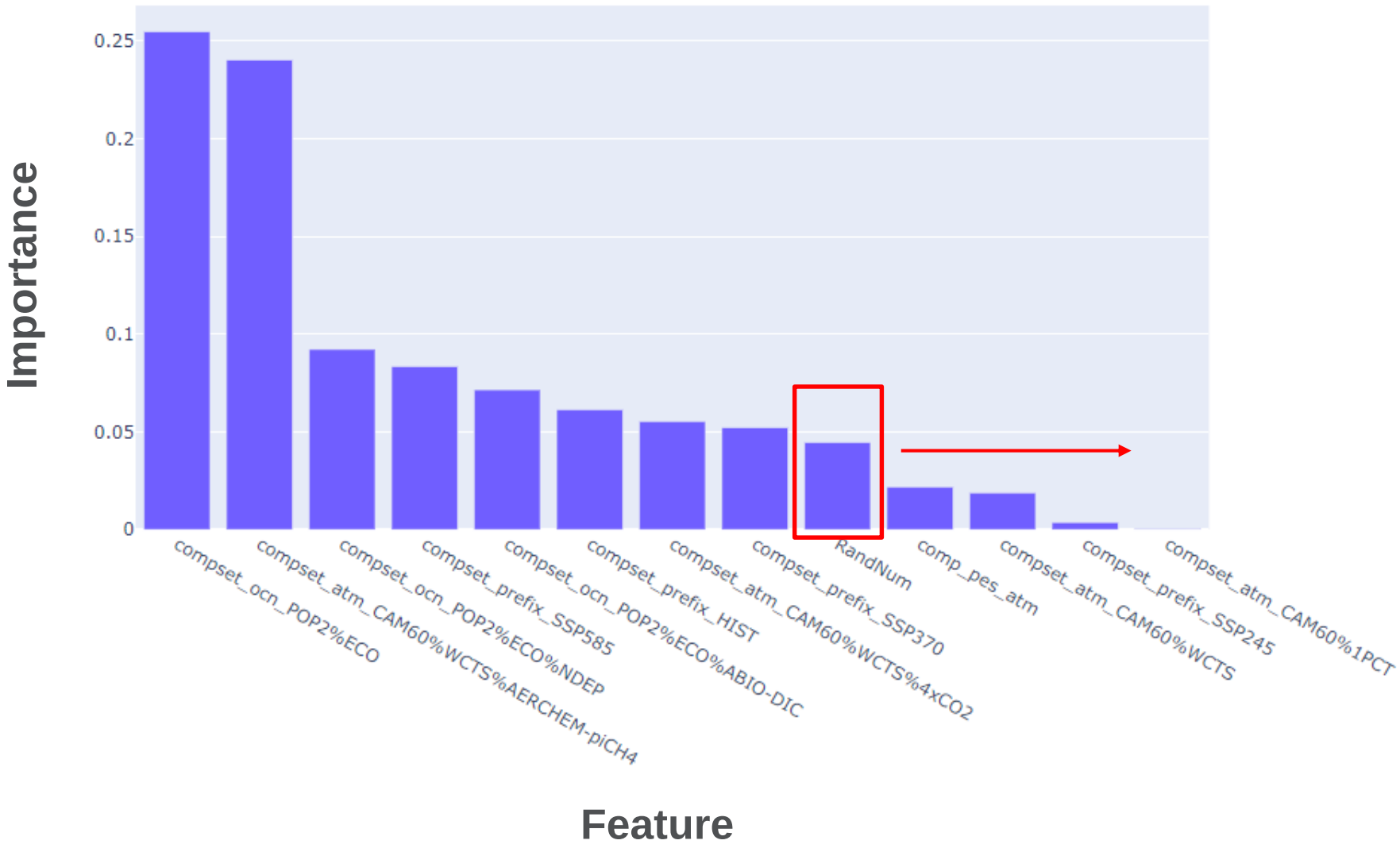
Analysis: System Upgrade

Feature Importance



Analysis: System Upgrade

Feature Importance



Analysis: System Upgrade

Final Report

	BaseNum		Change (%)	Prefix	ATM	OCN
Improved	101	b.e21.B1850G.f09_g17_gl4.CMIP6-piControl-withism	-10.94	1850	CAM60	POP2%ECO
	105	b.e21.BWSSP585cmip6.f09_g17.CMIP6-SSP5-8.5-WACCM	-3.8	SSP585	CAM60%WCTS	POP2%ECO%NDEP
	112	b.e21.B1850G.f09_g17_gl4.CMIP6-1pctCO2to4x-withism	-19.73	1850	CAM60%1PCT	POP2%ECO
Degraded	102	f.e21.FHIST_BGC.f09_f09_mg17.CMIP6-GMMIP	1.3	HIST	CAM60	DOCN%DOM
	104	b.e21.BWSSP370cmip6.f09_g17.CMIP6-SSP3-7.0-WACCM	11.86	SSP370	CAM60%WCTS	POP2%ECO%NDEP
	106	b.e21.BWCO2x4.f09_g17.CMIP6-G1-WACCM	11.7	1850	CAM60%WCTS%4XCO2	POP2%ECO%NDEP
	108	b.e21.B1850.f09_g17.CMIP6-DAMIP-hist-nat	27.87	1850	CAM60	POP2%ECO%ABIO_DIC
	111	b.e21.BSSP585_BPRPcmip6.f09_g17.CMIP6-esm-ssp585-ssp126-Lu	15.46	SSP585	CAM60	POP2%ECO%ABIO_DIC
	113	b.e21.BSSP245cmip6.f09_g17.CMIP6-SSP2-4.5	4.3	SSP245	CAM60	POP2%ECO%ABIO_DIC
	114	b.e21.B1850cmip6.f09_g17.DAMIP-hist-ghg	7.27	1850	CAM60	POP2%ECO%ABIO_DIC
Stayed the Same	103	f.e21.FWaerchem-piCH4.f09_g17.CMIP6-histSST-piCH4-WACCM	0.51	HIST	CAM60%WCTS%AERCHEM-piCH4	DOCN%DOM
	107	f.e21.F1850_BGC.f09_f09_mg17.CFMIP-piSST	1.59	1850	CAM60	DOCN%DOM

Analysis: Greedy CESM Data

9 years + 3 months

483,003 runs

38,062
Unique Cases

1,406,545
Simulated Years

1,054,615,678

CPU Hours

Analysis: Greedy CESM Data

9 years + 3 months

483,003 runs

38,062

Uniqu

Cases

6,5454

Simulated

Years

10 Parsers!

1,054,615,678

CPU Hours

Analysis: Greedy CESM Data

Predictive Modeling – Linear Regression

- Compset (parsed out)
- Grid (parsed out)
- Run type
- Simulated years

For each component:

- Instances
- Tasks
- Threads
- Root

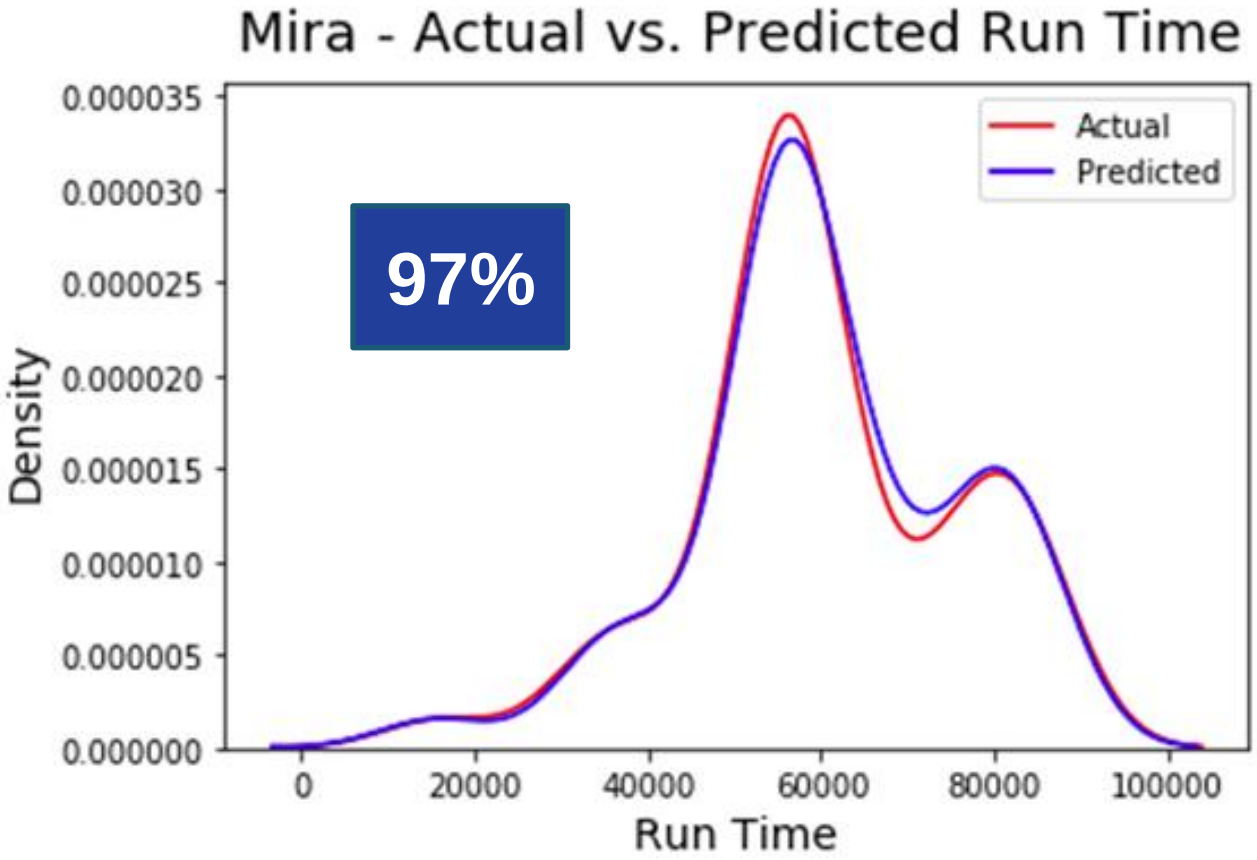
Can I predict total run time?

$$\text{Model cost} = \frac{\#\text{ of processors} \times \text{run time}}{\text{simulated years}}$$

Analysis: Greedy CESM Data

Predictive Modeling – Linear Regression

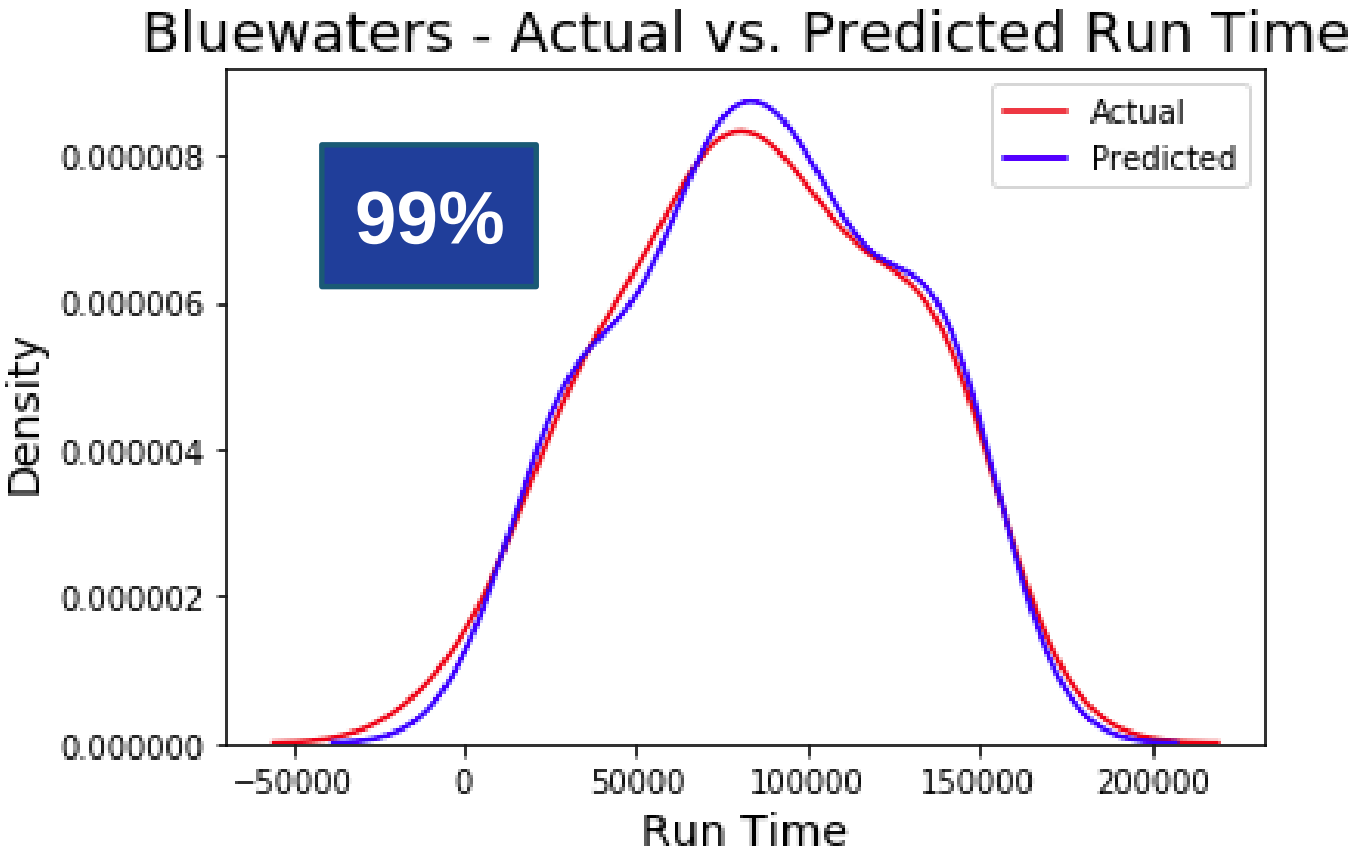
Mira (202 runs)



Analysis: Greedy CESM Data

Predictive Modeling – Linear Regression

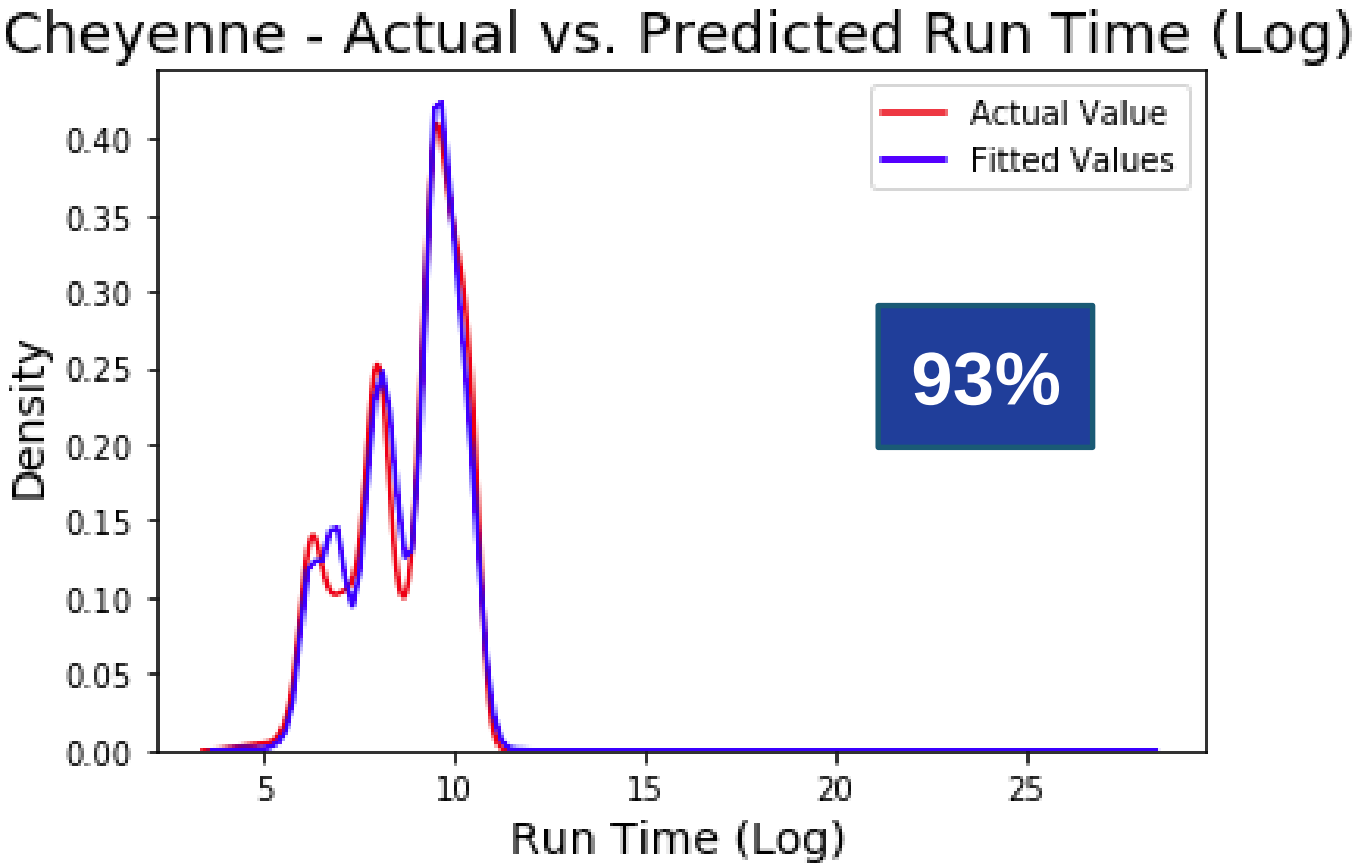
Bluewaters (305 runs)



Analysis: Greedy CESM Data

Predictive Modeling – Linear Regression

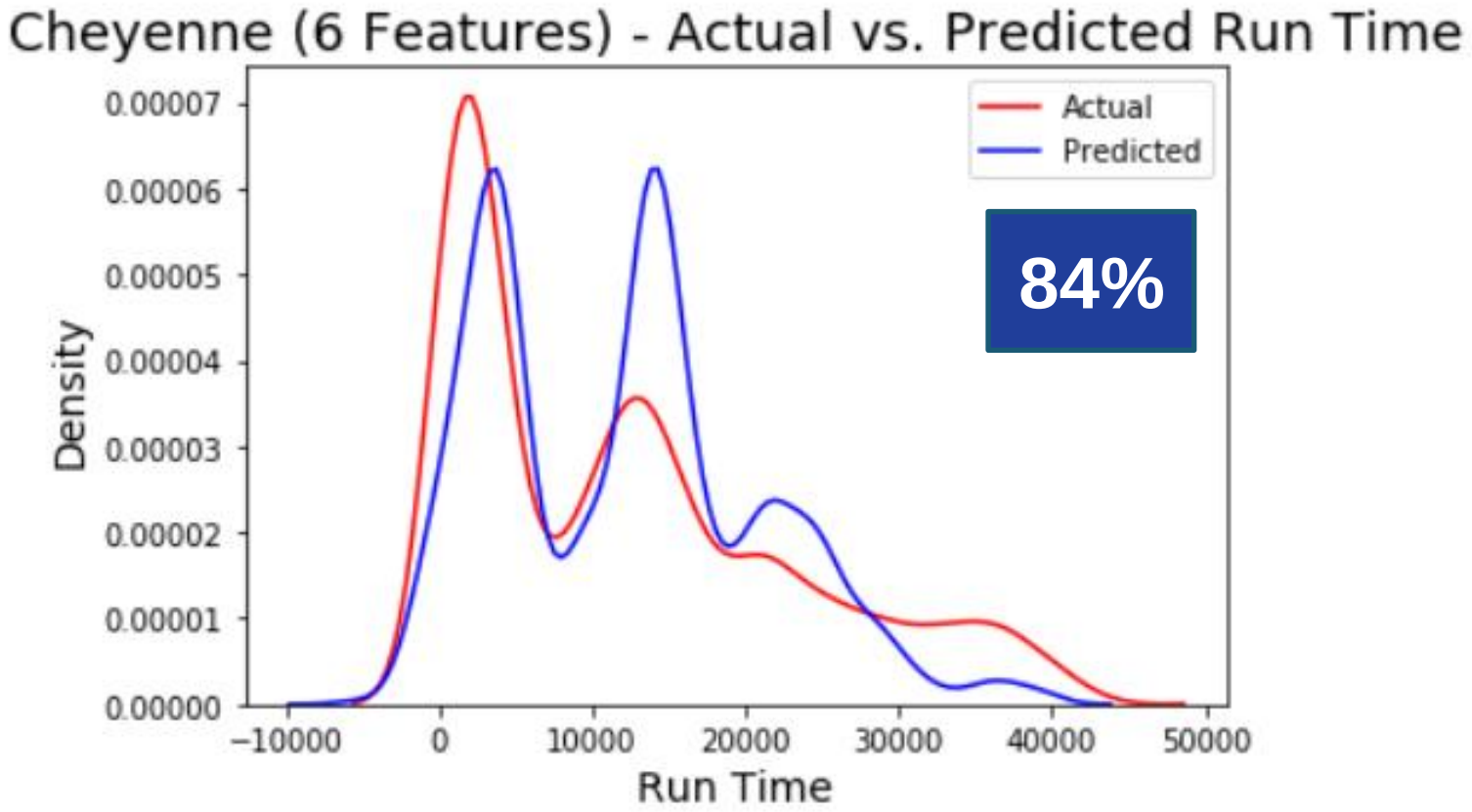
Cheyenne (48,313 runs)



Analysis: Greedy CESM Data

Predictive Modeling – Linear Regression

compset_init + compset_atm + compset_ocn +
grid_atm + grid_ocn + run_length_years ~ Run Time



Conclusion

Why do we care about predicting performance?

CPU hours are *expensive* and *limited*

If scientists can enter their configuration into a form and see the expected run time, they could:

- **Plan their computing allocation**
- **Eliminate the need for some performance test runs**

Conclusion

Why do we care about predicting performance?

CPU hours are *expensive* and *limited*

If scientists can enter their configuration into a form and see the expected run time, they could:

- **Plan their computing allocation**
- **Eliminate the need for some performance test runs**

Example: Cheyenne had 276,000 total runs;

103,000+ runs were less than 10 simulated days

= 4.4M CPU hours

Future Work

Ongoing analytics

- Model tuning on feature importance
- Track performance over time
- Track new version adoption rates

Automated tool that learns from performance data:

- Help inform scientist computing budgets
- Detect issues that reduce performance, such as misconfiguration, bad hardware, etc.

Acknowledgements

John Dennis
Brian Dobbins
NCAR mentors

Alice Bertini
NCAR SQL Training

AJ Lauer
Virginia Do
NCAR intern managers

Christy Pearson
Michael Busch
Nate George
Professors at Regis University

References

Balaji, et. al. CPMIP: Measurements of Real Computational Performance of Earth System Models in CMIP6. Geoscience Model Development Issue 10. January 02, 2017. https://www.geosci-model-dev.net/10/19/2017/

Images

Unless otherwise noted, graphics are from www.vecteezy.com

Questions?

Lolita Mannik – 4winds@mannik.com

This presentation, the data, and my Jupyter Notebooks will be posted at:
https://github.com/ChihuaWars/CESM-Analytics