Perturbed Parameter Ensemble (PPE)-based model calibration in the presence of model structural error

Qingyuan Yang[1,*], Addisu G Semie[1],
Gregory S Elsaesser[1,2], Brian Medeiros[1,3], Da Fan[1] and many others

1. Learning the Earth with Artificial Intelligence and Physics (LEAP) NSF Science and Technology Center 2. NASA Goddard Institute for Space Studies 3. NSF National Center for Atmospheric Research
*qy2288@columbia.edu

What is PPE

- PPE: a set of parameters and the corresponding climate model simulations
- Several tens of parameters perturbed
- Several hundreds of simulations

X: the parameters

	micro_mg_max_nicons	micro_mg_vtrmi_factor	micro_mg_iaccr_factor
1	1.000000e+08	1.200000	1.000000
2	1.000000e+08	1.000000	1.000000
3	1.000000e+08	1.366512	1.000000
4	4.272341e+09	0.961199	0.903760
5	2.412386e+09	1.154306	0.643733

…

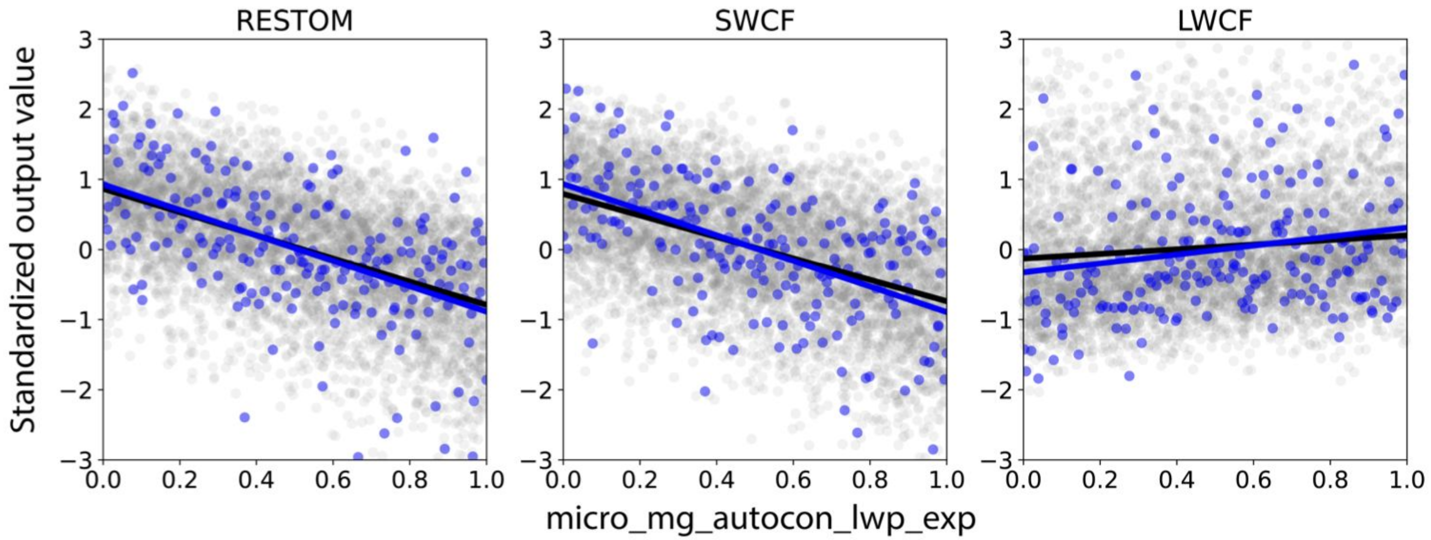
Y: processed, climate model output

ppe_ind	RESTOM_zonal_-75to-65	RESTOM_zonal_-65to-55	RESTOM_zonal_-55to-45
1	-100.057060	-70.938034	-48.529465
2	-100.405266	-71.161350	-48.245613
3	-100.752030	-71.381584	-48.530212
4	-97.482056	-64.994290	-36.864660
5	-99.494026	-69.113540	-43.387040

…

Why do we need PPEs?

- Help us understand the model
  - Relationship between model parameters and simulation results

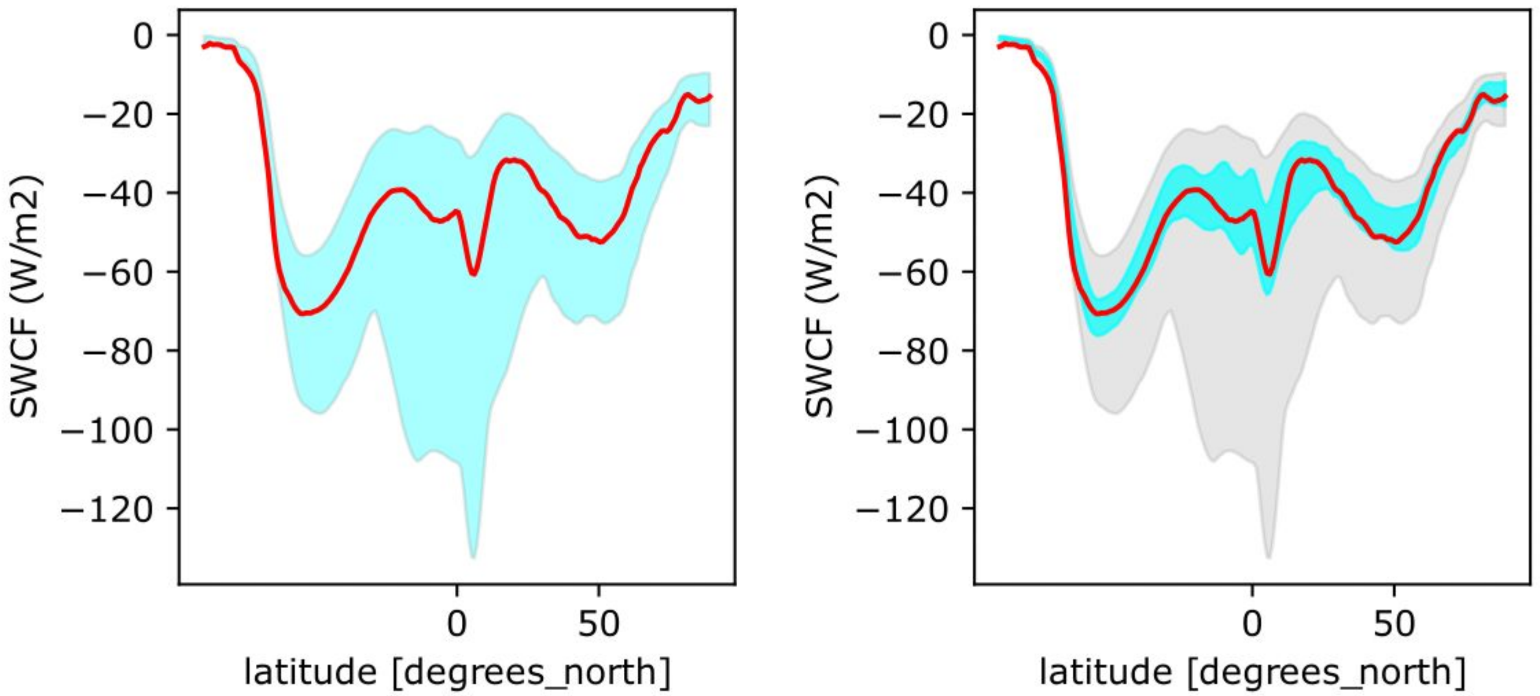


Eidhammer et al (2024)

- They can be used for model calibration

Why do we need calibration?

- Parameterizations are simplifications of the real physics
- The corresponding parameters could have a significant impact on the climate model output
- There are no correct parameter values, only optimum ones



Red: observation
Envelopes: min-max from the PPEs

Model calibration

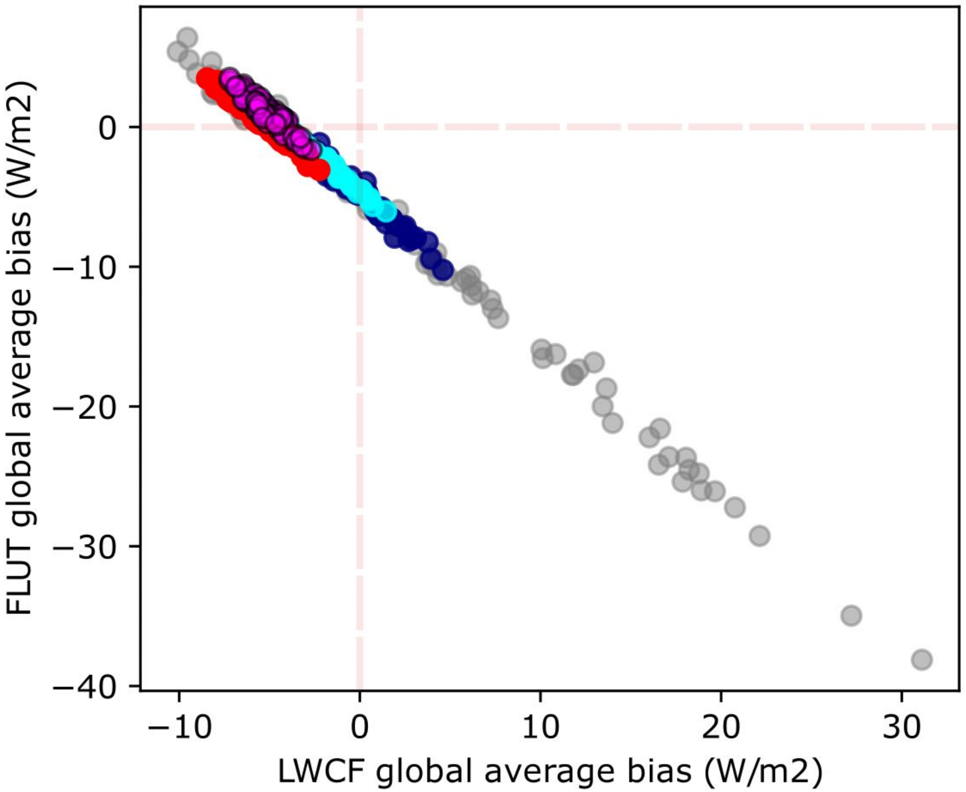
- Use the PPE as training data
  - The parameters are the X and the outputs are the Y (Y could be global averages, regional averages, seasonal averages …)
  - Build an emulator $Y = f(X)$
  - We have observations Y_{obs}

Find the best X that best match Y_{obs}

We can have infinite parameter value combinations, and then use them to find the optimum parameters, however…

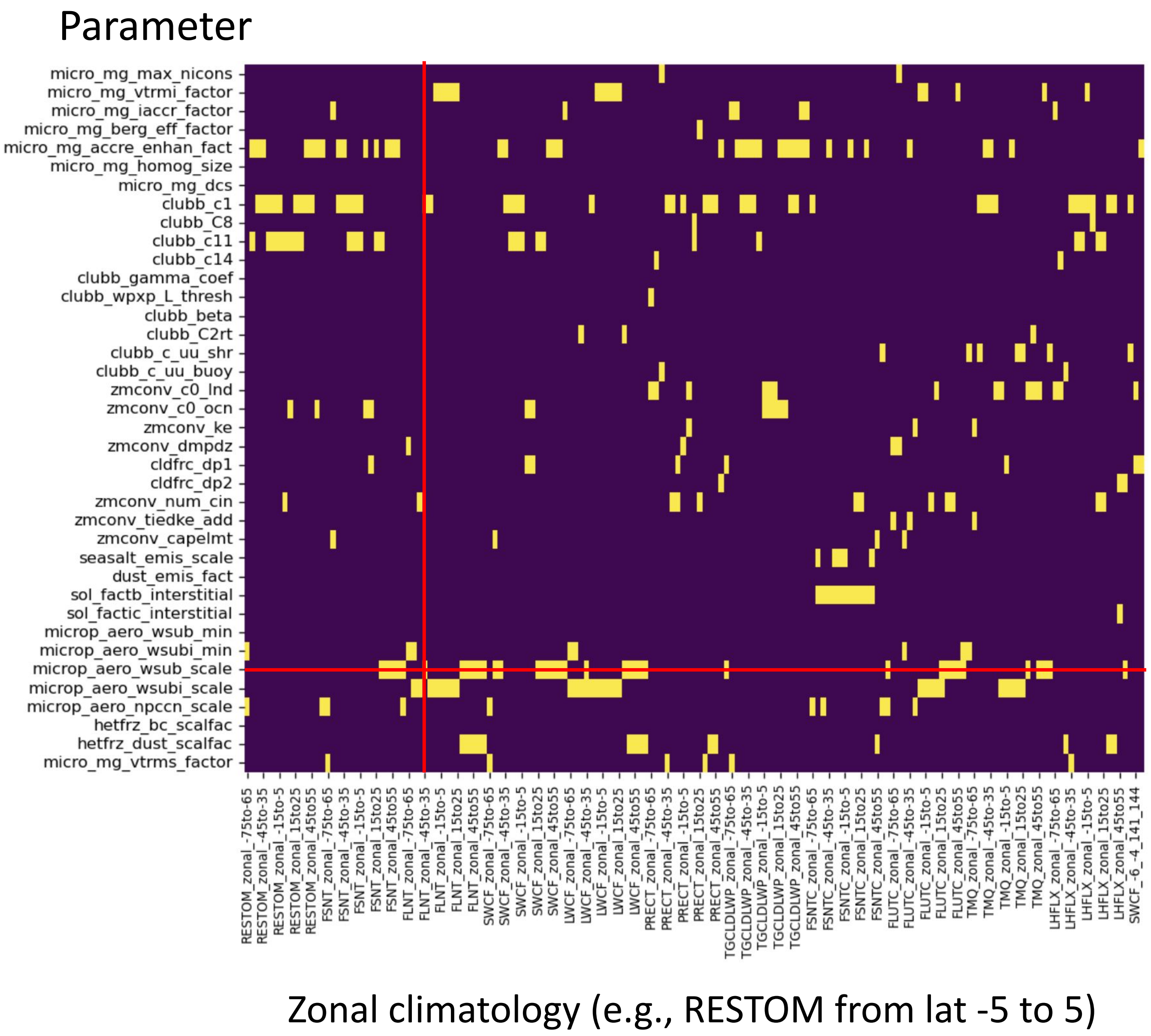
What makes this difficult for climate models?

- Limited data
  - If we have 30 parameters (e.g., 2^{30} is a huge number)
- We have many targets
  - RESTOM, Cloud Forcing at different latitude ranges
- Internal model structural error
  - Something we cannot resolve by varying the parameter values



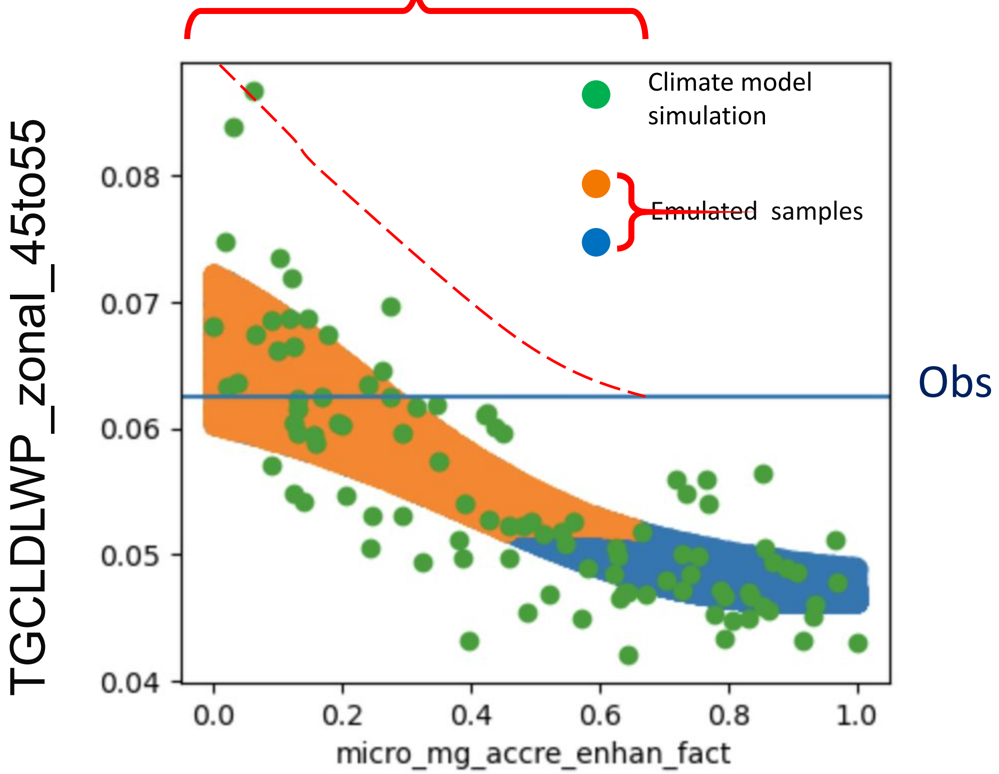
Our method

- Gaussian Process-based (allows uncertainty quantification), and each target y
- For each variable of interest, we construct our emulator as:
  - $y^i = GP^i\left(x_{para_i}, x_{para_j}\right)$
  - x_{para_i} and x_{para_j} are the two most sensitive parameters (automatically selected)



What's next?

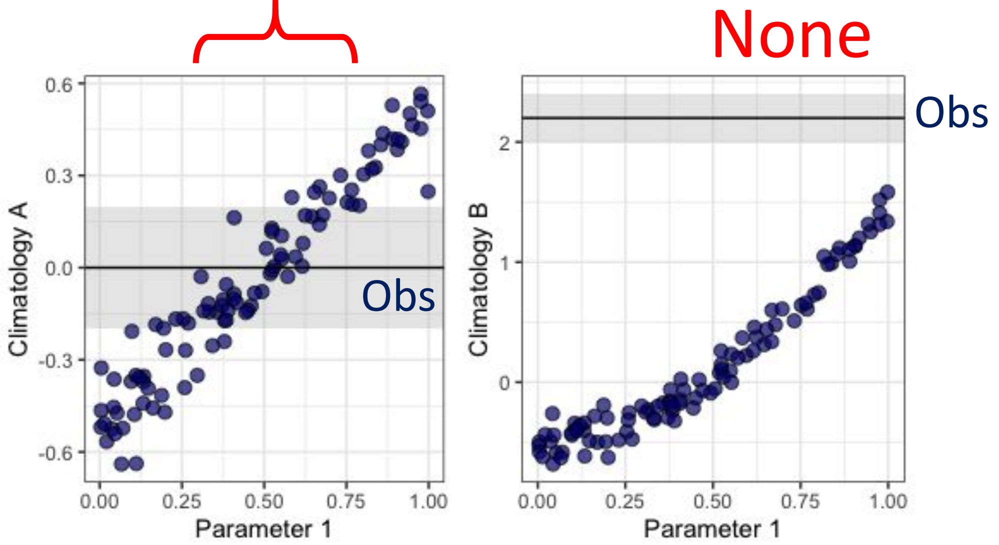
- Parameter estimation
  - History matching ✓ Bayesian method is not used
  - Likelihood is binary (yes or no)



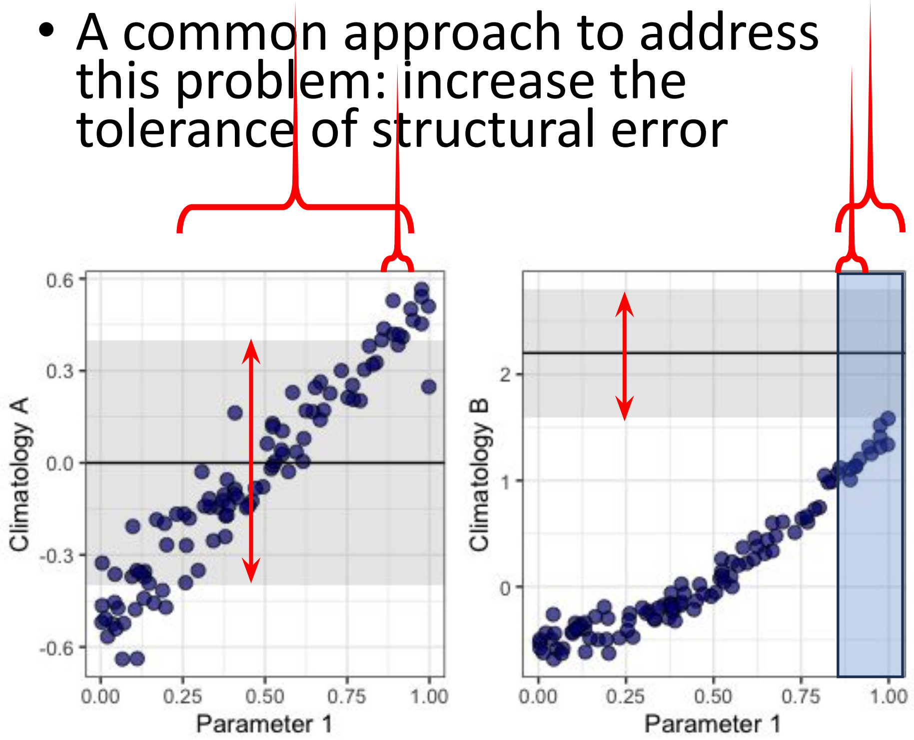
Similar procedure for many variables of interest

The challenge of structural error

- Should Parameter 1 be around 0.5 or close to 1.0?



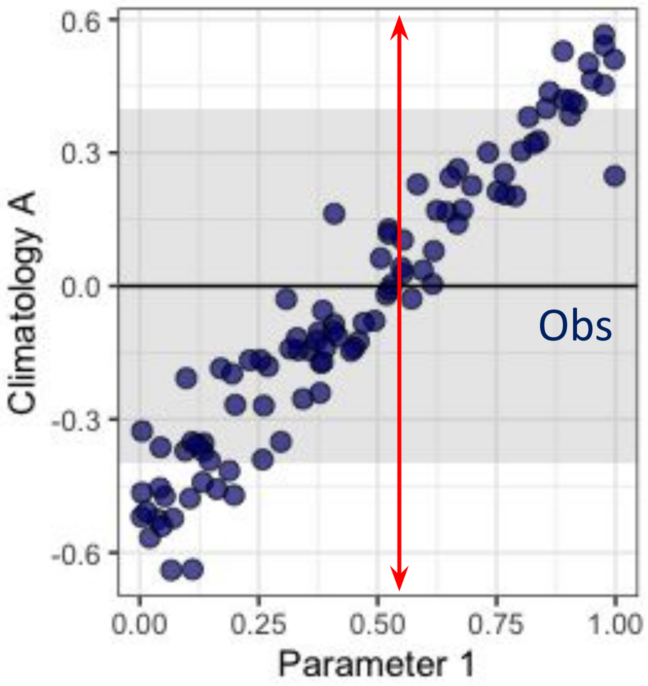
- A common approach to address this problem: increase the tolerance of structural error

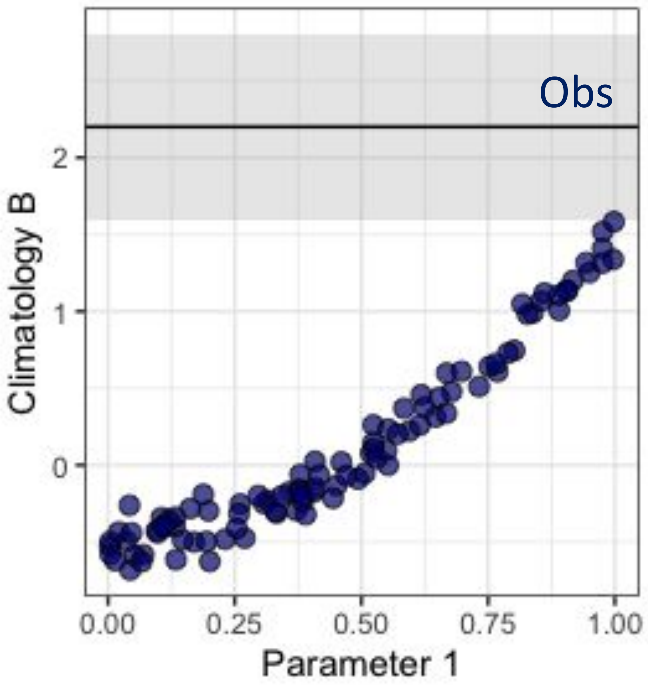


- Bias towards the observations that are harder to match

Additional risk

Too much tolerance leads to waste of useful observations





What should we do?
Avoid arbitrarily increasing the tolerance of structural error

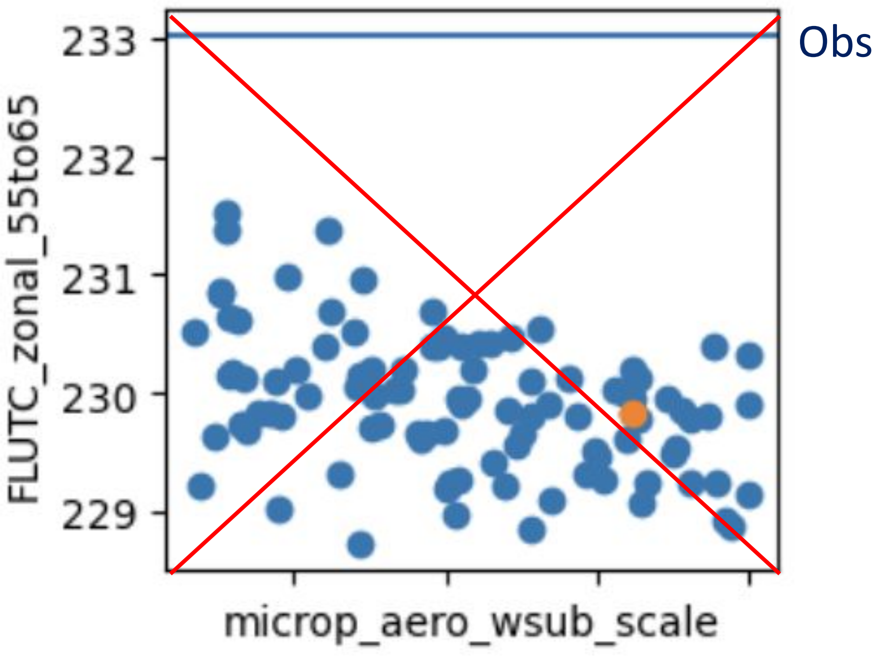
Report the occurrence of structural error when it occurs

Let's sweep it under the carpet:
Ignore the presence of structural error

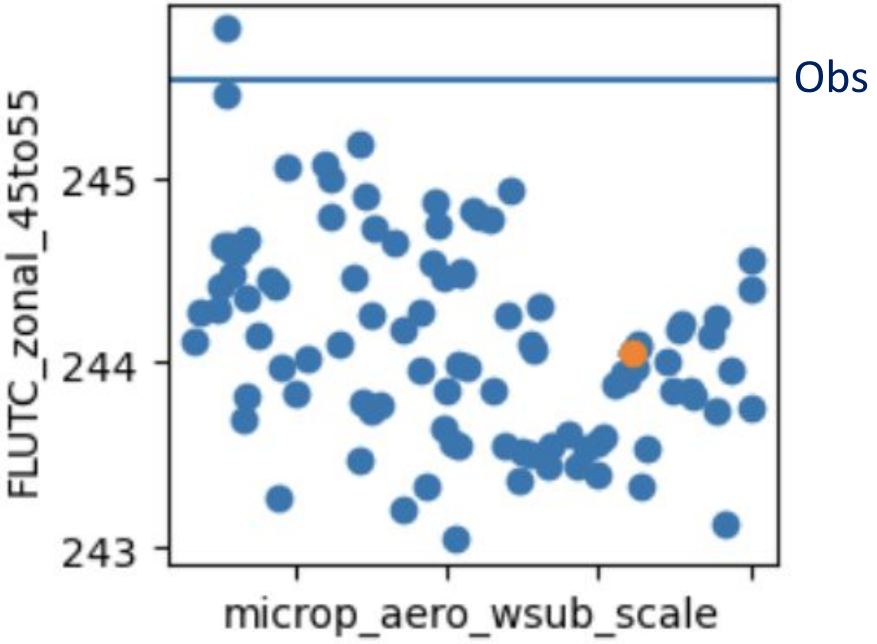
Why can we sweep it under the carpet?

If some local climatology is subject to strong structural error, the climatology nearby could have less severe structural error, but also similar to it.

FLUTC lat: 55-65 N



FLUTC lat: 45-55 N



Our method

- Emulator: Gaussian Process
- Parameter estimation: History matching
- Features:
  - Do not increase the tolerance for structural error
  - Detect structural error and neglect it

```
para_seq = list(test_case.grouped_hulls.keys())
check = orchestrate_test(para_seq, test_case.p_emu, test_case.tf_masks,
                         test_case.para_nm, test_case.grouped_hulls, test_case.paras_vars,
                         n_pts=  10000, n_threshold = 100,sample_threshold = 10**6, max_workers = 31)
```

```
Drop ['PRECT_zonal_45to55']
============================================================
Running ('clubb_c_uu_shr', 'microp_aero_wsubi_scale'), the 7th simulation
There is overlap for ('clubb_c_uu_shr', 'microp_aero_wsubi_scale'). Proceed
============================================================
```

To be made public on Github soon.

Test

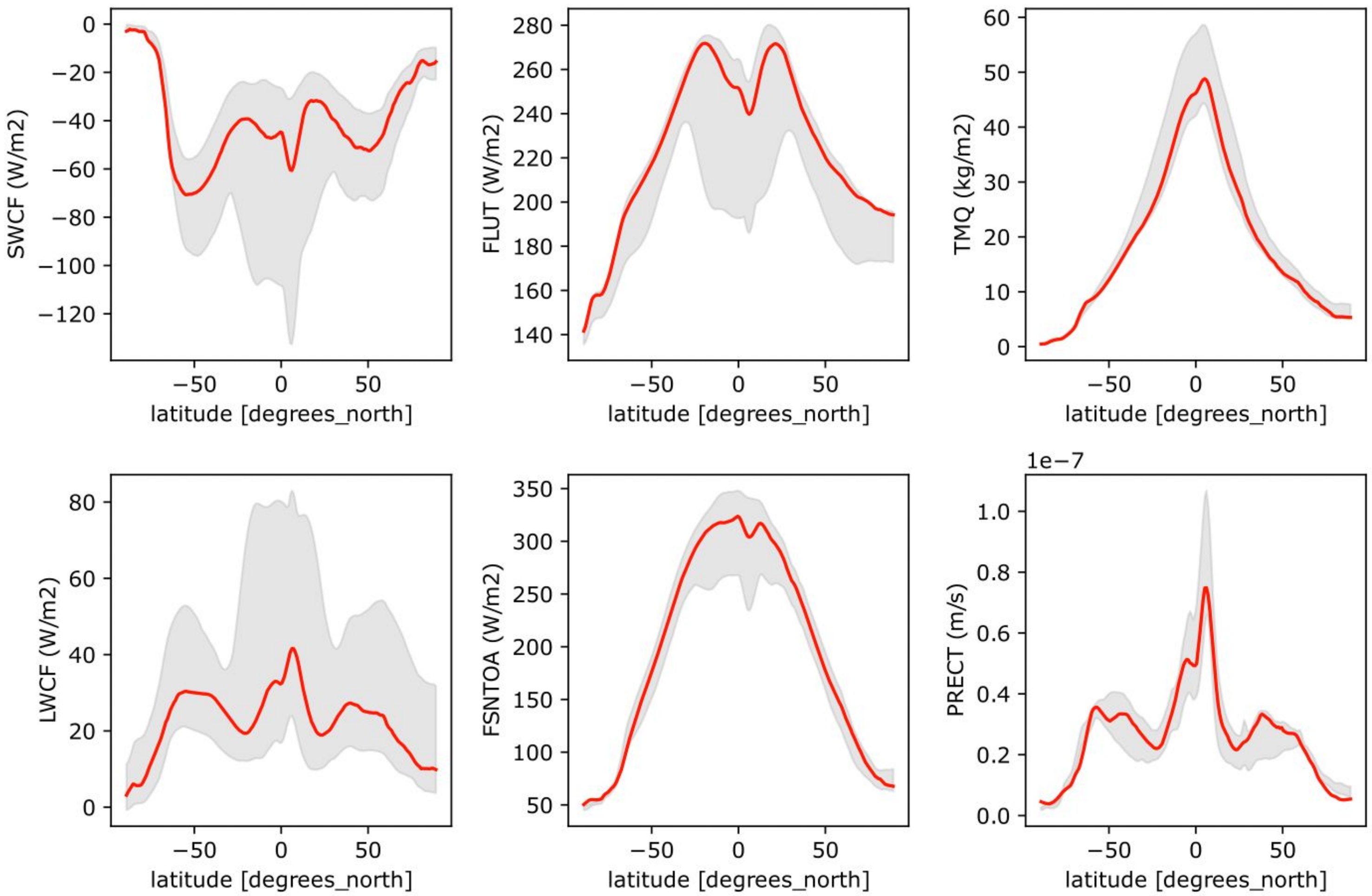
- Method test:

CAM6 with ML-based microphysics parameterization

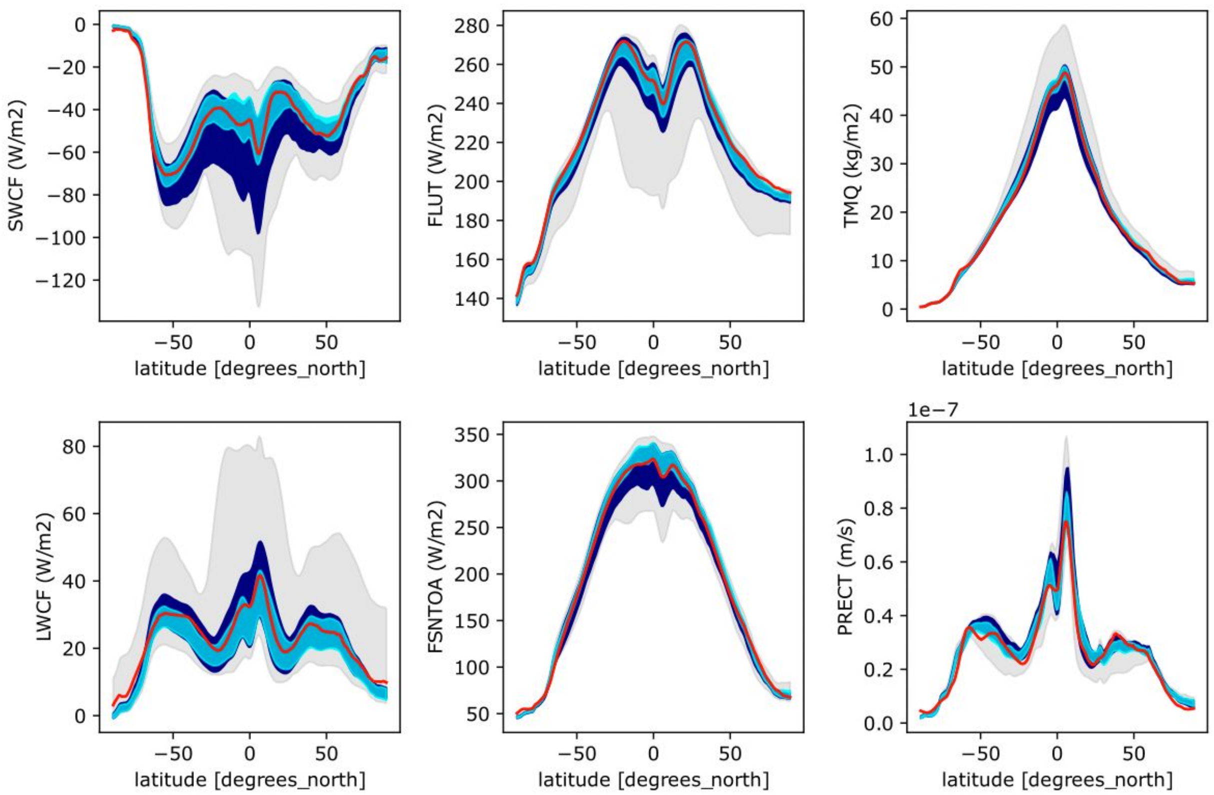
Uncoupled simulation for one year

100 parameters samples with 34 parameters perturbed

Targets: LWCF, SWCF, FLUT, FSNTOA, TMQ, PRECT 10°-zonal averages



Results



Type	Global bias		Global RMSE	
Iteration (ensemble number)	CAM6 default	Best member	CAM6 default	Best member
SWCF (W/m2)	-2.14	-0.77	10.60	8.05
LWCF (W/m2)	-1.58	-1.17	5.26	4.43
FSNTOA (W/m2)	0.73	1.80	10.10	8.65
FLUT (W/m2)	-2.82	-3.36	6.78	5.94
TMQ (kg/m2)	-0.13	0.58	1.82	1.76
PRECT (m/2)	9.23E-11	1.65E-10	1.10E-08	1.09E-08
TGCLDLWP (kg/m2)	0.01	-0.03	0.04	0.03
CLDTOT_ISCCP (%)	-13.45	-12.83	16.97	15.63

Testing for CAM7

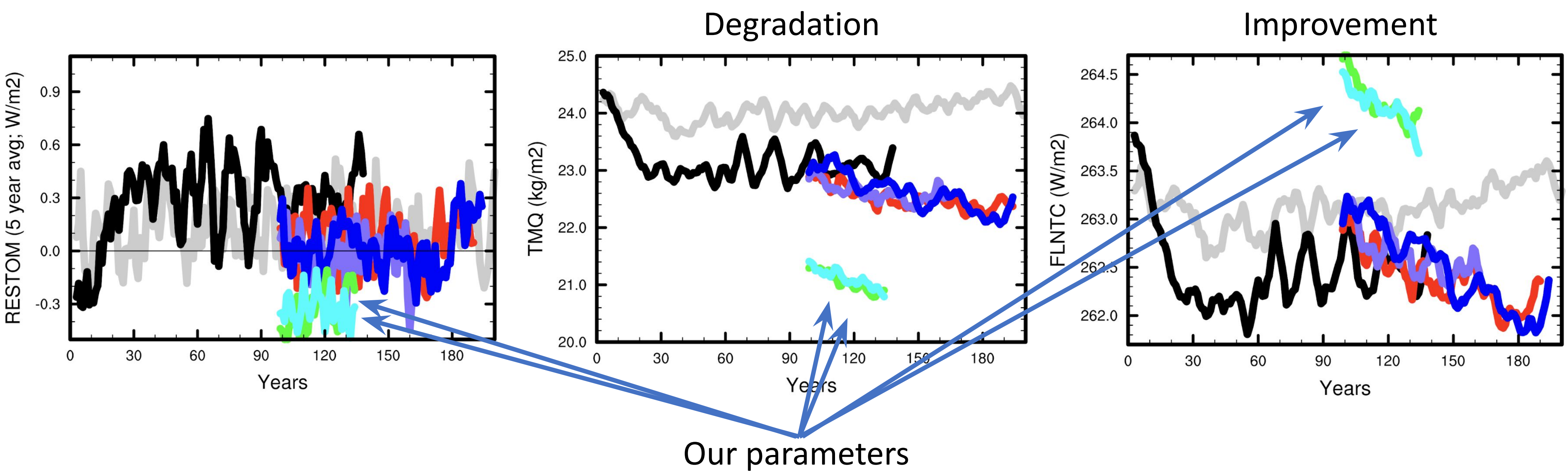
We generate a CAM7 PPE based on uncoupled simulations, and obtain two sets of optimum parameters and use them for the coupled simulation

Some improvement, but RESTOM is biased.

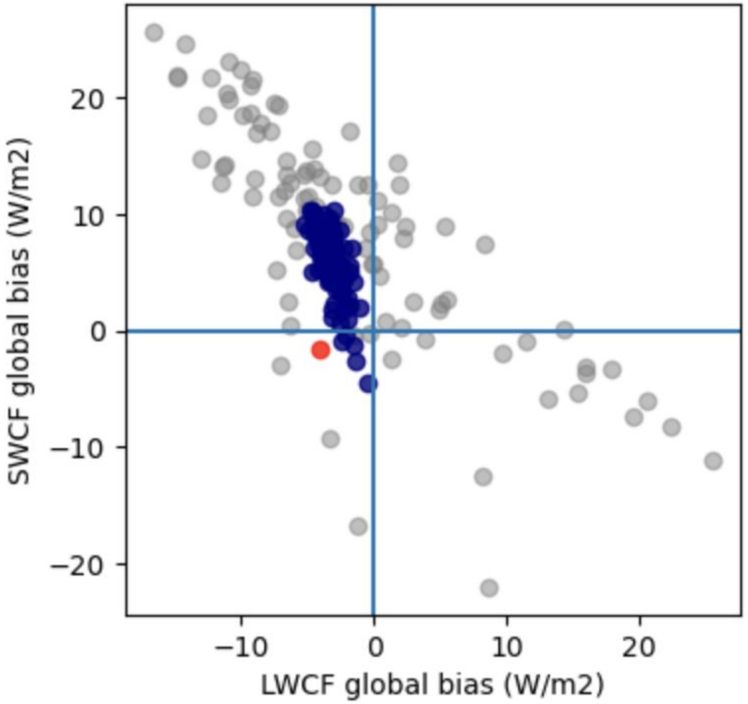
Main causes:

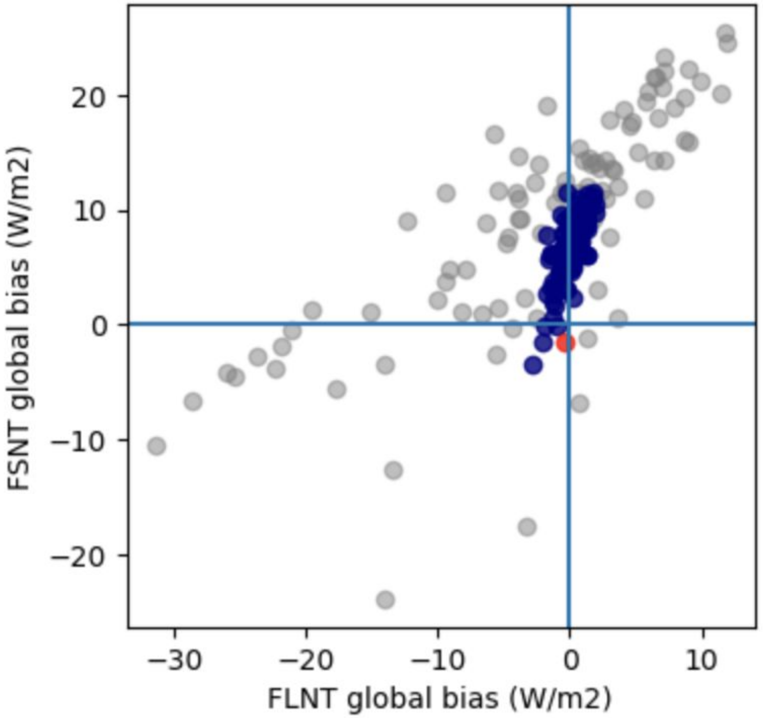
- Differences between the uncoupled and coupled runs

Selected coupled run results



Biases from uncoupled simulations





#271, the reference at the time of this analysis

Constrained parameters

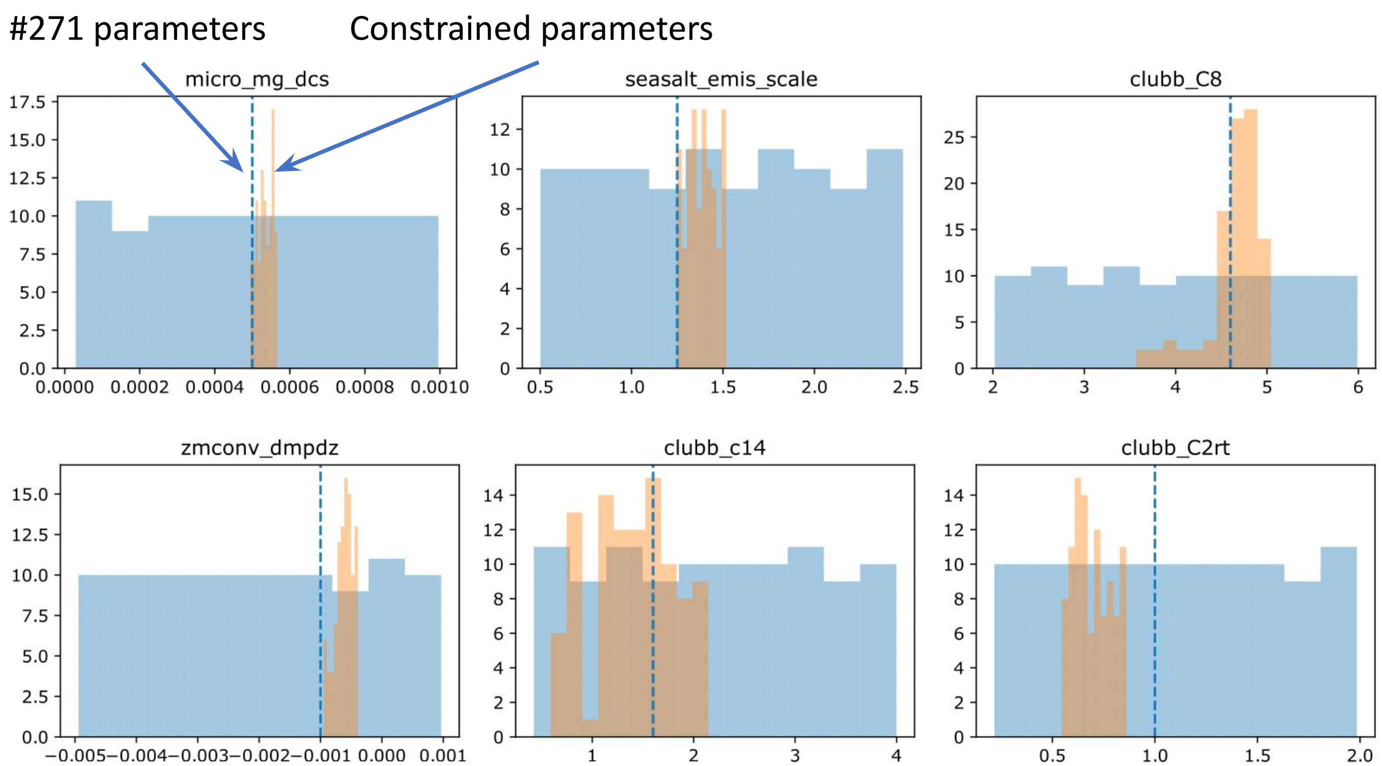
The difference is from the choice or what is considered more important

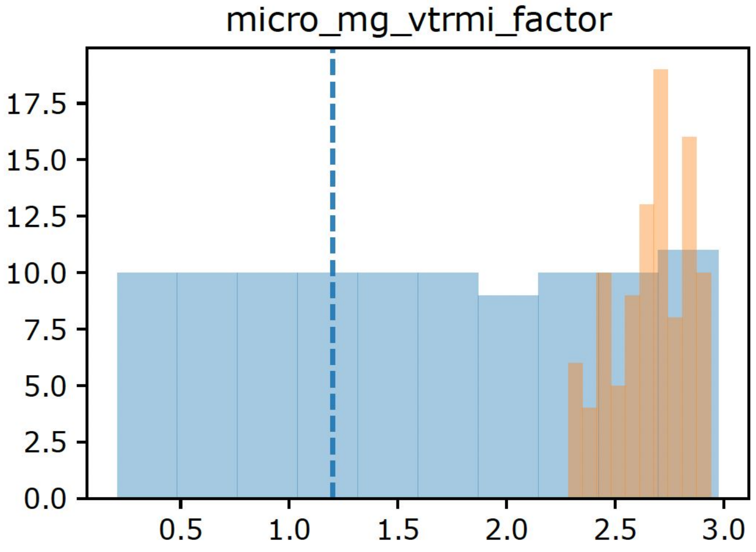
#271 parameters:

Care little about LWCF;

Our case:

We decide to drop a mixture of zonal LWCF and FLNT hoping to improve the overall performance --> Does not work well when used for coupled simulations

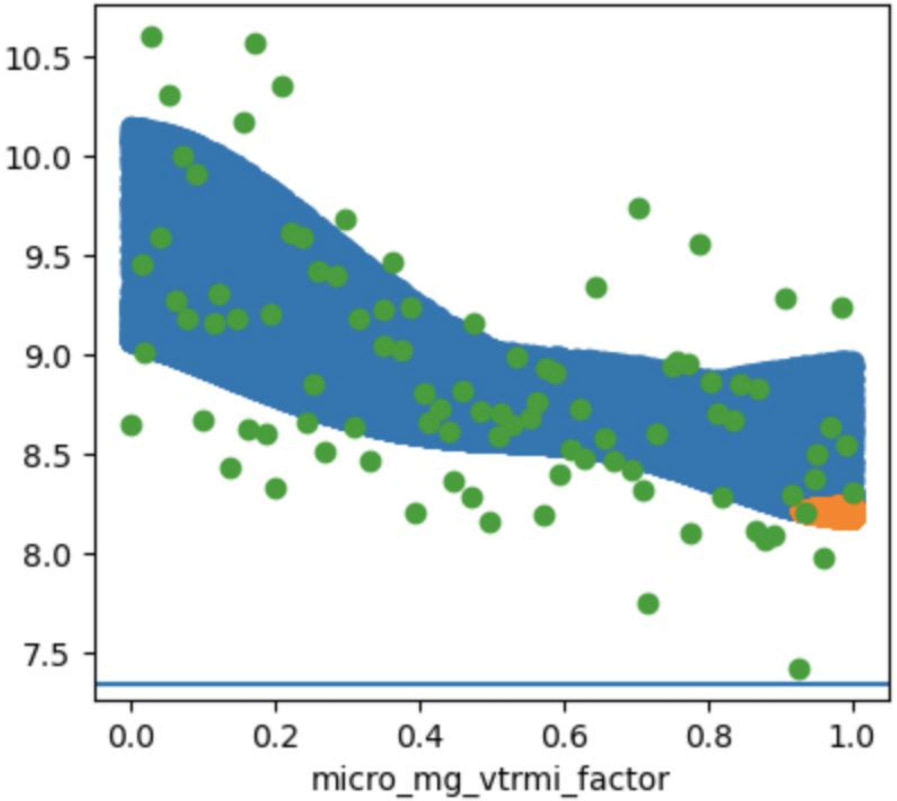





```
TMQ_zonal_25to35     : 0.00
TMQ_zonal_35to45     : 0.00
TMQ_zonal_45to55     : 0.29
TMQ_zonal_55to65     : 0.78
TMQ_zonal_65to75     : 0.93
LHFLX_zonal_-45to-35: 0.00
LHFLX_zonal_-25to-15: 0.00
LHFLX_zonal_-15to-5 : 0.00
LHFLX_zonal_-5to5   : 0.00
LHFLX_zonal_5to15   : 0.00
...
```

We can analyze what is causing the difference

We still accidentally consider TMQ_zonal_65to75 which should be discarded.



The difference is from the choice or what is considered more important

Conclusions and Lesson

- We present a new method for climate model calibration based on PPE datasets;
- It is designed to work in the presence of structural errors;
- It finds distinct parameter sets that lead to comparable or better model performance compared to the CAM default parameters in the uncoupled simulations;
- Challenges exist when this method is used for coupled simulations;

Lesson:

- **In between hand-tuning and auto-tuning, we might also need informed-tuning where we need to make a well-informed decision on what to prioritize for calibration.**

Thank you to everyone in AMWG!

qy2288@columbia.edu