Earth & Environmental Systems Modeling

State of Metrics and Benchmarking

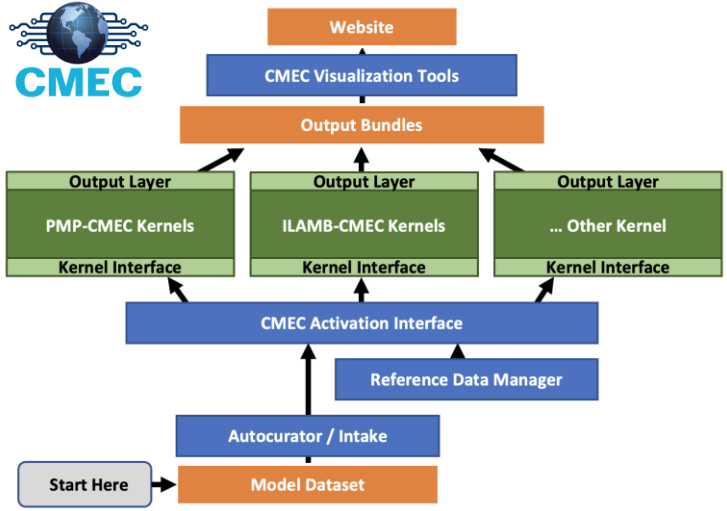
Paul Ullrich

U.S. DEPARTMENT OF ENERGY | Office of Science

2024 EESM PI Meeting
August 6–9, 2024

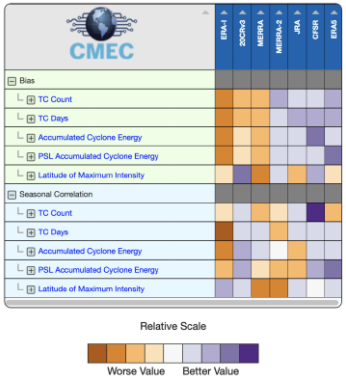
Coordinated Model Evaluation Capabilities (CMEC)

CMEC provides uniform standards and software tools that allow multiple metrics and diagnostics packages to be executed from a unified interface, and results explored in a unified manner.



Project Goals

1. Develop robust **standards** for the development of metrics and diagnostics packages.
2. Develop accompanying tools for **coordinated execution** of metric packages and **interactive analysis of** metrics and diagnostics package output.
3. Build **connections and standards across projects and agencies** related to model evaluation (e.g., MDTF).



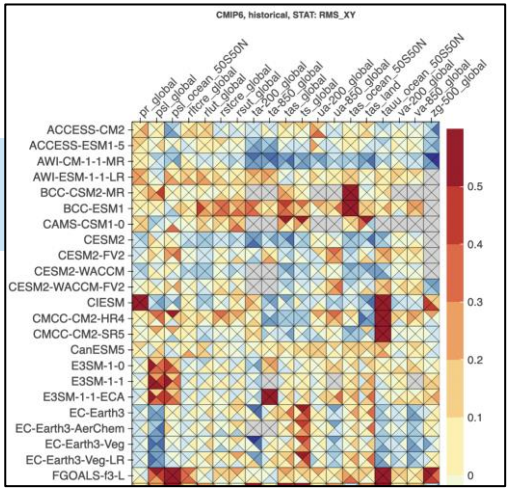
CMEC supports a comprehensive software suite for analysis of model evaluation output, including both climate data metrics and diagnostics.

https://cmec.llnl.gov/

Flagship RGMA Evaluation Packages

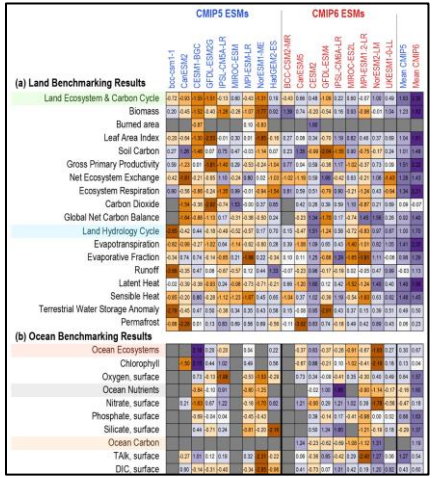
PCMDI Metrics Package

- Widely used for comprehensive evaluation of climate model systems by individual users, by national modeling centers, and within the international community.



International Land/Ocean Model Benchmarking

- Model–data intercomparison and integration packages for land and ocean hydrology and biogeochemistry. Designed to improve the performance of land and ocean models, ILAMB and IOMB also inform improved design of new measurement campaigns to reduce uncertainties associated with key land carbon cycle and ocean processes.



Standalone Evaluation Packages (Examples)

Coastal Storms (CyMEP)

An evaluation package for cyclonic storms (e.g., tropical cyclones) including basic statistics, spatial pattern evaluation, higher-order intensity metrics, and seasonality.

8° x 8°	$r_{xy,track}$	$r_{xy,gen}$	$r_{xy,u10}$	$r_{xy,slp}$	$r_{xy,ace}$	$r_{xy,pace}$
OBS	1.00	1.00	1.00	1.00	1.00	1.00
ERAI	0.89	0.86	0.39	0.23	0.83	0.92
20CRv3	0.93	0.90	0.79	0.83	0.90	0.97
MERRA	0.93	0.89	0.20	0.06	0.85	0.93
MERRA2	0.86	0.77	0.61	0.41	0.89	0.92
JRA	0.97	0.93	0.63	0.46	0.94	0.98
CFSR	0.97	0.93	0.72	0.32	0.93	0.97
ERA5	0.96	0.93	0.59	0.50	0.92	0.96

Worse Performance — Better Performance

Figure: Spatial skill scores of simulated TC track, genesis, and intensity in a set of reanalysis products against IBTrACS observations.

Drought Metrics

An evaluation package for model representation of drought, including probability of initiation/termination, areal coverage and seasonal character.

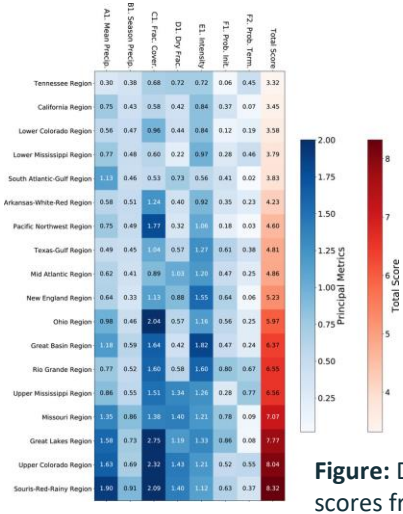


Figure: Drought skill scores from ERA5 across U.S. Regions.

Scales of Precipitation

An evaluation package to evaluate spatio-temporal connectivity of precipitation in models.

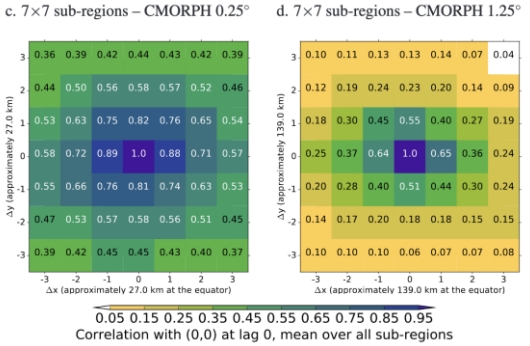
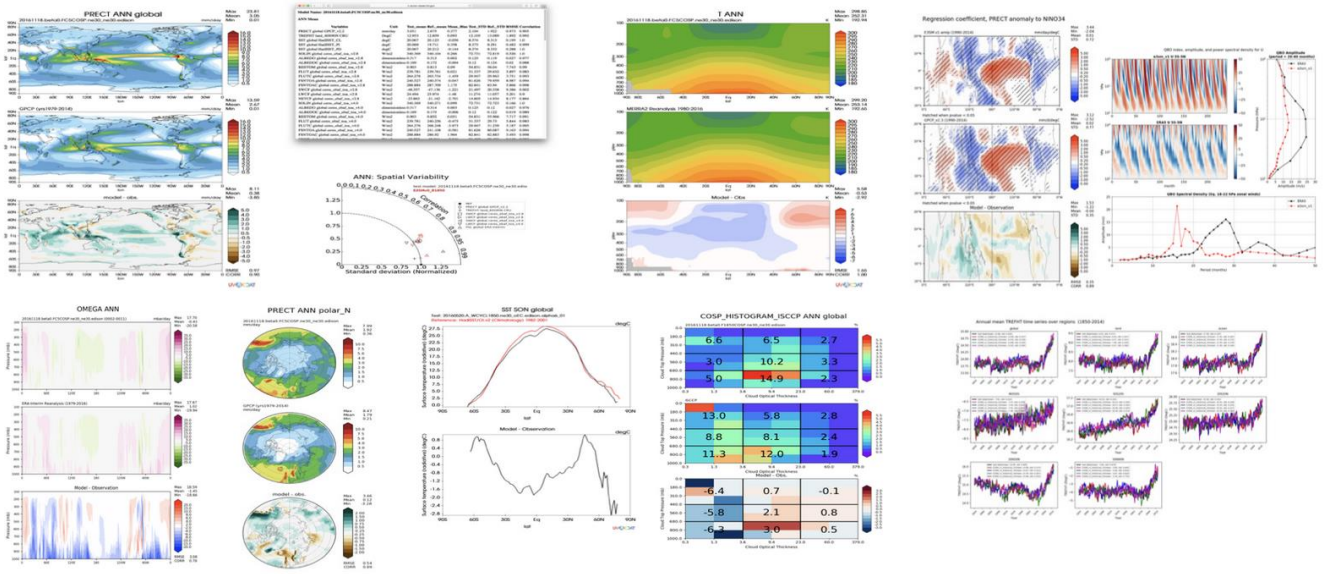


Figure: Correlations between precipitation intensity on a rectangular grid obtained from CMORPH at two scales.

E3SMDiags

https://github.com/E3SM-Project/e3sm_diags

Summary: A diagnostics and benchmarking package used within the E3SM project for comprehensive model evaluation and tuning.

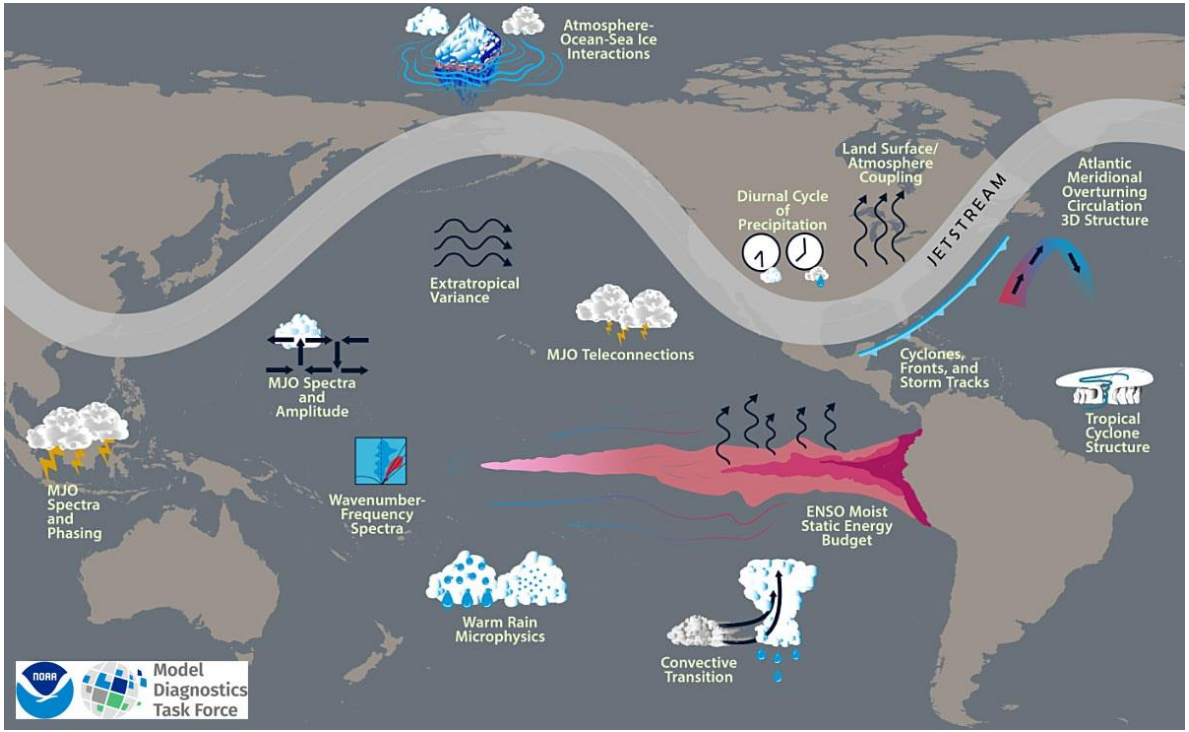


Collaborations

https://www.gfdl.noaa.gov/mdtf-diagnostics/

CMEC compatible

Summary: A NOAA-funded framework focused on model benchmarking via process-oriented diagnostics (PODs).

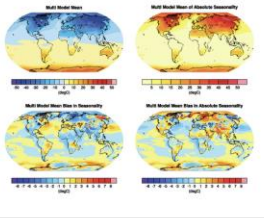


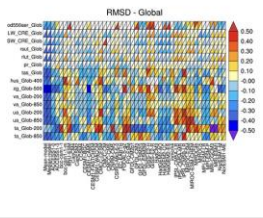
International Efforts

ESMValTool
Earth System Model Evaluation Tool

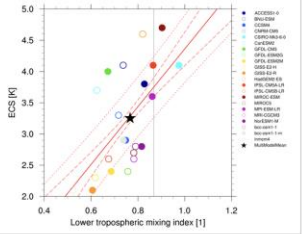
https://esmvaltool.org

Summary: A grab bag of different evaluation capabilities integrated from a variety of sources covering all aspects of the Earth system.

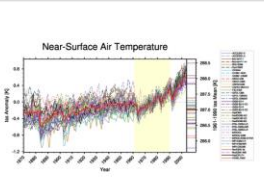

recipe_flato13ipcc.yml



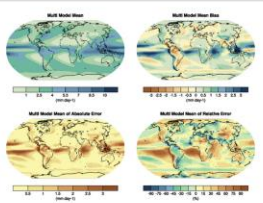
recipe_perfmetrics_CMIP5.yml

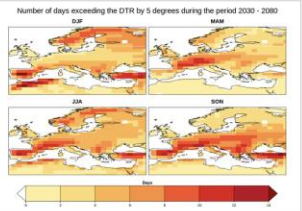


recipe_ecs_scatter.yml

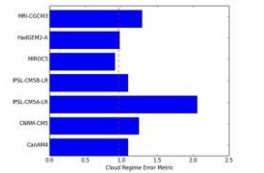


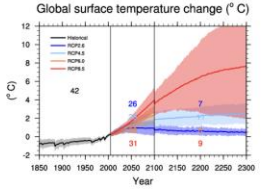
recipe_flato13ipcc.yml


recipe_flato13ipcc.yml

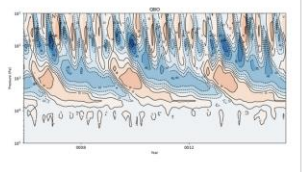


recipe_diurnal_index.yml


recipe_crem.yml



recipe_collins13ipcc.yml


recipe_autoassess_stratosphere.yml

DOE's contribution to CMIP7

Engagement with the CMIP Climate Model Benchmarking Task Team

Birgit Hassler
TT Co-Lead

Forrest Hoffman
TT Co-Lead

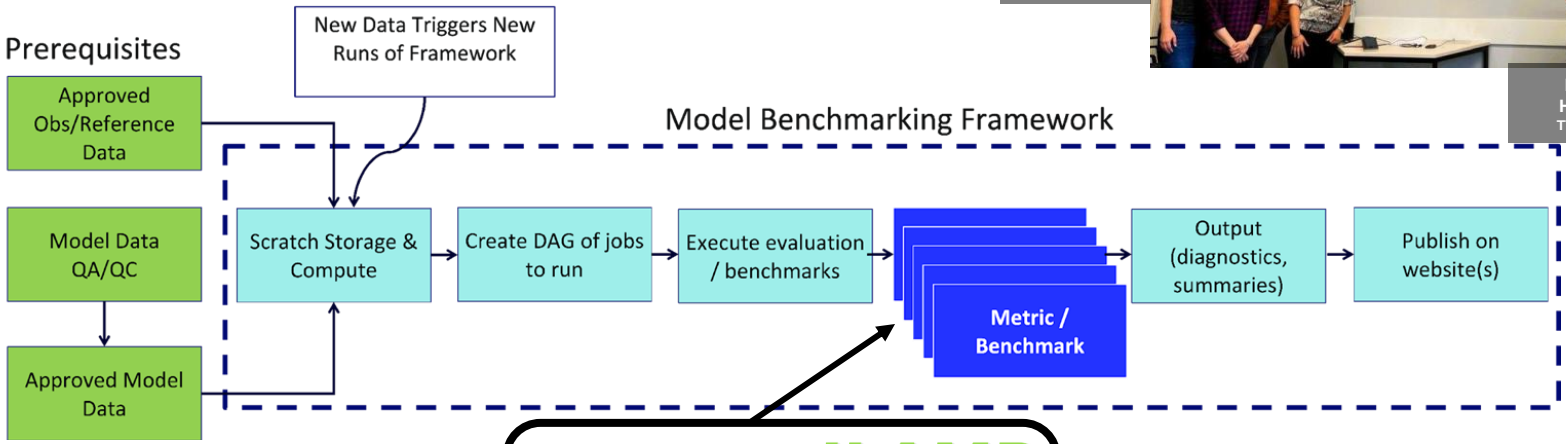
Jiwoo Lee

CMIP Model Benchmarking

[Task Team Meeting @Germany, DLR (2024 May)]

WCRP ESMO

Proposed **CMIP7 Rapid Evaluation Framework (REF)**



Earth System Model Evaluation and Benchmarking with the *PCMDI Metrics Package (PMP)*

Current Core Team members:

Jiwoo Lee, Ana Ordonez, Peter Gleckler, Paul Ullrich, Bo Dong, Kristin Chang (LLNL, PCMDI)

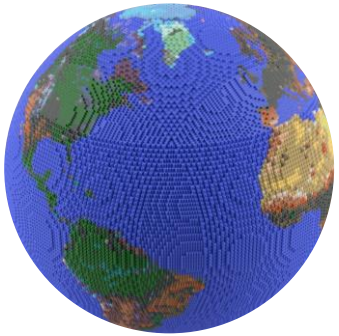
Along with contributors:

Mark Zelinka, Li-Wei Chao, Tom Vo, Paul Durack (LLNL), Min-Seop Ahn (NASA GSFC), Yann Planton (Monash U), Michael Wehner (LBNL), Daehyun Kim (SNU), Elina Valkonen (NASA GSFC), Julie Caron (NCAR), and many others!

@AOGS, 2024 June 25th

Systematic Evaluation for Diverse Models

There are 1000s simulations from 100s of diverse models in CMIPs!



Different resolution

Dynamics / Physics

Or, something unique!

How can we **objectively evaluate** and **efficiently document** their **performance**?

PCMDI Metrics Package (PMP)

Lee et al. 2024: Systematic and Objective Evaluation of Earth System Models: PCMDI Metrics Package (PMP) version 3. *Geoscientific Model Development*, 17, 3919–3948, doi: 10.5194/gmd-17-3919-2024

PMP is:

Open-source Python package for **objective evaluation and benchmarking** of physical climate as simulated by models

GitHub

PMP does :

Assess model **performance** using diverse metrics

Ensure **reproducibility** with detailed **provenance** and **version** control (codes, data, and operating conditions)

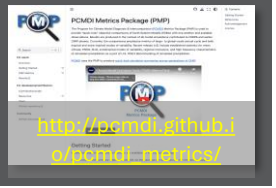
Link to reference datasets from **obs4MIPs** for more robust evaluation and reproducibility

obs4MIPs

PMP provides:

Reuseable software with documentation

Pre-calculated **database** of statistics and metrics for the CMIP archive



Impacts:

Quantitatively measure the performance evolution across **CMIP** generations

Provide objective goals for **model development**

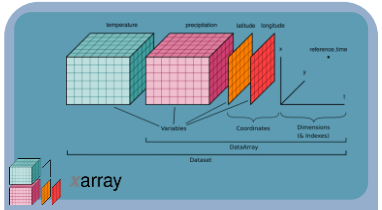
Downloaded > *33,000* times and used for evaluation of DOE and other agencies' models

E3SM Energy Exascale Earth System Model · NASA · NOAA · GFDL · National Institute of Meteorological Sciences · IPSL

PMP's primary building component: xCDAT

T. Vo, S. Po-Chedley, J. Boutte, J. Zhang, J. Lee (LLNL)

xCDAT encourages reusable code and reproducible science



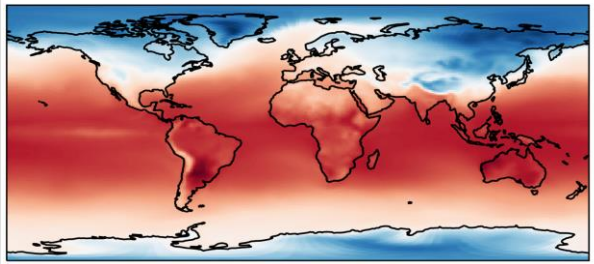
- Geospatial weighted averaging
- Temporal averaging, climatologies, departures
- Dataset bounds and CF metadata handling
- Horizontal and vertical regridding

Pure Xarray

```python
import numpy as np
import xarray as xr

# 1. Open the dataset.
dpath = (
    "/p/user_pub/work/CMIP6/CMIP/E3SM-Project/"
    "E3SM-2-0/historical/r1i1p1f1/Amon/ts/gr/v20220830/"
)
ds = xr.open_mfdataset(dpath + "*.nc")

# 2. Calculate monthly departures.
ts_mon = ds.ts.groupby("time.month")
ts_mon_clim = ts_mon.mean(dim="time")
ts_anom = ts_mon - ts_mon_clim

# 3. Compute global average.
coslat = np.cos(np.deg2rad(ds.lat))
ts_anom_wgt = ts_anom.weighted(coslat)
ts_anom_global = ts_anom_wgt.mean(dim="lat").mean(dim="lon")

# 4. Calculate annual averages.
# ncar.github.io/esds/posts/2021/yearly-averages-xarray/
mon_len = ts_anom_global.time.dt.days_in_month
mon_len_by_year = mon_len.groupby("time.year")
wgts = mon_len_by_year / mon_len_by_year.sum()

temp_sum = ts_anom_global * wgts
temp_sum = temp_sum.resample(time="AS").sum(dim="time")
denom_sum = (wgts).resample(time="AS").sum(dim="time")

ts_anom_global_ann = temp_sum / denom_sum

```

xCDAT

```python
import xcdat as xc

# 1. Open the dataset.
dpath = (
    "/p/user_pub/work/CMIP6/CMIP/E3SM-Project/"
    "E3SM-2-0/historical/r1i1p1f1/Amon/ts/gr/v20220830/"
)
ds = xc.open_mfdataset(dpath)

# 2. Calculate monthly departures.
ds_anom = ds.temporal.departures("ts", freq="month")

# 3. Compute global average.
ds_anom_global = ds_anom.spatial.average("ts")

# 4. Calculate annual averages
ds_anom_global_ann = ds_anom_global.temporal.group_average
("ts", freq="year")

```



More examples available at https://xcdat.readthedocs.io

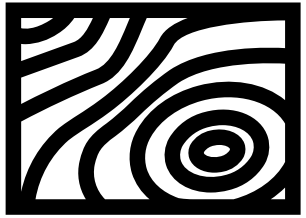
Vo et al. (2024) xCDAT: A Python package for simple climate data analysis on structured grids. Journal of Open Source Software. DOI: 10.21105/joss.06426

A comparison of code to calculate global-mean, monthly anomalies

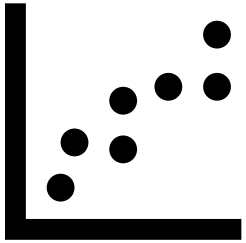
What Do We Evaluate?

Average

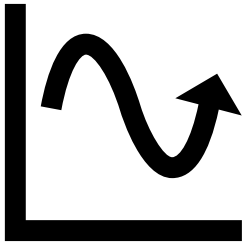
Pattern



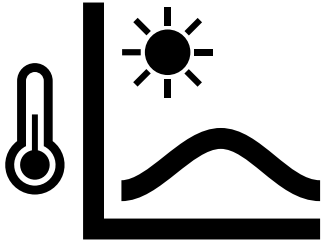
Physical Relationships



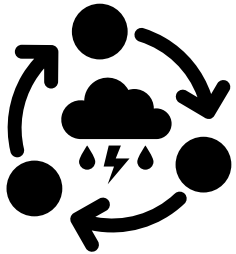
Variability



Extremes

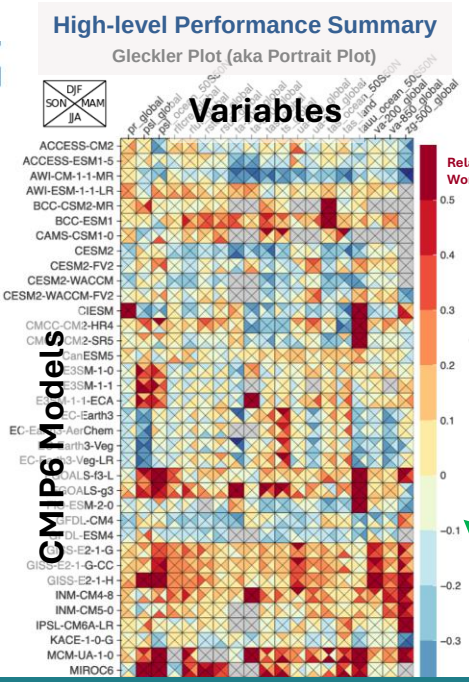


Processes



Evaluation needs to include diverse aspects of the simulated physical climate

Evaluation of **Climatology**



Sea-surface Temperature

Precipitation

Precipitable Water

Wind

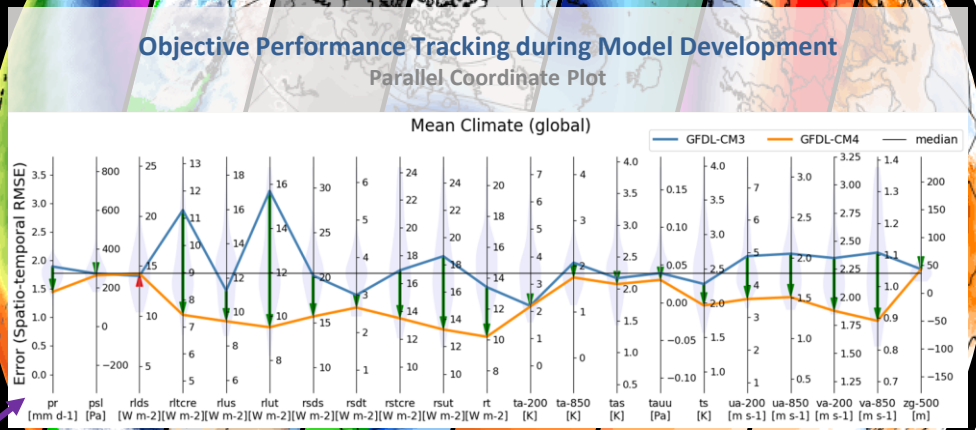
Geopotential Height

Sea Level Pressure

... and many others!

Relative Performance

Absolute error



Collective evaluation of multiple climate fields enables objective performance tracking

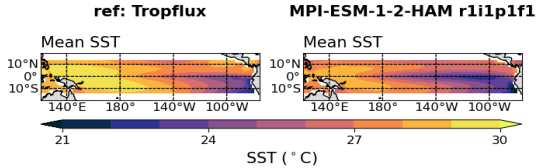
Field image : ClimateReanalyzer.org

Lawrence Livermore National Laboratory

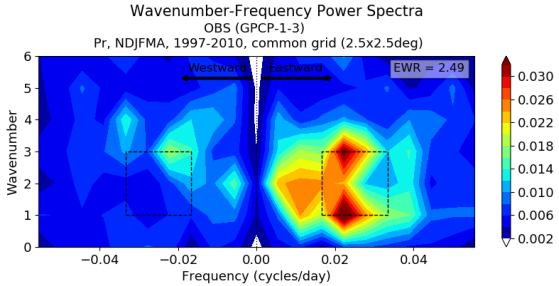
PCMDI

Climate Variability

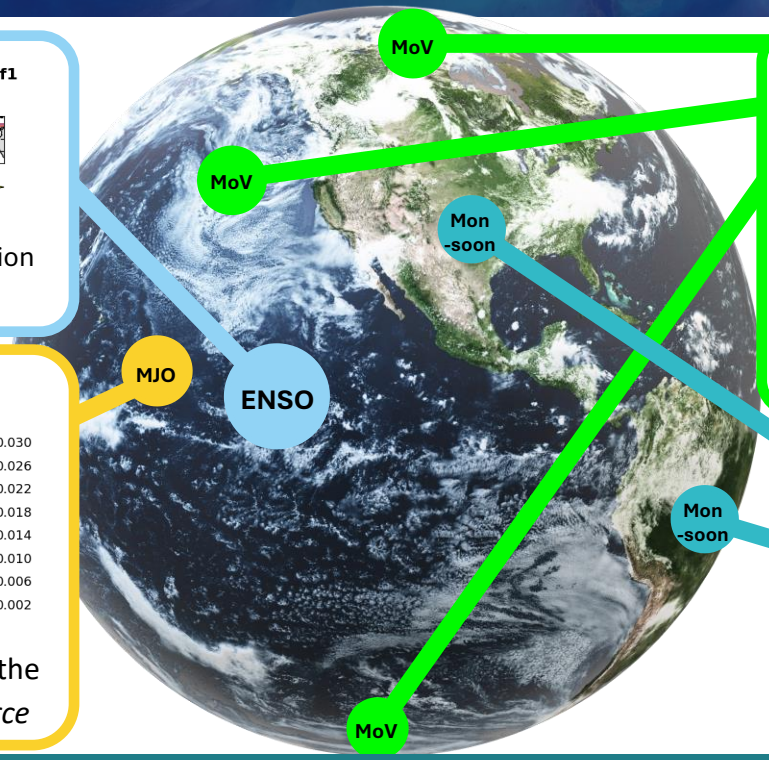
Tropics



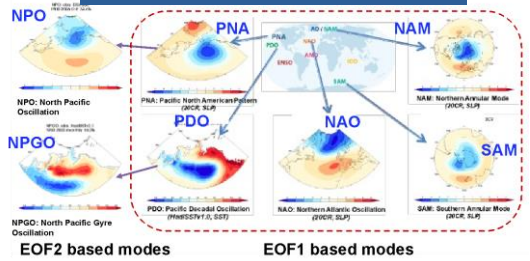
ENSO metrics developed by the collaboration with *CLIVAR Pacific Region Panel*



MJO propagation metrics developed by the collaboration with *CLIVAR MJO Task Force*

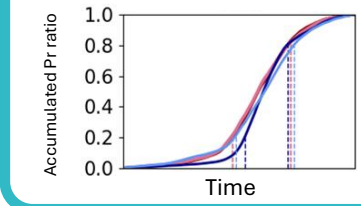


Mid-to-High Latitude



Extratropical **Modes of Variability** *using EOF and CBF*

Monsoon

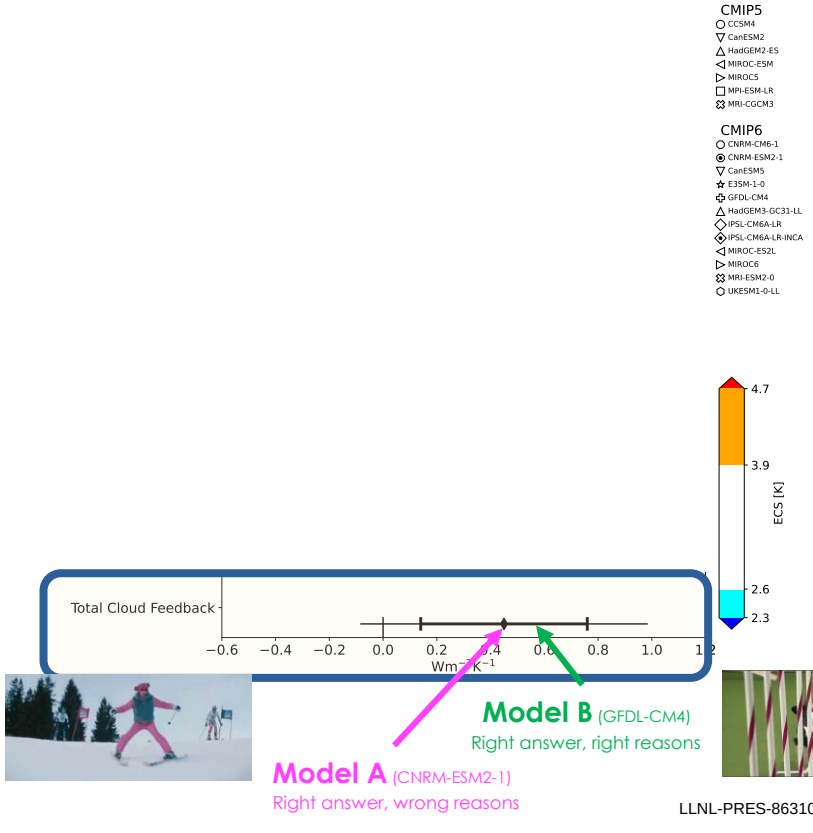


Evaluation of climate variability allows us to explore connectivity in the climate system

Cloud Feedbacks

Implementation contributed by Mark Zelinka and Li-Wei Chao

- **Cloud feedbacks** are broken down into individual components quantified in the WCRP Climate Sensitivity assessment. (Sherwood et al. 2020)
- An overall cloud feedback error metric is computed for each model based on the RMSE across the individual cloud feedback components.
- **Mean-state cloud property** error metrics (Klein et al. 2013) are also computed as part of this package.

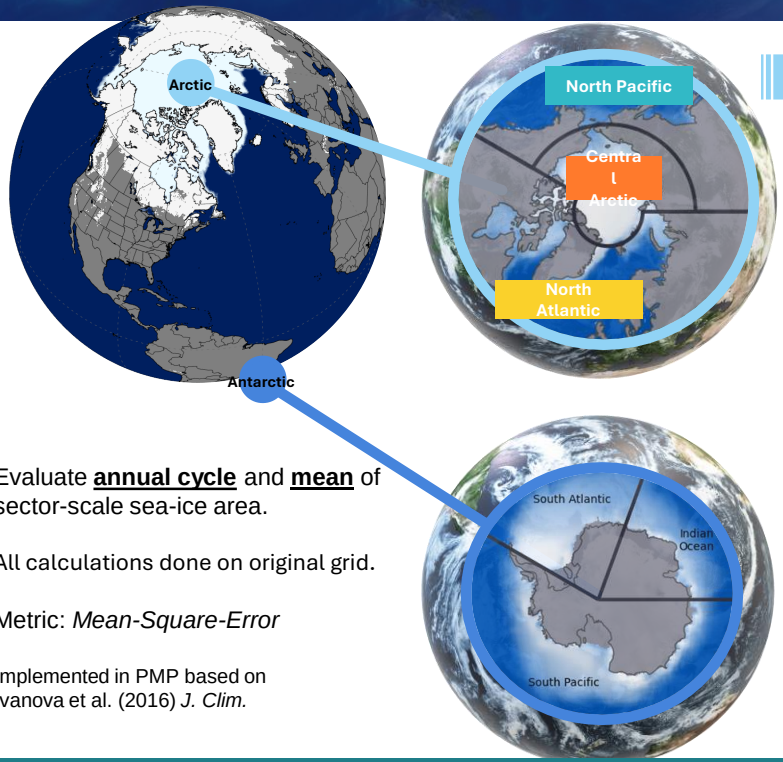


Recent addition: Sea Ice Area

Sector scale analysis: Moving beyond total hemispheric sea-ice extent



Ana Ordonez, Jiwoo Lee, Paul Durack, Peter Glekler

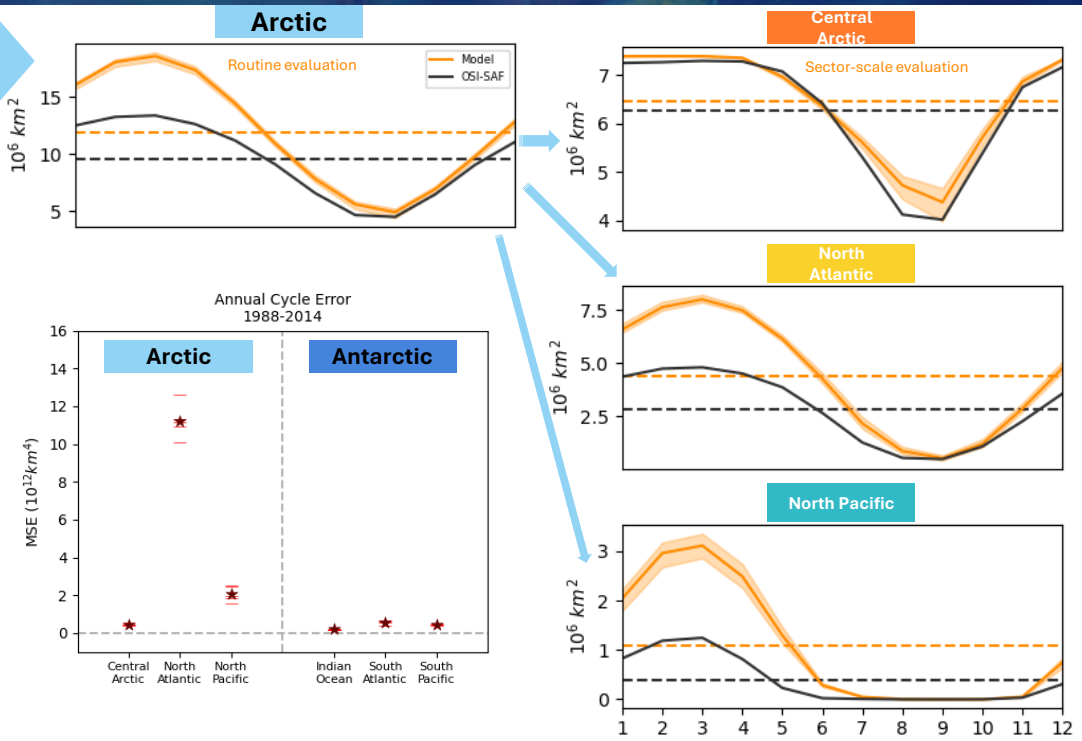


Evaluate **annual cycle** and **mean** of sector-scale sea-ice area.

All calculations done on original grid.

Metric: *Mean-Square-Error*

Implemented in PMP based on Ivanova et al. (2016) *J. Clim.*

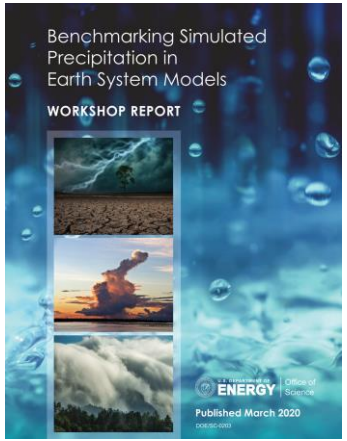


Evaluation efforts have expanded to include more components of the climate system

Please visit Ana Ordonez's poster for details!

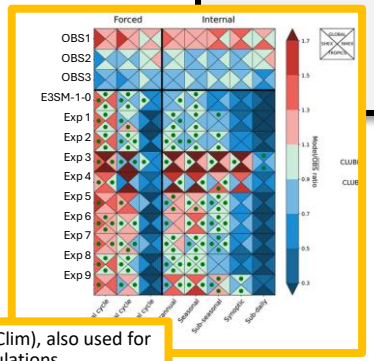
Precipitation Benchmarking (including Extremes)

Min-Seop Ahn, P. J. Gleckler, J. Lee, A. G. Pendergrass, C. Jakob, A. Ordonez, P. Ullrich, M. Wehner, and many others

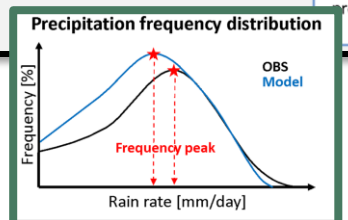


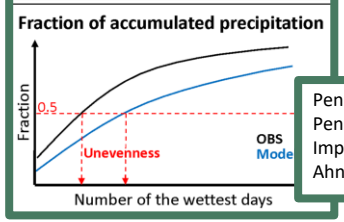
U.S. DEPARTMENT OF ENERGY July 2019 Workshop @DC

Spatial Distribution of Mean State	Seasonal Cycle	Variability Across Timescales	Intensity / Frequency Distributions	Extremes
RMS Error / MAE of mean state	Amplitude + phase of seasonal cycle	Standard deviation at different timescales	Simple Daily Intensity Index (SDII)	Rx1day
Pattern correlation			Unevenness	Rx5day
Monthly mean snow	Monthly score (RMS error)	Diurnal cycle – phase and amplitude	Mean and variance of daily precip.	20y return values (from GEV)
			Perkins score	Rx3h
			Fraction of precipitation days	Seasonal breakdown

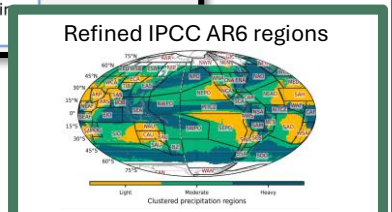


Ahn et al. (2021, J. Clim), also used for examine E3SM simulations.

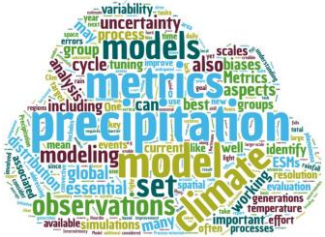


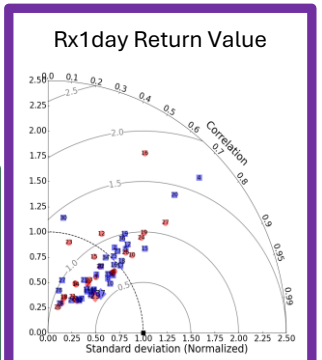


Pendergrass and Deser (2017), Pendergrass and Knutti (2018) Implementation and application by Ahn et al. (2023, GMD)



Ahn et al. 2023 *GMD*

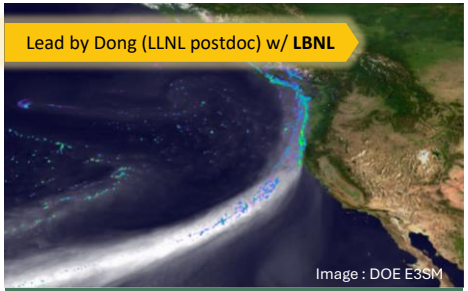




Precipitation Extreme (Wehner et al., 2020, 2021)

Collaborative additions (on-going and planned)

Large-scale Meteorology Driving Extremes



Atmospheric Rivers

Narrow and stretched strong rain band

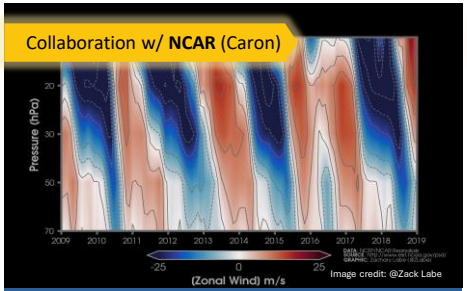
Landfall frequency,
Length and width, etc.

Atmospheric Blocking

Traffic jam in the atmosphere

Frequency, duration, etc.

Stratosphere-troposphere Connections

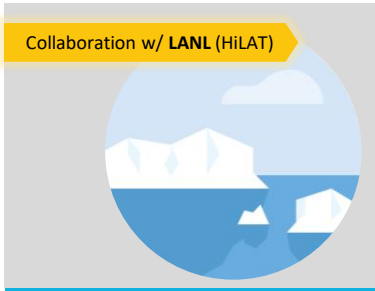


Quasi-Biennial Oscillation (QBO)

Stratosphere oscillation impacting weather

Amplitude, QBO-MJO teleconnection, etc.

Cryosphere



Polar Metrics

Diverse metrics for Arctic/Antarctic regions

We are working on identifying potential candidate metrics

Leveraging TEMPEST EXTREMES

We are leveraging collaborations with the community to incorporate advanced performance measures

Reference datasets

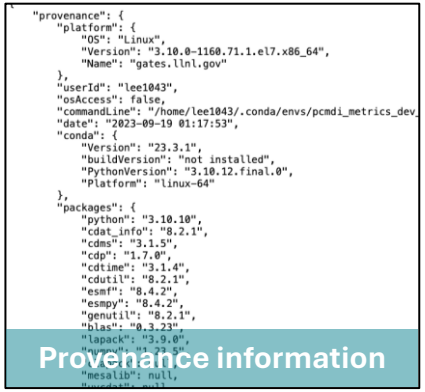
The PMP leverages data products provided by obs4MIPs

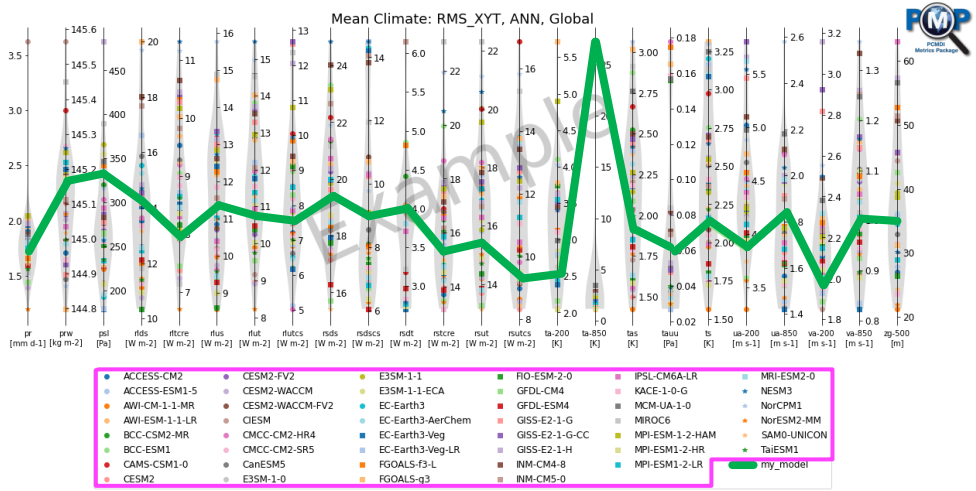
- Obs4MIPs accelerates model evaluation, research and development, via:
  - Technical alignment of **observations** and selected **reanalysis** with CMIP
  - Detailed **provenance** including product origins, data preparation, and unified version control
  - Delivery on ESGF side-by-side with CMIP
- PMP uses dozens of obs4MIPs datasets including daily and 3hr products
- A suite of new obs4MIPs compliant products are now staged for **ESGF** publication
- As a WCRP project, obs4MIPs is expected to be a critical resource for **CMIP benchmarking**
- PCMDI, NASA and ESA are providing leadership

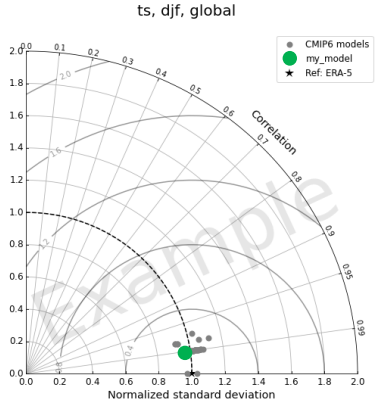
For further information contact Peter Gleckler (gleckler1@llnl.gov)

Database of pre-calculated metrics provides a foundation for benchmarking

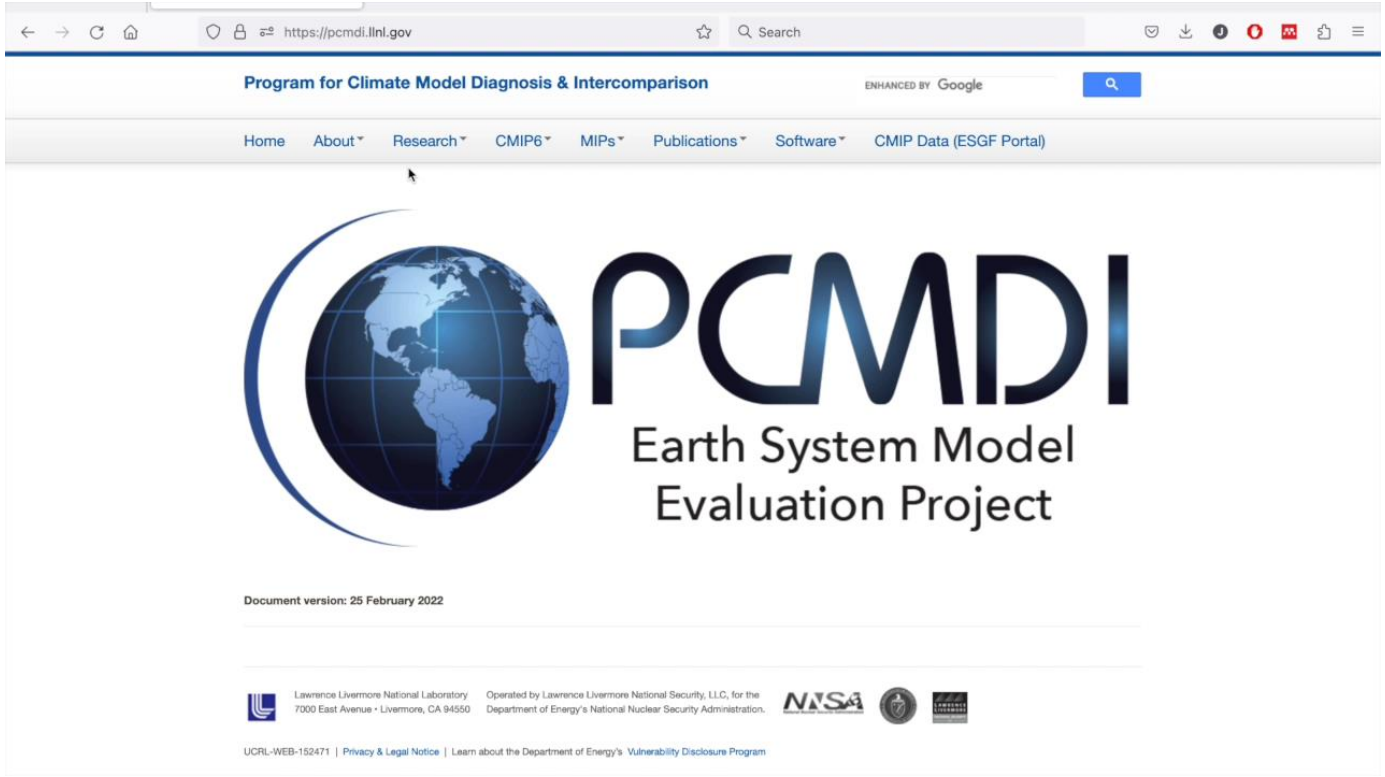
- We release a **database of pre-calculated PMP metrics for CMIP5/6** https://github.com/PCMDI/pcmdi_metrics_results_archive
- which enables **objective benchmarking** (i.e., comparing the user's model with other CMIP models) *without* massive downloading of CMIP models to calculate the same metrics for them.
- **Provenance** information ensures robust **reproducibility** of the statistical numbers of metrics.







PMP results interactive visualization



https://pcmdi.llnl.gov/metrics/

Thank You!

http://pcmdi.github.io/pcmdi_metrics/

Lee et al. 2024
Systematic and Objective Evaluation of Earth System Models: PCMDI Metrics Package (PMP) version 3. *Geoscientific Model Development*, 17, 3919–3948, doi: 10.5194/gmd-17-3919-2024

Lawrence Livermore National Laboratory

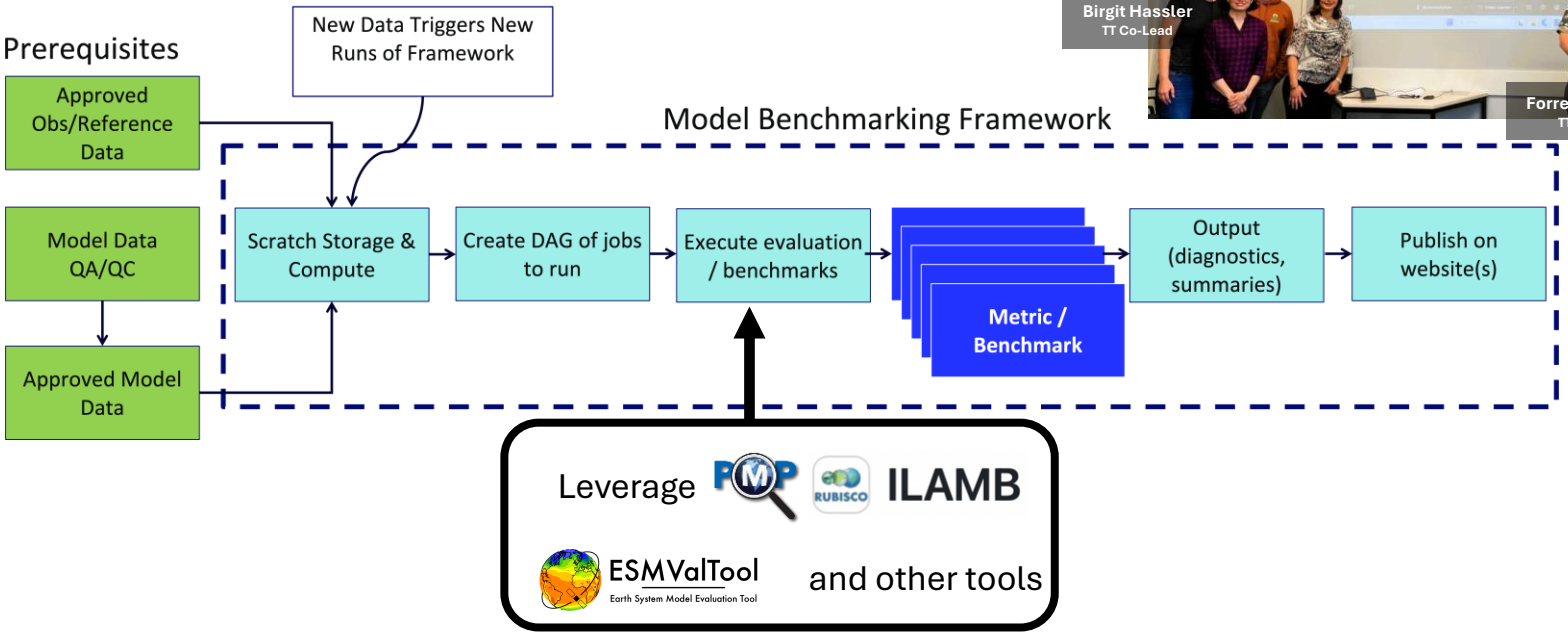
U.S. DEPARTMENT OF ENERGY | Office of Science

PCMDI
Program for Climate Model Diagnosis and Intercomparison

PMP contributes to CMIP Model Benchmarking

Engagement with the CMIP Benchmarking Task Team

Proposal for **CMIP7 Rapid Evaluation Framework**



CMIP Model Benchmarking

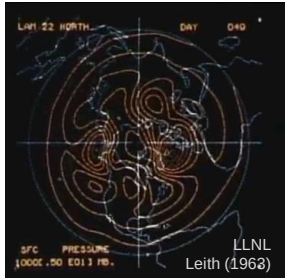
Task Team Meeting @Germany, DLR (2024 May)

WCRP ESMO

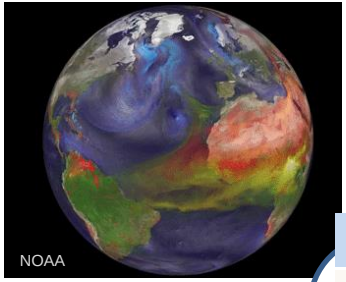
Climate Models are Essential Tools for Understanding Climate Change

Evolutions of Earth System/Climate Models

In 1960s

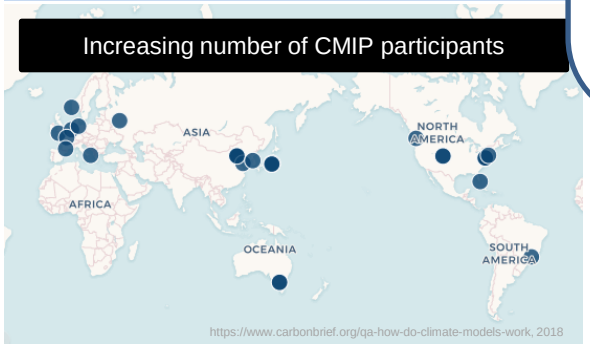


Current



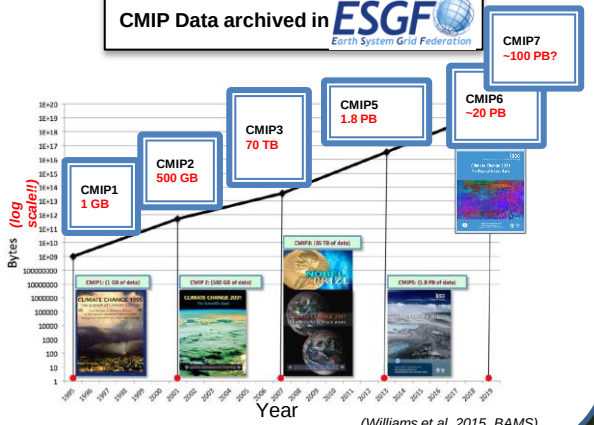
Diversified Earth System/Climate Models

Increasing number of CMIP participants



Outburst of Data

CMIP Data archived in ESGF Earth System Grid Federation



(Williams et al. 2015, BAMS)

Earth & Environmental Systems Modeling

Nathan Collier, Forrest Hoffman, Dave Lawrence

Methodological Developments in the International Land Model Benchmarking (ILAMB) Effort

U.S. DEPARTMENT OF ENERGY | Office of Science

2024 EESM PI Meeting
August 6–9, 2024

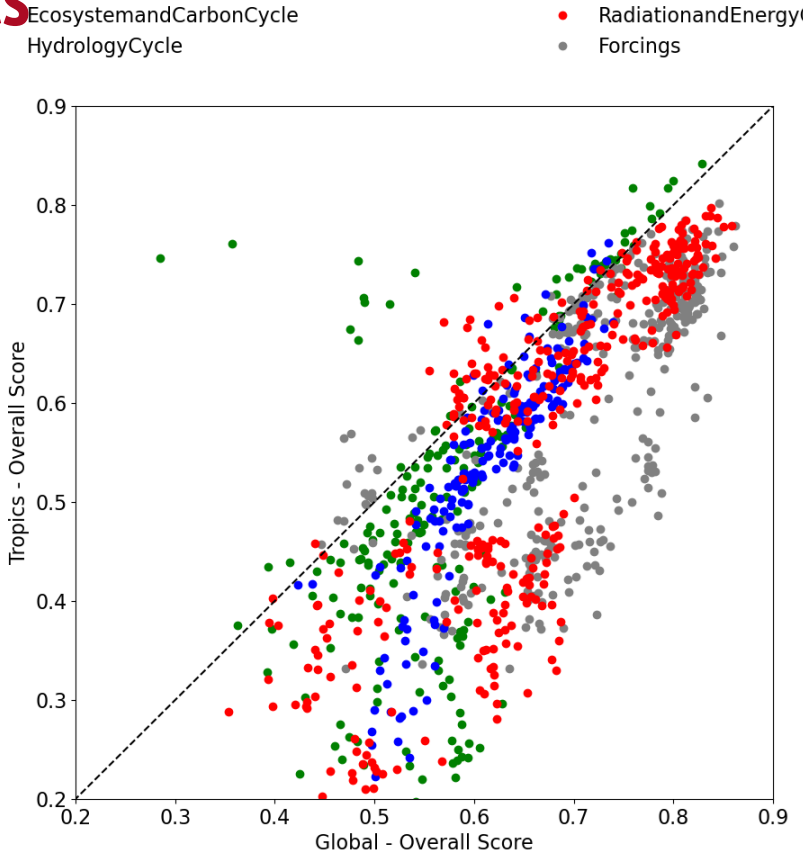
Methodological Developments in the International Land Model Benchmarking (ILAMB) Effort

Nathan Collier, Forrest Hoffman, Dave Lawrence

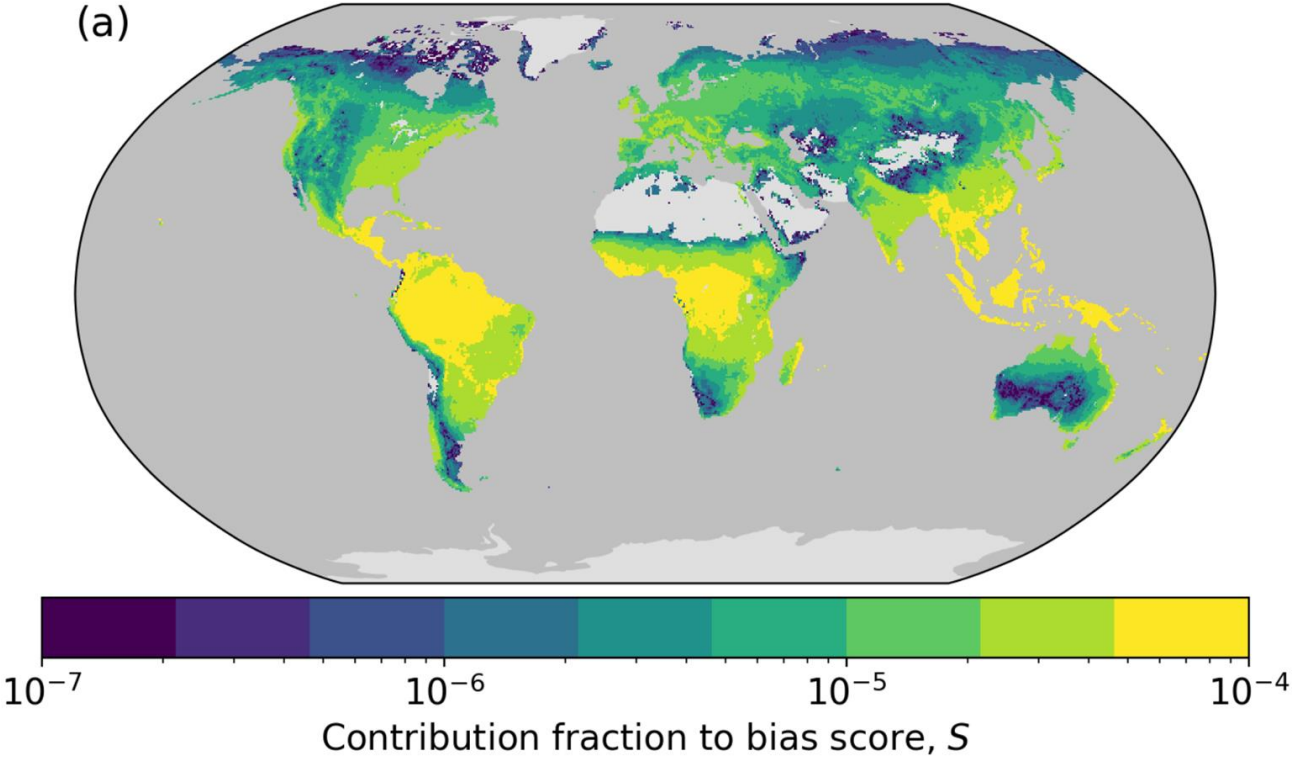
RUBISCO

Heavily Influenced by Tropics

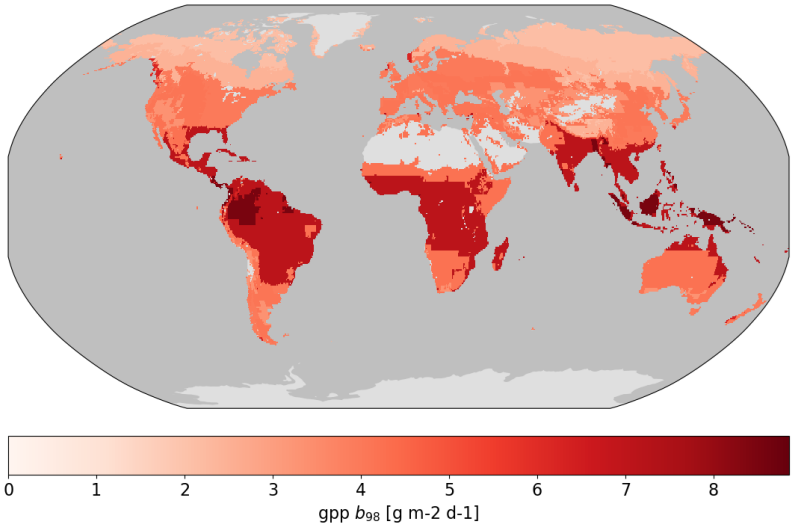
- We have observed that our current scoring methodology favors performance in the tropics.
- Plot shows that for many datasets, the tropics score correlates strongly to the global score
- This is due to our choice of normalizing errors by the variability of the reference data and the use of *mass weighting*



Bias score influence map: gpp | FLUXCOM



A Change in How We Normalize Errors

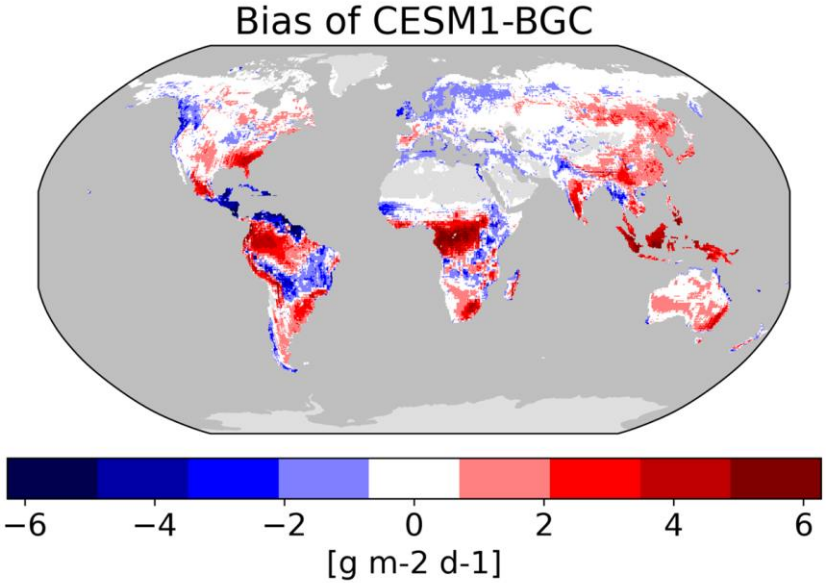


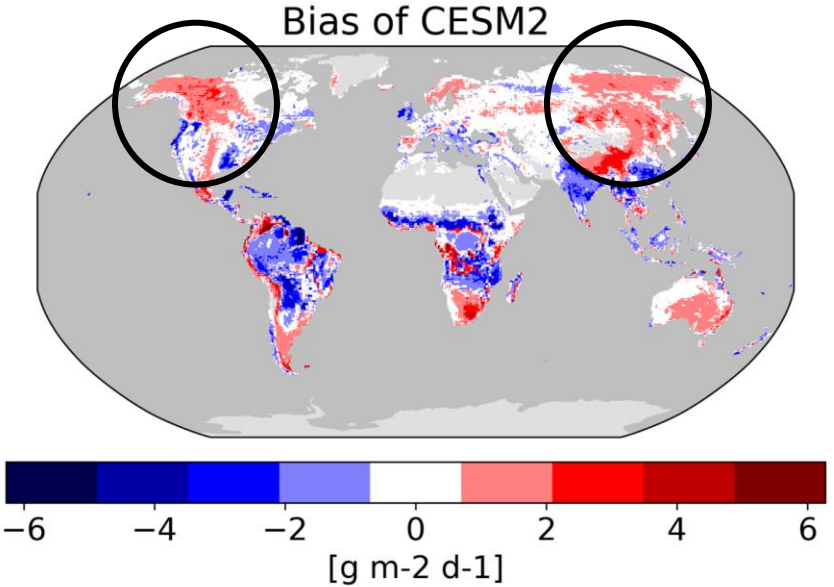
The gpp 98th error quantile within Whittaker biomes across CMIP5v6 models.

- The goal is to make errors from different areas of the globe comparable.
- Select a set of regions which represent biomes in which errors can be treated as commensurate in order of magnitude.
- Inside each region, for each variable, and across a selection of models, compute the 98th quantile of |bias(x)| with respect to all datasets for that variable.

CESM2 | gpp | FLUXCOM

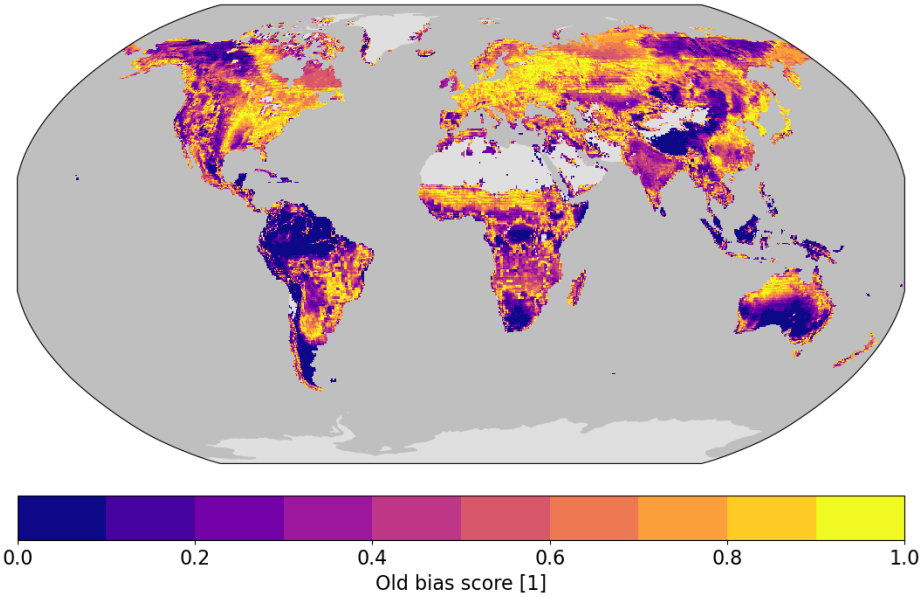
Notice larger bias in high latitudes, anomalous among CMIP models

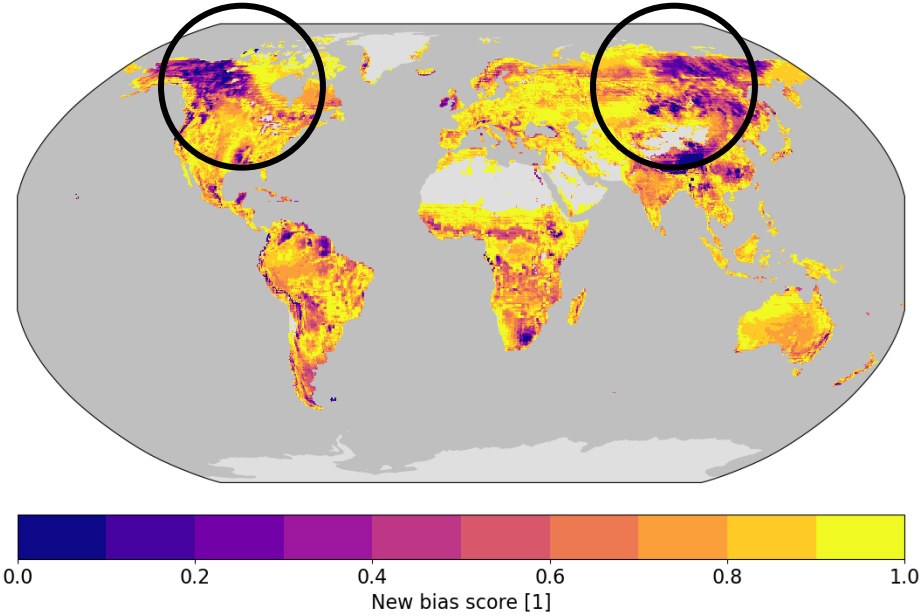




CESM2 | gpp | FLUXCOM

With the new methodology, these areas light up clearly





Other Ways to Use ILAMB Data

```
In [1]: import intake
   ...: cat = intake.open_catalog("https://raw.githubusercontent.com/nocollier/intake-ilamb/main/ilamb.yaml")

In [2]: cat['
```

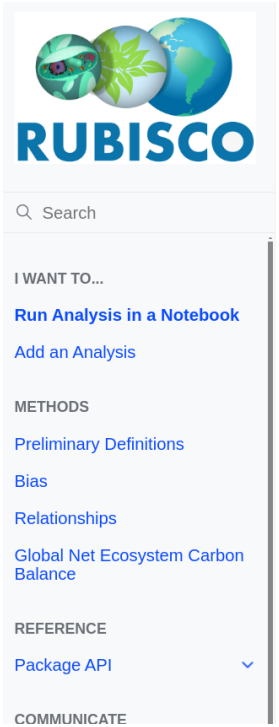
'albedo | CERESed4.1'	'lai | MODIS'	'rlns | GEWEX.SRB'
'albedo | GEWEX.SRB'	'mrro | CLASS'	'rlns | WRMC.BSRN'
'biomass | ESACCI'	'mrro | Dai'	'rlus | CERESed4.1'
'biomass | NBCD2000'	'mrro | LORA'	'rlus | FLUXNET2015'
'biomass | Thurner'	'mrsos | WangMao'	'rlus | GEWEX.SRB'
'biomass | Tropical'	'nbp | GCP'	'rlus | WRMC.BSRN'
'biomass | US.FOREST'	'nbp | Hoffman'	'rns | CERESed4.1'
'burntFractionAll | GFED4.1S'	'nee | FLUXCOM'	'rns | CLASS'
'cSoil | HWSD'	'nee | FLUXNET2015'	'rns | FLUXNET2015'
'cSoil | NCSCDV22'	'pfext | NSIDC'	'rns | GEWEX.SRB'
'co2 | NOAA.GMD'	'pr | CLASS'	'rns | WRMC.BSRN'
'dtr | CRU4.02'	'pr | CMAPv1904'	'rsds | CERESed4.1'
'evspsbl | GLEAMv3.3a'	'pr | FLUXNET2015'	'rsds | FLUXNET2015'
'evspsbl | MOD16A2'	'pr | GPCCv2018'	'rsds | GEWEX.SRB'
'evspsbl | MODIS'	'pr | GPCPv2.3'	'rsds | WRMC.BSRN'
'fBNF | DaviesBarnard'	'reco | FLUXCOM'	'rsns | CERESed4.1'
'gpp | FLUXCOM'	'reco | FLUXNET2015'	'rsns | FLUXNET2015'
'gpp | FLUXNET2015'	'regions_continental | ILAMB'	'rsns | GEWEX.SRB'
'gpp | WECANN'	'regions_continental | IPCC'	'rsns | WRMC.BSRN'
'hfdsl | CLASS'	'regions_global_land | ILAMB'	'rsus | CERESed4.1'
'hfls | CLASS'	'regions_global_land_no_ant | ILAMB'	'rsus | FLUXNET2015'
'hfls | DOLCE'	'regions_whittaker_biomes | ILAMB'	'rsus | GEWEX.SRB'
'hfls | FLUXCOM'	'rhums | CRU4.02'	'rsus | WRMC.BSRN'
'hfls | FLUXNET2015'	'rhums | ERA5'	'swe | CanSISE'
'hfls | WECANN'	'river_basins | Dai'	'tas | CRU4.02'
'hfss | CLASS'	'rlds | CERESed4.1'	'tas | FLUXNET2015'
'hfss | FLUXCOM'	'rlds | FLUXNET2015'	'tasmax | CRU4.02'
'hfss | FLUXNET2015'	'rlds | GEWEX.SRB'	'tasmin | CRU4.02'
'hfss | WECANN'	'rlds | WRMC.BSRN'	'twsa | GRACE'
'lai | AVH15C1'	'rlns | CERESed4.1'	
'lai | AVHRR'	'rlns | FLUXNET2015'	

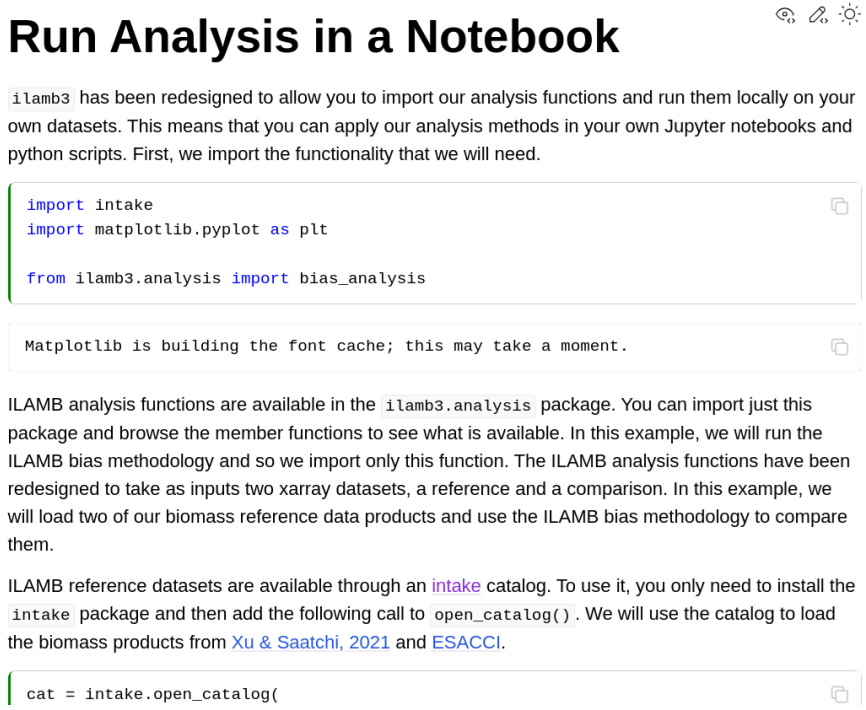
Other Ways to Use ILAMB Data

```
In [1]: import intake
   ...: cat = intake.open_catalog("https://raw.githubusercontent.com/nocollier/intake-ilamb/main/ilamb.yaml")

In [2]: gpp = cat['gpp | WECANN'].read()

In [3]: gpp
Out[3]:
<xarray.Dataset>
Dimensions:      (time: 108, nb: 2, lat: 180, lon: 360)
Coordinates:
  * time         (time) object 2007-01-16 12:00:00 ... 2015-12-16 12:00:00
  * lat          (lat) float64 89.5 88.5 87.5 86.5 ... -86.5 -87.5 -88.5 -89.5
  * lon          (lon) float64 -179.5 -178.5 -177.5 -176.5 ... 177.5 178.5 179.5
Dimensions without coordinates: nb
Data variables:
    time_bounds  (time, nb) object 2007-01-01 00:00:00 ... 2016-01-01 00:00:00
    gpp          (time, lat, lon) float64 9.969e+36 9.969e+36 ... 9.969e+36
Attributes:
    title:         Water, Energy, and Carbon with Artificial Neural Networks ...
    version:       1
    institutions:  Columbia University
    source:        Solar Induced Fluorescence (SIF), Air Temperature, Precipi...
    history:       \n2020-11-02: downloaded https://avdc.gsfc.nasa.gov/pub/da...
    references:    \n@ARTICLE{Alemohammad2017,\n  author = {Alemohammad, S. H...
    comments:      \ntime_period: 2007-01 through 2015-11; temporal_resolutio...
    convention:    CF-1.8
```

ILAMB v3: On the Horizon

- Shift to xarray as a base object.
- Adapt to the way researchers are working.
- Working from the bottom up and making soft *releases* as we go.
- Each new capability will be fully documented.
- Great time to get me your wish lists.





Run Analysis in a Notebook

`ilamb3` has been redesigned to allow you to import our analysis functions and run them locally on your own datasets. This means that you can apply our analysis methods in your own Jupyter notebooks and python scripts. First, we import the functionality that we will need.

```
import intake
import matplotlib.pyplot as plt

from ilamb3.analysis import bias_analysis
```

```
Matplotlib is building the font cache; this may take a moment.
```

ILAMB analysis functions are available in the `ilamb3.analysis` package. You can import just this package and browse the member functions to see what is available. In this example, we will run the ILAMB bias methodology and so we import only this function. The ILAMB analysis functions have been redesigned to take as inputs two xarray datasets, a reference and a comparison. In this example, we will load two of our biomass reference data products and use the ILAMB bias methodology to compare them.

ILAMB reference datasets are available through an intake catalog. To use it, you only need to install the `intake` package and then add the following call to `open_catalog()`. We will use the catalog to load the biomass products from Xu & Saatchi, 2021 and ESACCI.

```
cat = intake.open_catalog(
```

https://github.com/rubisco-sfa/ilamb3