Addressing the calibration bottleneck using machine learning: Application to the CNRM-CM6-1 model

Romain Roehrig

CNRM, Météo-France and CNRS, Toulouse, France

6th WGNE Workshop on Systematic Errors in Weather and Climate Models, 2 November 2022, Reading, UK

Climate model calibration

Climate model = a software
+ external forcings
+ a horizontal/vertical grid
+ a scientific content (e.g., parameterizations)
+ values for model internal/uncertain parameters >> ***calibration***

Calibration (or tuning)

- Common to most modelling frameworks
- Can be seen as an optimisation procedure under constraints (or ***metrics***), possibly with priorities/weights.
- Need for high-quality ***references/observations***
- +1 W m^{-2} at TOA ~ +0.5–1.5 K of global mean near-surface temperature (*Hourdin et al. 2017*).
    - ➢ *Given current uncertainties, present-day global-mean temperature in a climate model is mostly a result of tuning.*

A bottleneck for climate model development

- ***High dimensionality*** of the parameter space ~O(10)
- Climate model numerical simulations are ***computationally expensive***
    - ➢ An exhaustive exploration of the parameter space is not directly possible.
- Large number and variety of metrics O(10-100)
- ***Overfitting*** issue, treatment of ***uncertainties***

Calibration of CNRM-CM6-1 (*Voldoire et al. 2019, Roehrig et al. 2020*)

- Manual calibration, 1 or 2 parameters at the same time, mixing well-defined metrics and more subjective considerations
- Calibration of stand-alone components before coupling, priorities among metrics
- Often questioning the model physical content. But difficult to disentangle true model structural limits from “just” a poor calibration?

A rationale for addressing the calibration bottleneck

History matching

- Determine the plausible sets of model parameter values rather than optimize
- Or equivalently rule out the ***implausible*** sets of model parameter values
- For a given set of ***quantitative metrics***
- Considering ***reference/observations uncertainties***
- And introducing priors for ***model structural errors*** (interpreted at first as a tolerances to error)

Machine learning

- ***Emulate*** the model behaviour (i.e. the dependence of metrics to model parameters), to explore at very weak cost the parameter space.
- Consider also the ***emulators' uncertainty***
- ***Gaussian processes*** nicely provide predictions with an uncertainty estimate

Iterative refocussing

- Be as ***parsimonious*** as possible in terms of true simulations
- Start with a 'few' number of simulations and progressively add new ones to improve the emulator quality, but only where it is needed
- Possibly add new metrics along the way, based on more expansive simulations (pre-conditioning with cheaper configurations)

- ➢ *A formalized calibration procedure, transparent and reproducible.*
- ➢ *More rigorous comparison between parameterizations, quantifying more rapidly the true benefit of a new development.*
- ➢ *Possibly not a single acceptable configuration but several: being able to explore the model parametric uncertainty of its emergent properties.*

Williamson et al. (2015, 2017)

The technical framework

1. Define targeted (scalar) ***metrics*** f, their ***reference values*** r_f and associated ***uncertainties*** $\sigma_{r,f}$
2. Identify the relevant model ***parameters*** $\boldsymbol{\lambda}$, and their "acceptable" range >> ***input parameter space*** Λ
3. Define a simulation strategy, build an experimental design, run simulations >> learning dataset
4. ***Emulate*** $f(\boldsymbol{\lambda})$ for each metric (Gaussian Processes)
5. Identify the sub-space of Λ which is compatible with references for all metrics

   >> ***Not-Ruled-Out-Yet – NROY – space***

   considering
   - The reference uncertainty
   - The emulator uncertainty
   - The model structural error (tolerance to error) $\sigma_{d,f}$

   >> ***Implausibility*** measure I_f, ***cutoff*** T

$$I_f(\boldsymbol{\lambda}) = \frac{|r_f - \mathrm{E}\left[f(\boldsymbol{\lambda})\right]|}{\sqrt{\sigma_{r,f}^2 + \sigma_{d,f}^2 + \mathrm{Var}\left[f(\boldsymbol{\lambda})\right]}}.$$

$$\mathrm{NROY}_f^1 = \{\boldsymbol{\lambda} \mid I_f(\boldsymbol{\lambda}) < T\}$$

$$\mathrm{NROY}^1 = \bigcap_f \mathrm{NROY}_f^1 = \{\lambda \mid I_f(\lambda) < T, \text{for all } f\}$$

Based on HTExplo (Couvreux et al. 2020)

Williamson et al. (2015, 2017)

The technical framework

1. Define targeted (scalar) ***metrics*** f, their ***reference values*** r_f and associated ***uncertainties*** $\sigma_{r,f}$
2. Identify the relevant model ***parameters*** $\boldsymbol{\lambda}$, and their "acceptable" range >> ***input parameter space*** Λ
3. Define a simulation strategy, build an experimental design, run simulations >> learning dataset
4. ***Emulate*** $f(\boldsymbol{\lambda})$ for each metric (Gaussian Processes)
5. Identify the sub-space of Λ which is compatible with references for all metrics

   >> ***Not-Ruled-Out-Yet – NROY – space***

   considering

   - The reference uncertainty
   - The emulator uncertainty
   - The model structural error (tolerance to error) $\sigma_{d,f}$

   >> ***Implausibility*** measure I_f, ***cutoff*** T

$$I_f(\boldsymbol{\lambda}) = \frac{|r_f - \mathrm{E}\left[f(\boldsymbol{\lambda})\right]|}{\sqrt{\sigma_{r,f}^2 + \sigma_{d,f}^2 + \mathrm{Var}\left[f(\boldsymbol{\lambda})\right]}}.$$

$$\mathrm{NROY}_f^1 = \{\boldsymbol{\lambda} \mid I_f(\boldsymbol{\lambda}) < T\}$$

$$\mathrm{NROY}^1 = \bigcap_f \mathrm{NROY}_f^1 = \{\lambda \mid I_f(\lambda) < T, \text{for all } f\}$$

6. Iterate over several waves to reduce the emulators' uncertainty in NROY^{N-1}, until convergence

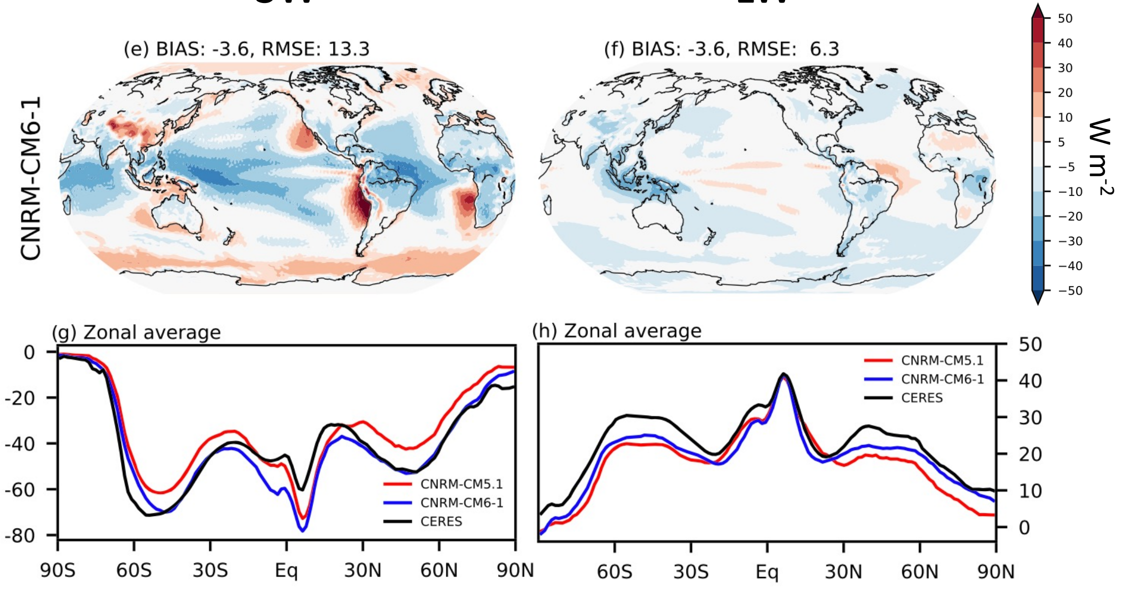
Based on HTExplo (Couvreux et al. 2020)

Williamson et al. (2015, 2017)

The starting configuration: CNRM-CM6-1

CRE at TOA

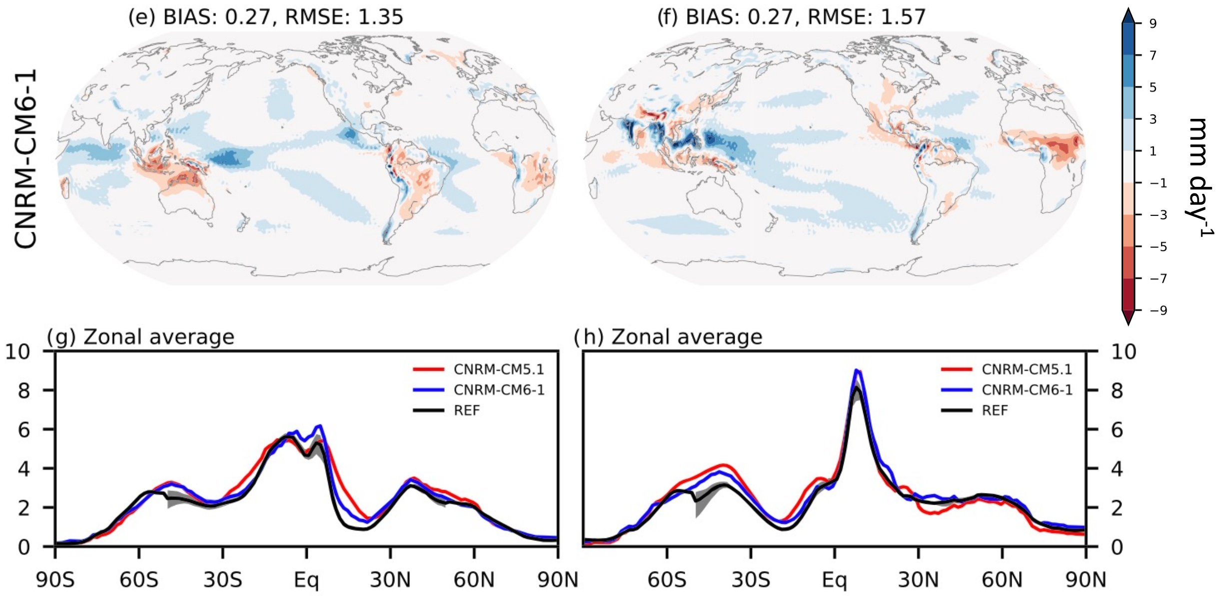
SW LW



Reference: CERES-EBAF Ed. 4.1

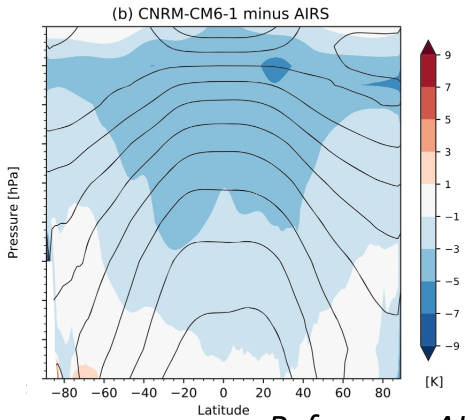
Precipitation

DJFM JJAS



Reference: MSWEP v1.2 + GPCP v2.3 + TRMM 3B42 v7

Annual Temperature



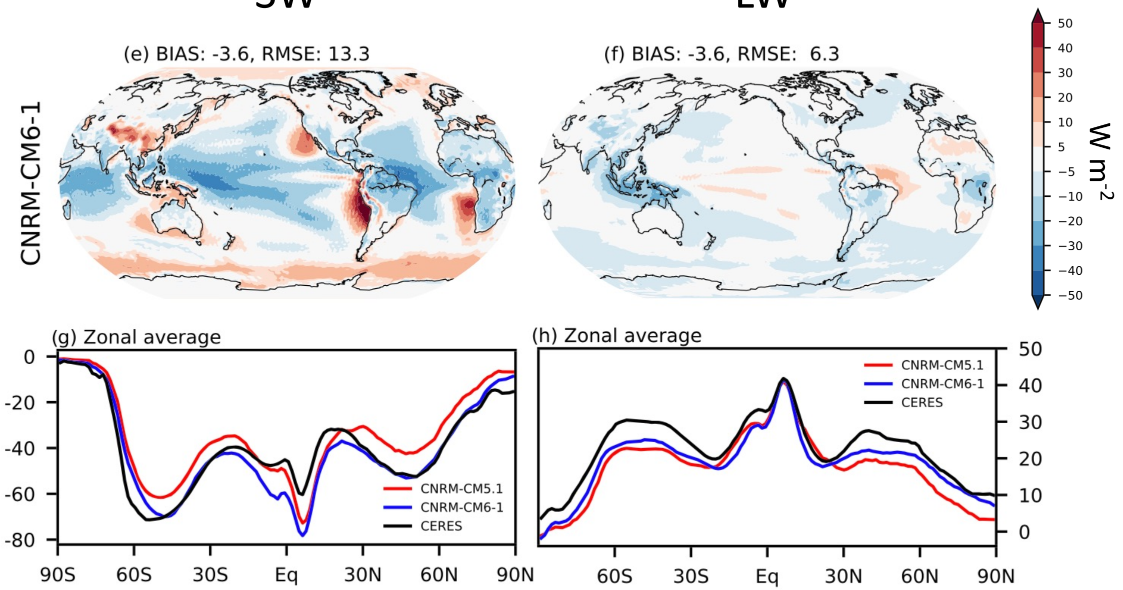
Reference: AIRS

Roehrig et al. (2020)

The starting configuration: CNRM-CM6-1

CRE at TOA

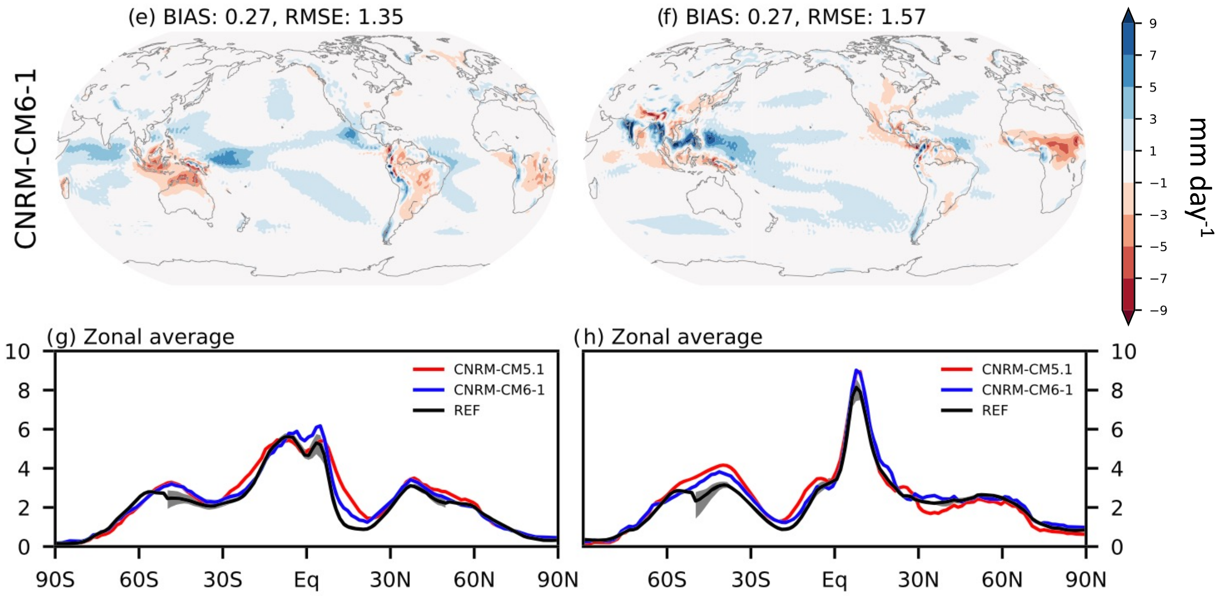
SW LW



Reference: CERES-EBAF Ed. 4.1

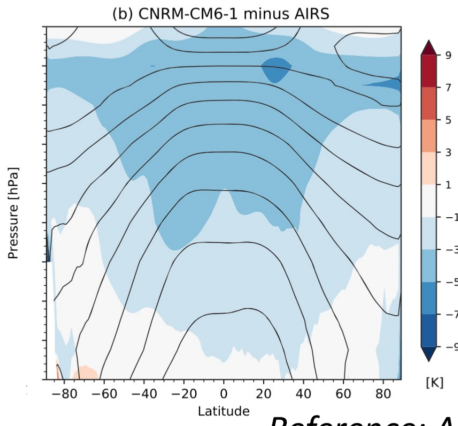
Precipitation

DJFM JJAS



Reference: MSWEP v1.2 + GPCP v2.3 + TRMM 3B42 v7

Annual Temperature



Reference: AIRS

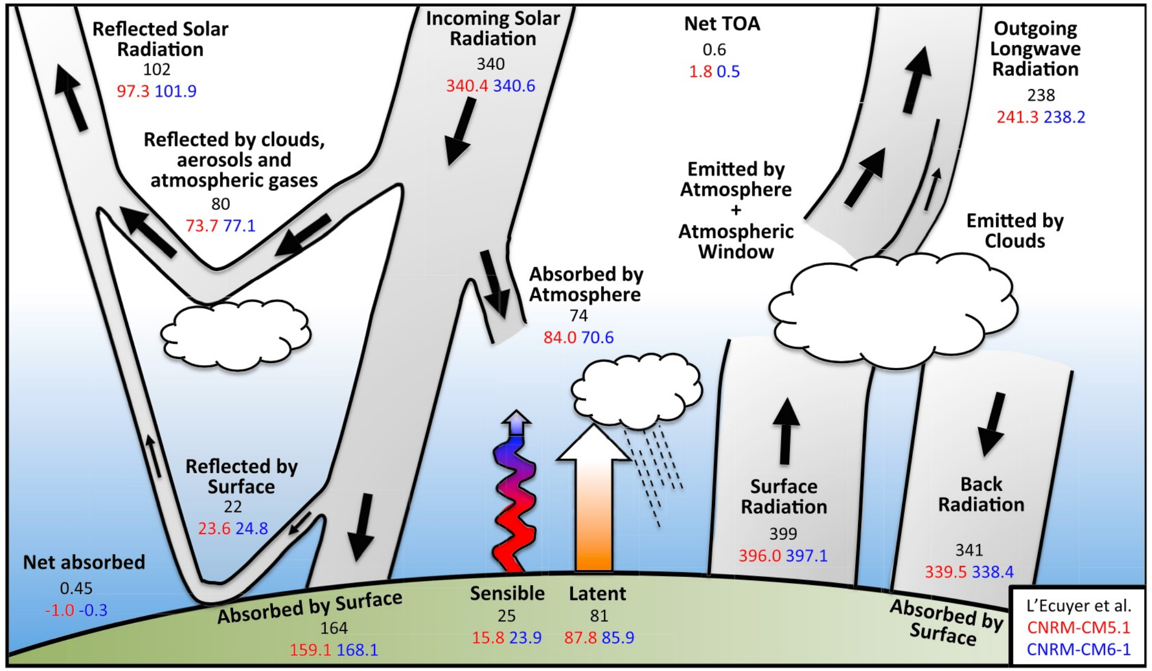
➢ ***Can we reduce CNRM-CM6-1 biases through better calibration?***

Roehrig et al. (2020)

Metrics, references and uncertainty estimates

Following the CNRM-CM6-1 tuning strategy, 3 classes of metrics:

1. ***Global averages of the energy budget components***
    - at TOA: OLR, OSR, Net, SW/LW CRE
    - at the surface ocean: Net, SWdn, LWdn
    - Values from CERES-EBAF, uncertainties based on the literature
    - Except Net at surface/TOA = 0 +/- 0.1 W m^{-2} : the model has to be equilibrated.
    - Tolerance to error: 0.5 W m^{-2}



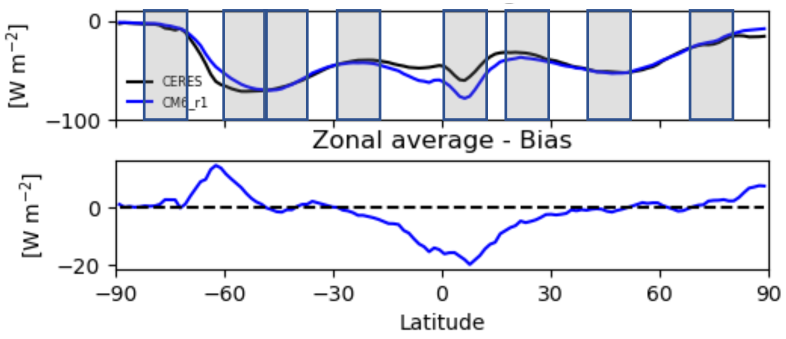
Roehrig et al. (2020)

Metrics, references and uncertainty estimates

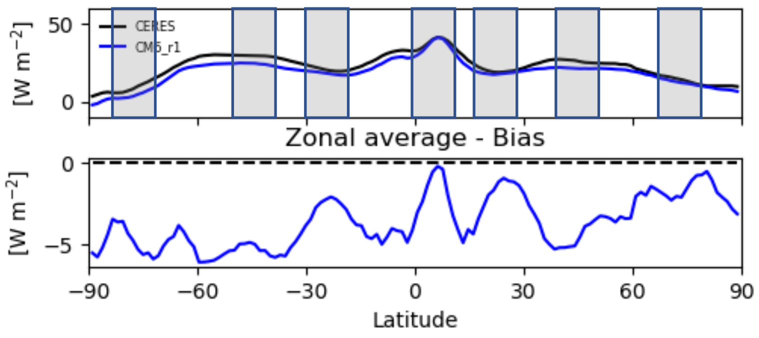
Following the CNRM-CM6-1 tuning strategy, 3 classes of metrics:

1. Global averages of the energy budget components
2. ***Zonally average profiles*** of SW/LW CRE + Temperature at 200 hPa
   - SW/LW CRE: CERES-EBAF with uncertainty of 2 W m^{-2} + tolerance of 1 W m^{-2}
   - T200: based on ERA5/JRA55/MERRA/CFSR ensemble mean and std, tolerance 1.5 K

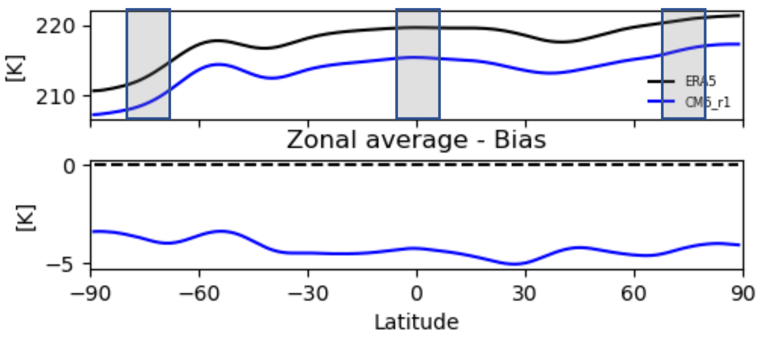
SW CRE



LW CRE



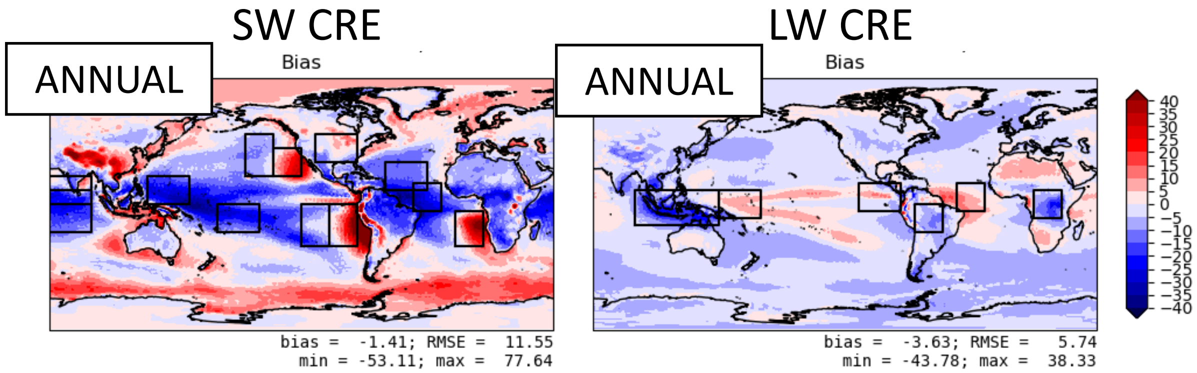
T at 200 hPa

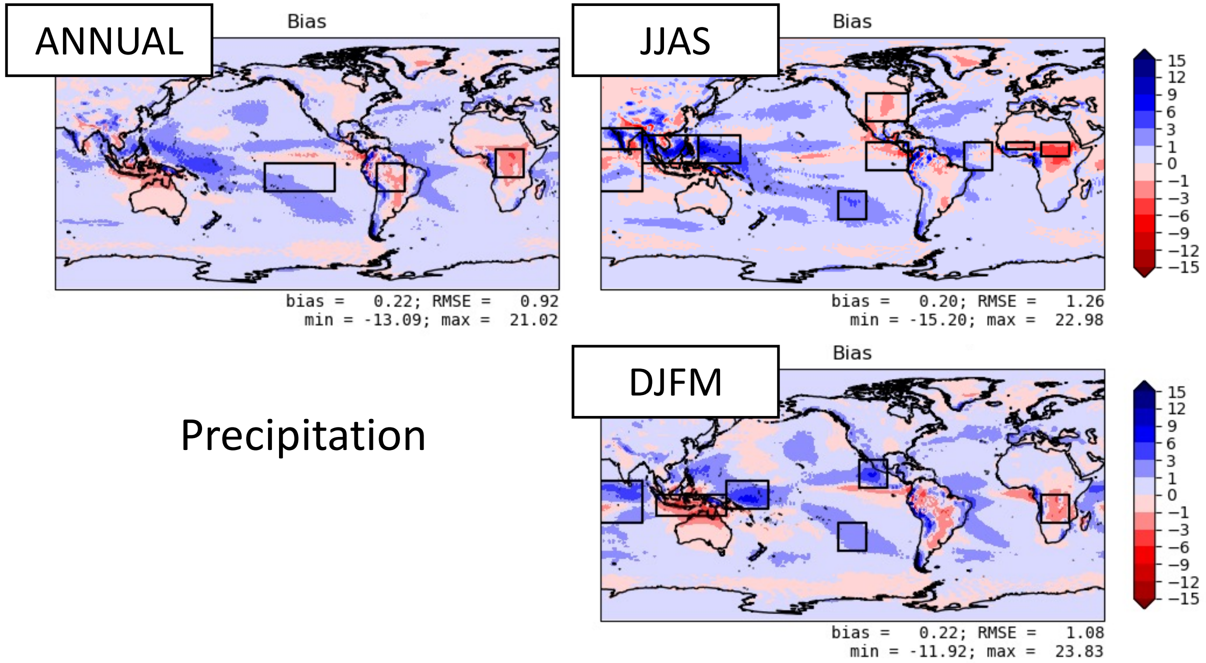


Metrics, references and uncertainty estimates

Following the CNRM-CM6-1 tuning strategy, 3 classes of metrics:

1. Global averages of the energy budget components
2. Zonally average profiles of SW/LW CRE + Temperature at 200 hPa
3. ***Regional and seasonal averages*** of SW/LW CRE and precipitation
   - SW/LW CRE: CERES-EBAF, uncertainty of 2 W m^{-2}, tolerance of 5 W m^{-2}
   - Precipitation: MSWEP/GPCP/TRMM 3B42 ensemble mean and std, tolerance between 0.5 and 1 mm day^{-1}





Metrics, references and uncertainty estimates

Following the CNRM-CM6-1 tuning strategy, 3 classes of metrics:

1. Global averages of the energy budget components
2. Zonally average profiles of SW/LW CRE + Temperature at 200 hPa
3. Regional and seasonal averages of SW/LW CRE and precipitation

- ***63 (scalar) metrics***

Model parameters and simulation strategy

46 model parameters

- 7 from turbulence (TKE scheme + PBL-top entrainment)
- 16 from microphysics (1-moment, 5 hydrometeors)
- 19 from the unified dry, shallow and deep convection scheme
- 4 from cloud radiative properties (heterogeneity)

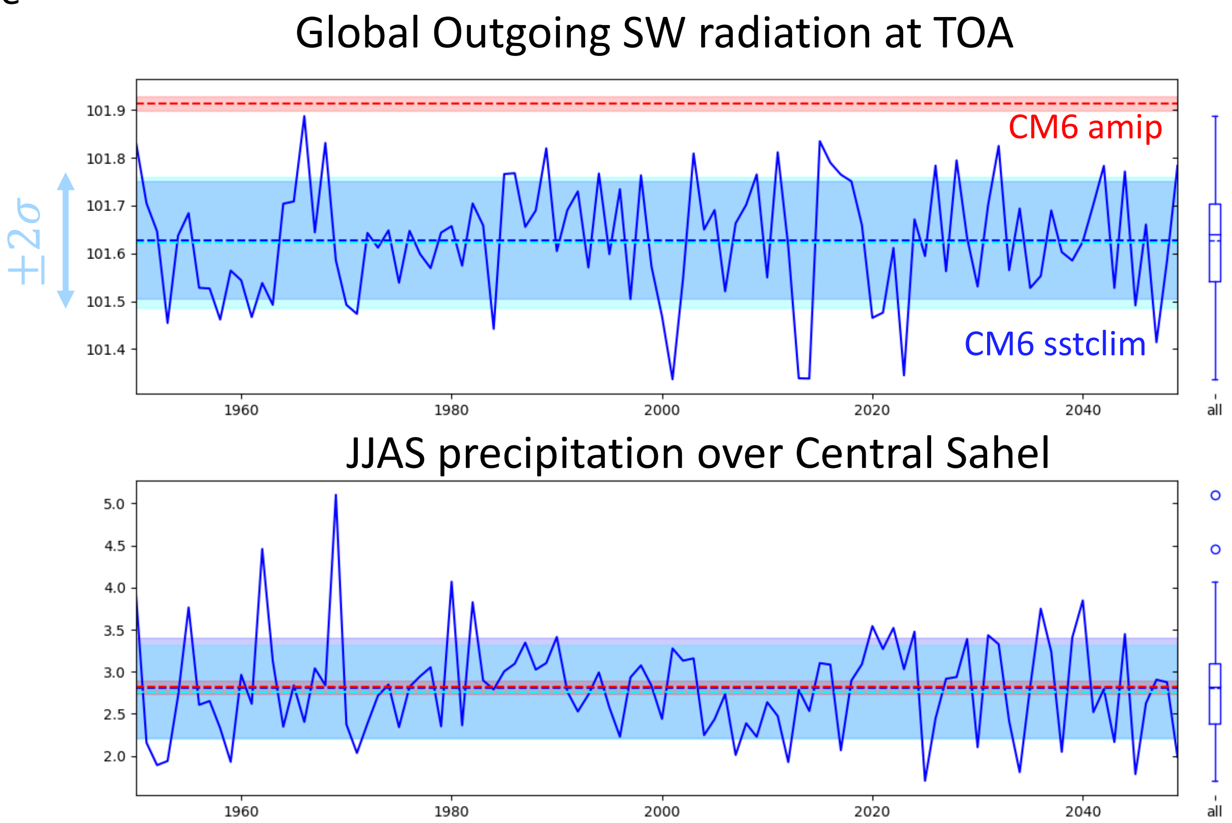
Model parameters and simulation strategy

46 model parameters

- 7 from turbulence (TKE scheme + PBL-top entrainment)
- 16 from microphysics (1-moment, 5 hydrometeors)
- 19 from the unified dry, shallow and deep convection scheme
- 4 from cloud radiative properties (heterogeneity)

Waves of 400 simulations

- 1-year ***sstclim*** simulations + 3-month spin-up
- sstclim vs amip correction of the reference target
- Consideration of ***internal variability uncertainty*** based on a 100-year sstclim simulation with CNRM-CM6-1.
- Latin Hypercube sampling for 1st wave



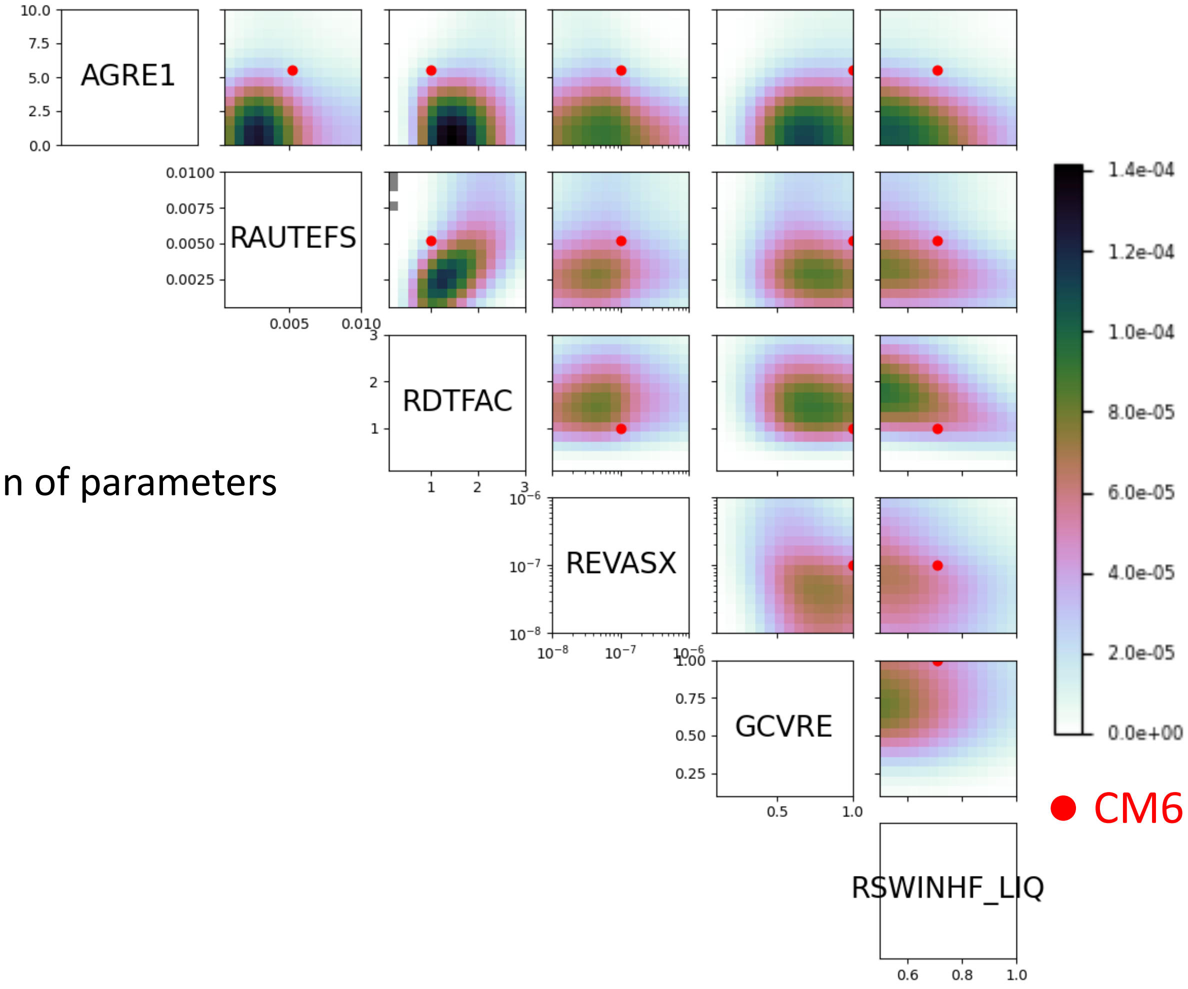
Wave 1 results

- NROY[1] space: 0.66% of the input space
- ***Numerical characterization*** of the NROY space:
  - (very) large sampling (LHS) of the input parameter space
  - Use emulators to compute associated implausibilities
  - Compute densities of points within NROY space as a function of parameters
  - 1D or 2D representations

Wave 1 results

- NROY[1] space: 0.66% of the input space
- ***Numerical characterization*** of the NROY space:
  - (very) large sampling (LHS) of the input parameter space
  - Use emulators to compute associated implausibilities
  - Compute densities of points within NROY space as a function of parameters
  - 1D or 2D representations

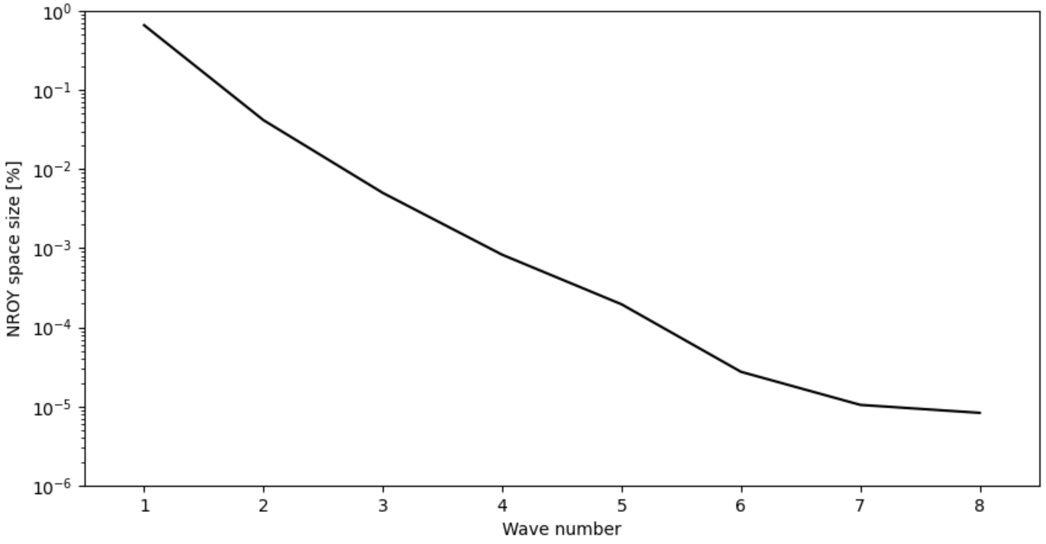
NROY[1] density within input parameter space
For some of the dominant parameters



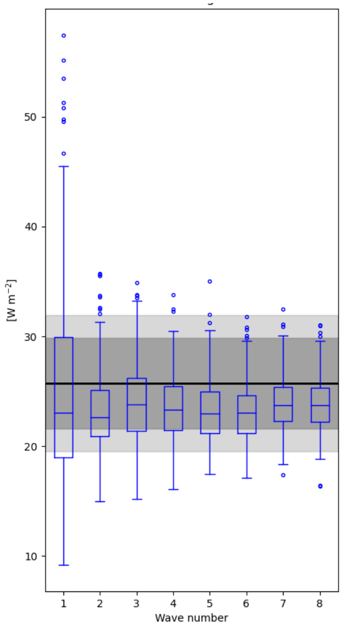
From Wave 1 to Wave 8

- Strong reduction of the NROY space size (8 orders of magnitude)
- Some metrics have converged, some are more demanding

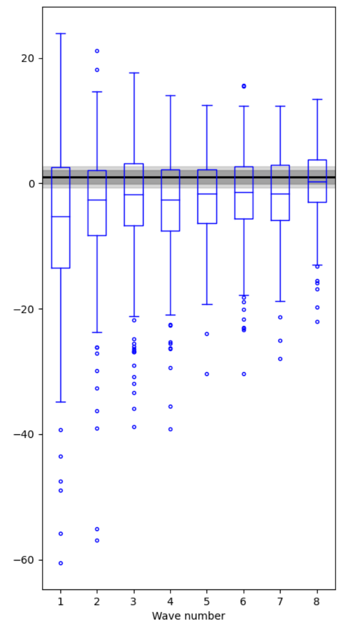
NROY space fraction of the input space



Global LW CRE at TOA

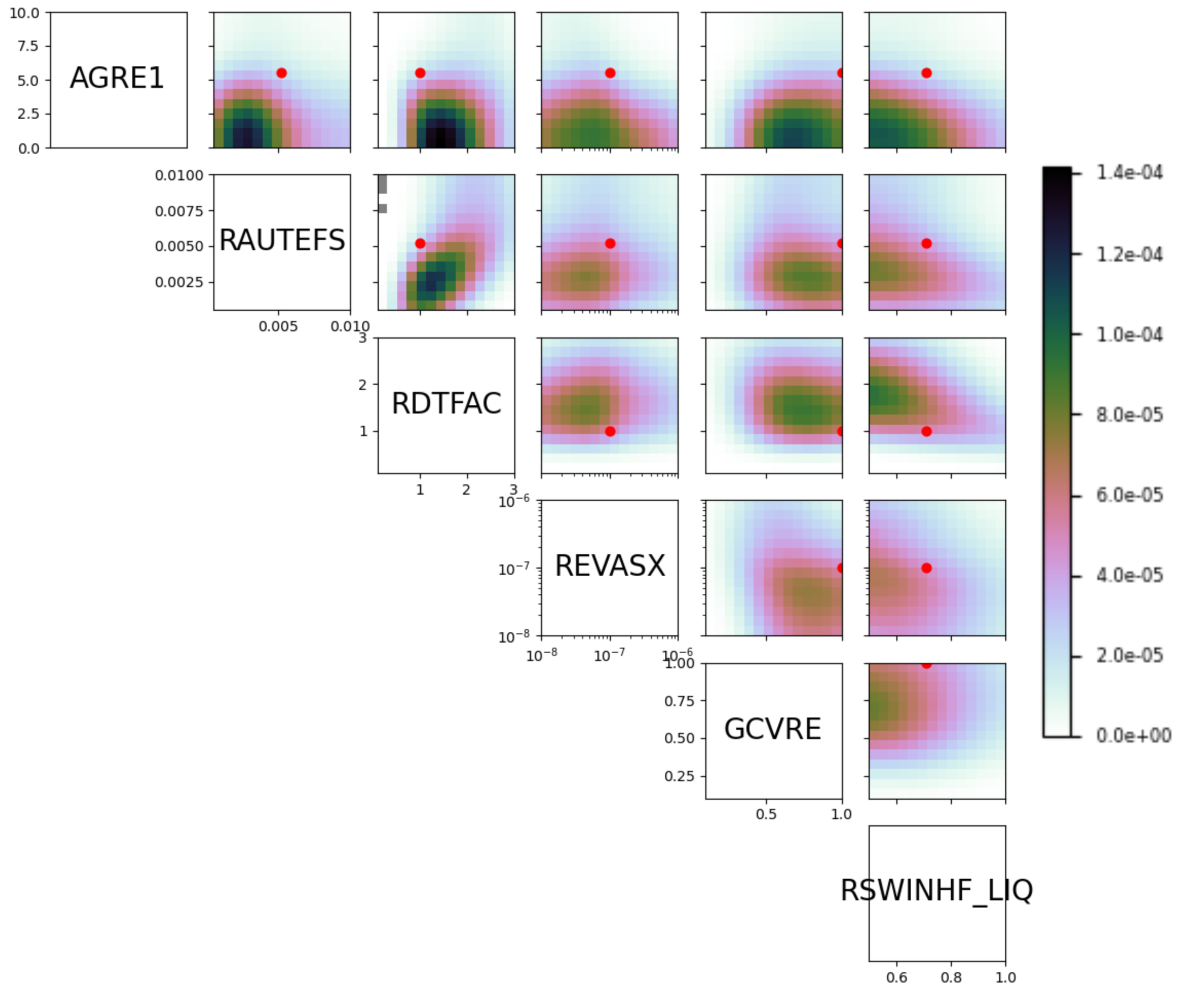


Ocean net energy flux

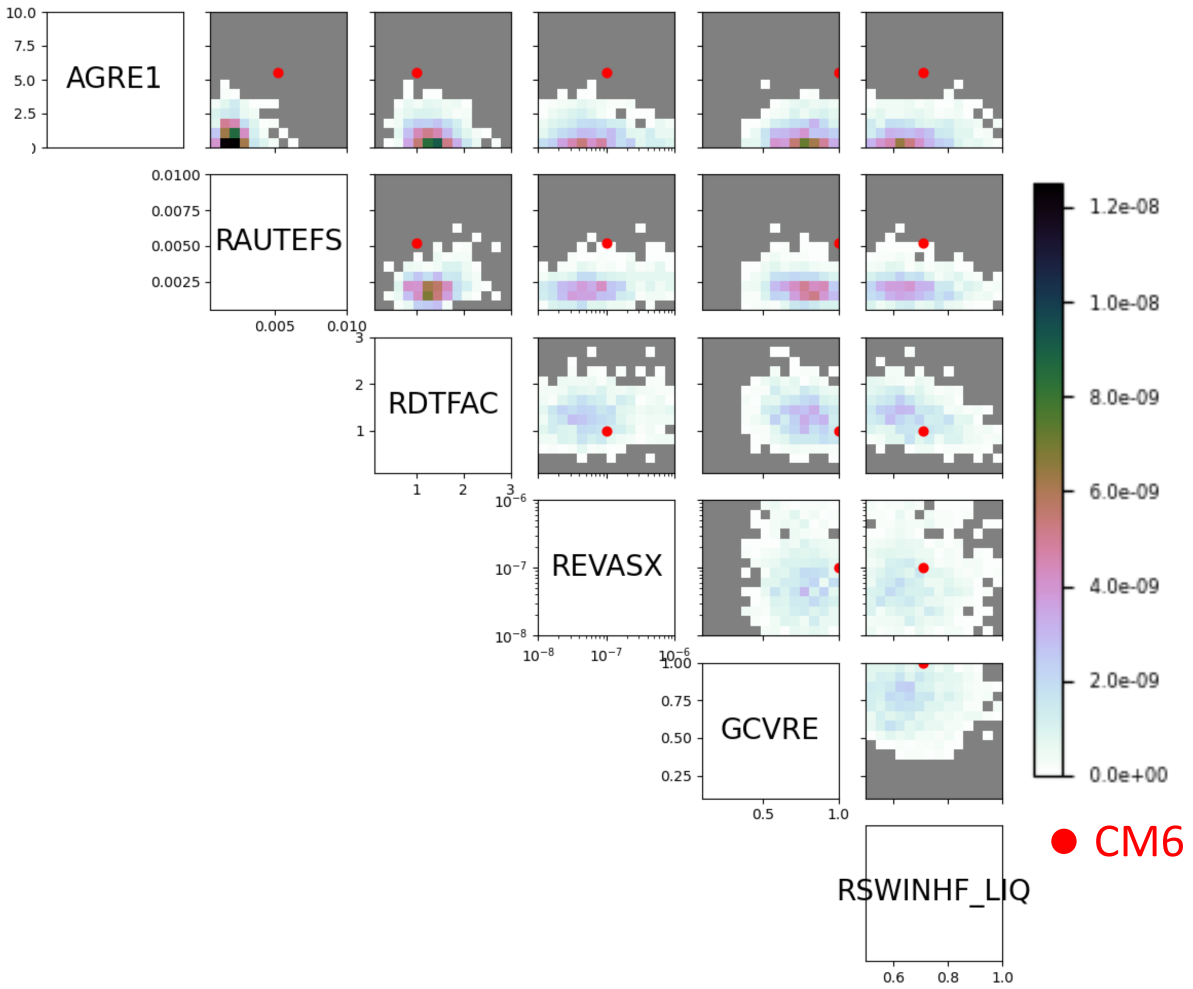


From Wave 1 to Wave 8

NROY[1] density within input parameter space
For some of the dominant parameters



NROY[8] density within input parameter space
For some of the dominant parameters



+ some new dominant parameters have emerged

Choosing configurations of interest

Unfortunately, none of Wave 1-8 simulations fulfils all the metrics

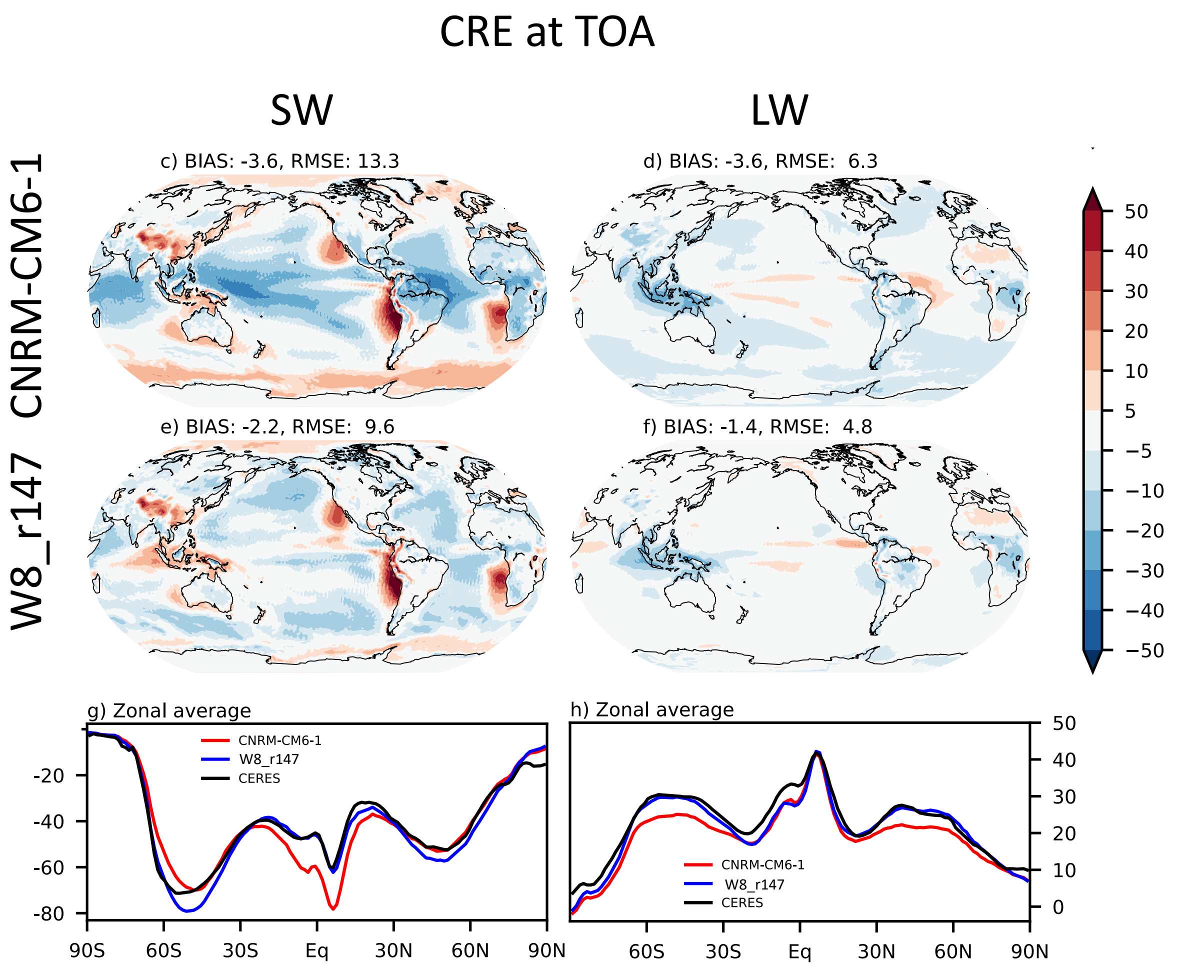
- Convergence is not yet achieved
- Appropriate sampling of small NROY spaces is difficult and requires further work
- Some tolerances to error are likely too weak and require to be revisited.

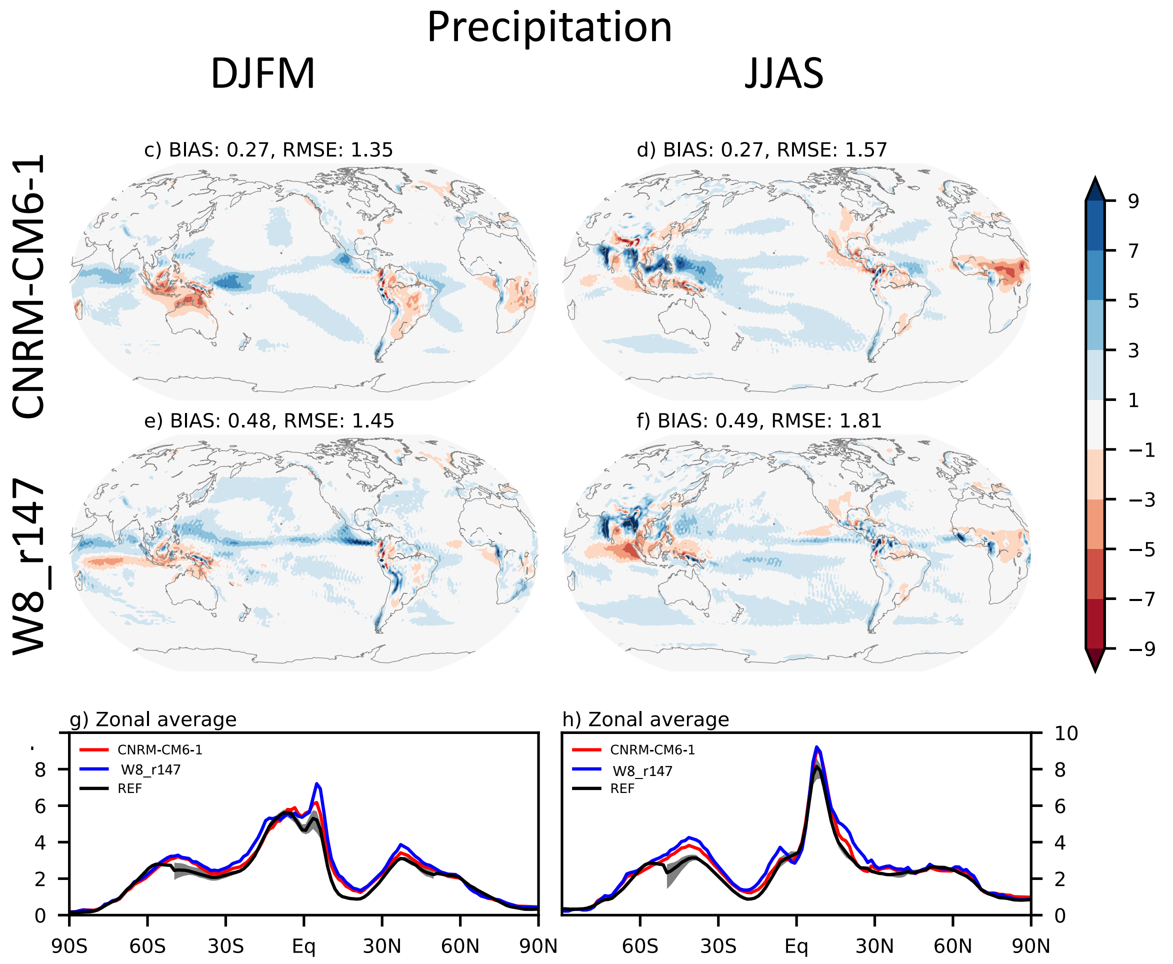
Nevertheless

- A few simulations fulfil all the metrics but one (for a cutoff of 3)
- A few have interestingly low RMSEs for targeted variables

➢ A selection of these simulations is further analysed with amip-style simulations (10 years).

Improved performance?





- ***Improved or similar performance*** on several mean state features
- Some errors seems truly structural: clouds/radiation over eastern part of ocean basins

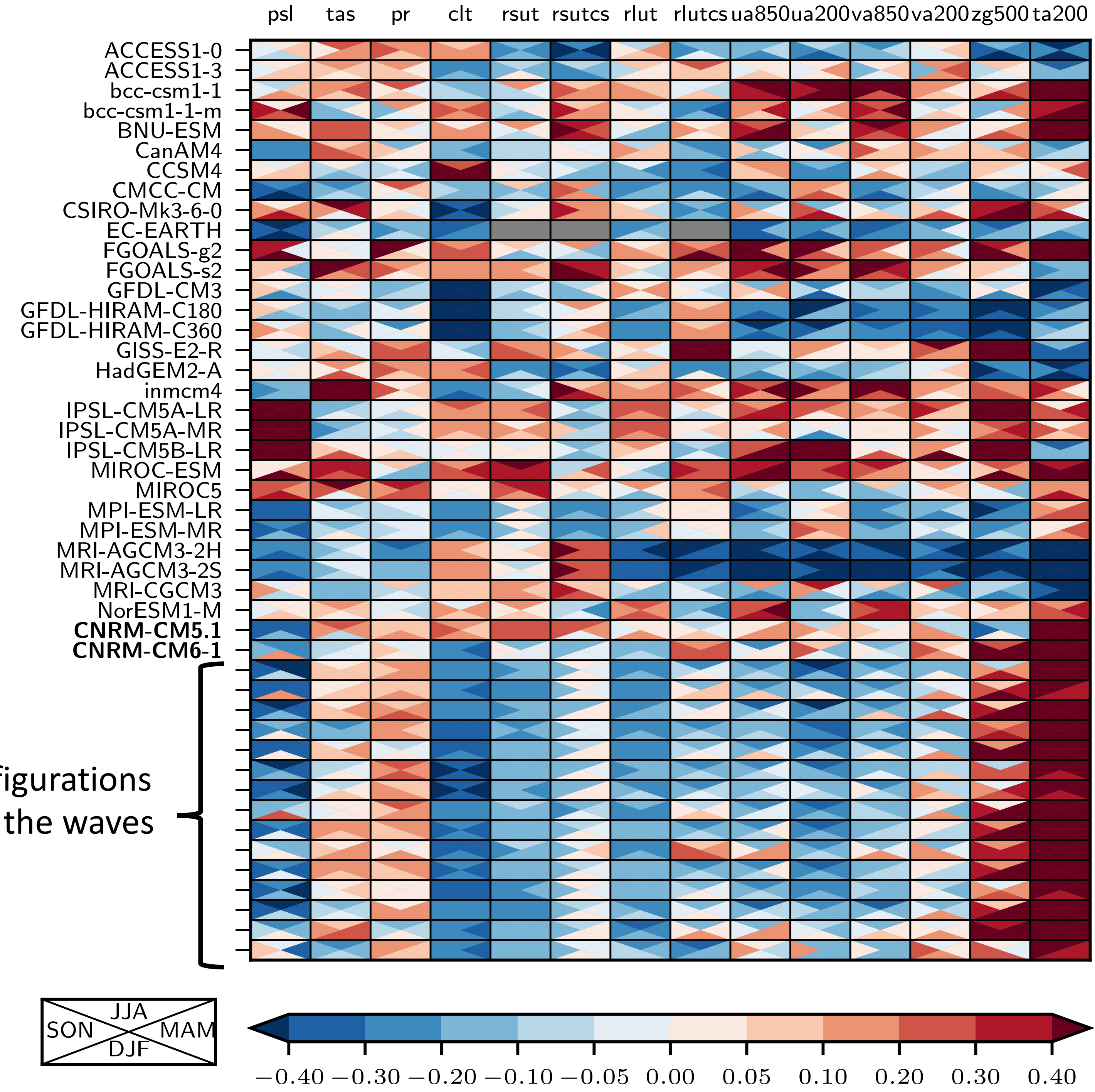
Improved performance?

Relative score within the CMIP5 ensemble

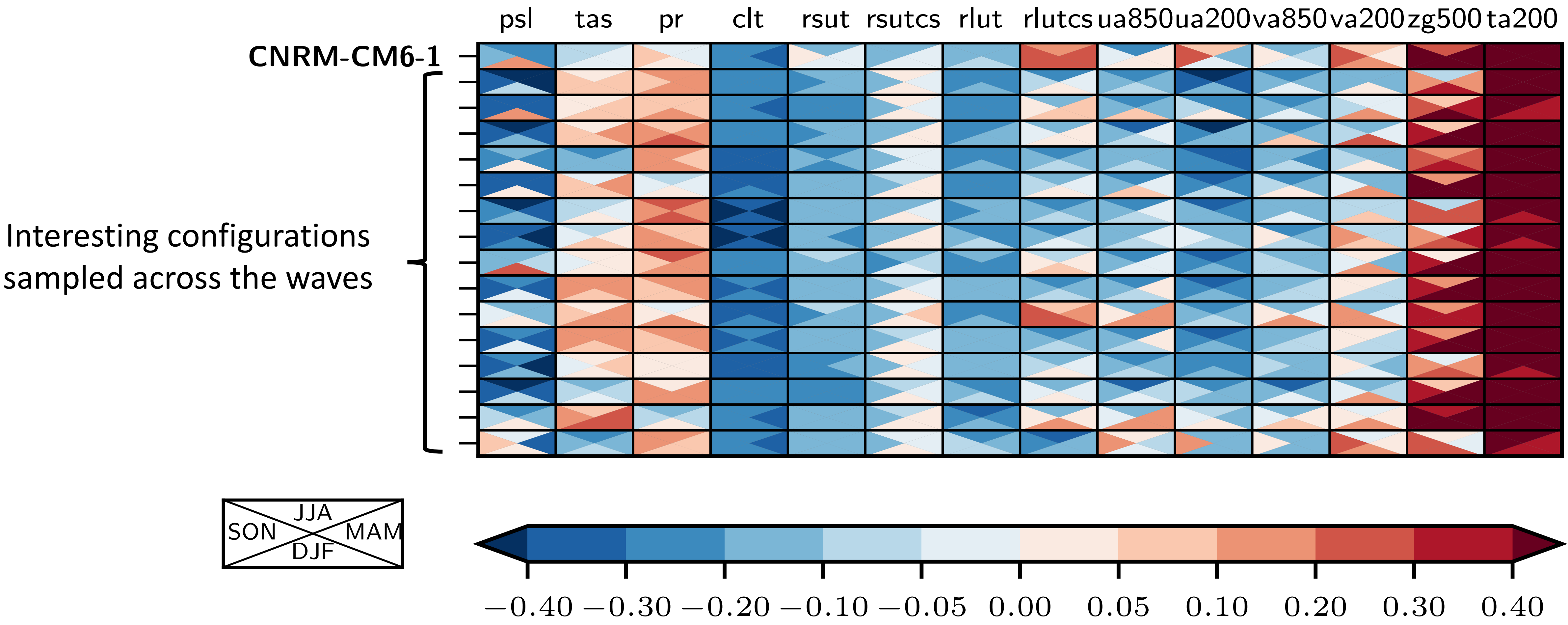
(Gleckler et al. 2016)

$$\text{score} = \frac{\text{RMSE} - \text{RMSE}_{\text{median}}}{\text{RMSE}_{\text{median}}}$$

Interesting configurations sampled across the waves



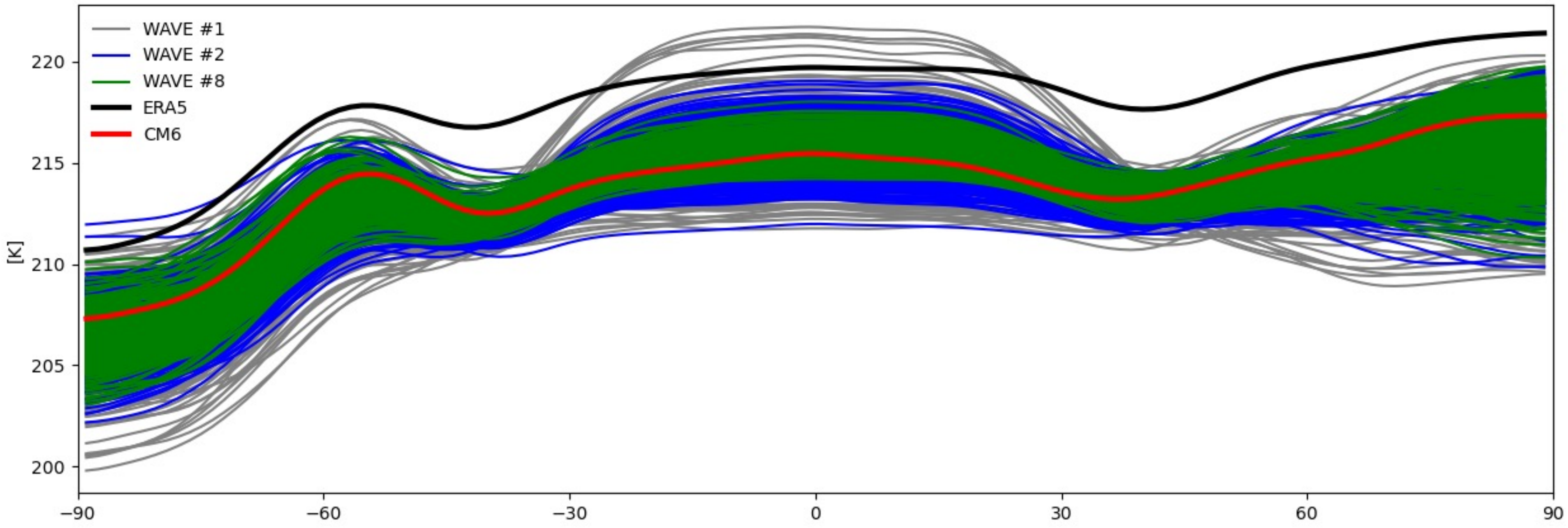
Improved performance?



- ***Improved or similar performance*** on several mean state features
- Some errors seems truly structural: clouds/radiation over eastern part of ocean basins, upper-tropospheric temperature
- Some ***trade-offs*** are required

Upper-tropospheric temperature bias: structural limit?

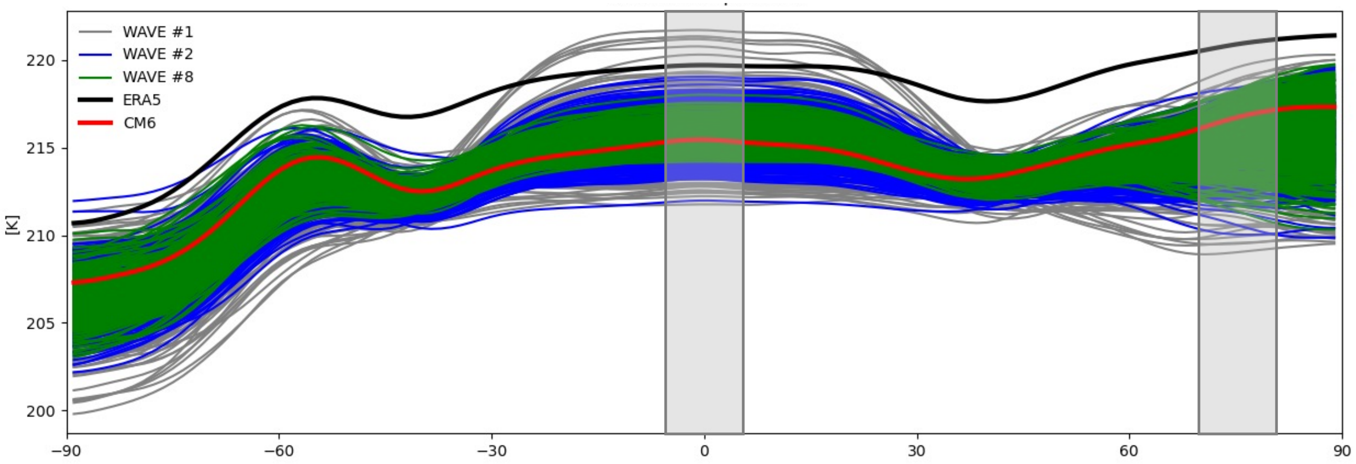
Zonally-average 200-hPa Temperature

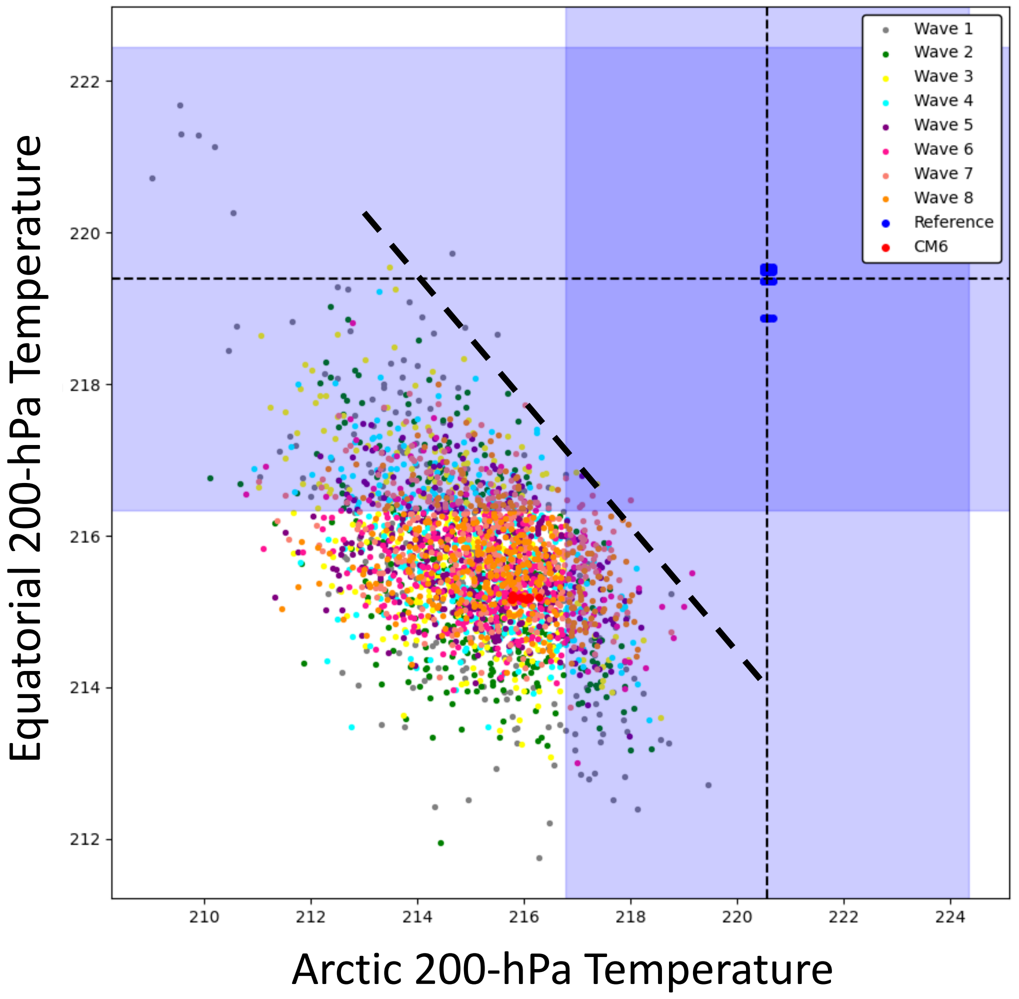


- Simulations performing well (beyond the chosen tolerance to error) in the equatorial regions disappear with successive waves
  - ➢ Incompatible with other metrics

Upper-tropospheric temperature bias: structural limit?

Zonally-average 200-hPa Temperature





- Simulations performing well (beyond the chosen tolerance to error) in the equatorial regions disappear with successive waves
    - ➢ Incompatible with other metrics
- The model can most likely not capture both equatorial and arctic upper-tropospheric temperatures within uncertainty ranges, while remaining compatible with other metrics.

Conclusions and next steps

History matching with iterative refocussing

- Provides a ***relevant and efficient framework for model calibration in the presence of uncertainties***
- Can help accelerate model development by comparing calibrated model version
    - ➢ *Assessing the true added value of a new development*
- Can help better ***identify model structural errors***, and thereby help focus bias understanding/model development

Next steps

- Play with tolerances to error to better identify/quantify model structural errors and trade-offs to be made
- Add new metrics (e.g., variability)
- Pre-conditioning with cheaper model configurations
  e.g., 1D/LES for preserving process-level performance *(Couvreux et al. 2020, Hourdin et al. 2020)*.
- Towards calibration of ocean-atmosphere coupled configurations:
  accelerating spin-up, use of intermediate resolutions, fast/slow processes…
- Develop physical interpretations of what is happening in the calibration process?

More on the technical/statistical aspects

- Going beyond scalar metrics, emulate directly vectors/maps (using EOFs, *Salter et al. 2019*)
- Develop strategies to better sample particularly small NROY spaces.