# Calibrating ensembles for model independence

Gab Abramowitz

Nadja Herger

UNSW Sydney /

ARC Centre of Excellence for Climate Systems Science

ARC CENTRE OF EXCELLENCE FOR CLIMATE SYSTEM SCIENCE

UNSW AUSTRALIA

# Structure

- Context
- Epistemic vs aleatory uncertainty
- Ensemble interpretation paradigms
- Why weighting / sub-selecting for dependence and/or performance is a calibration exercise
- Should calibration be application-specific or holistic?

# Framing dependence

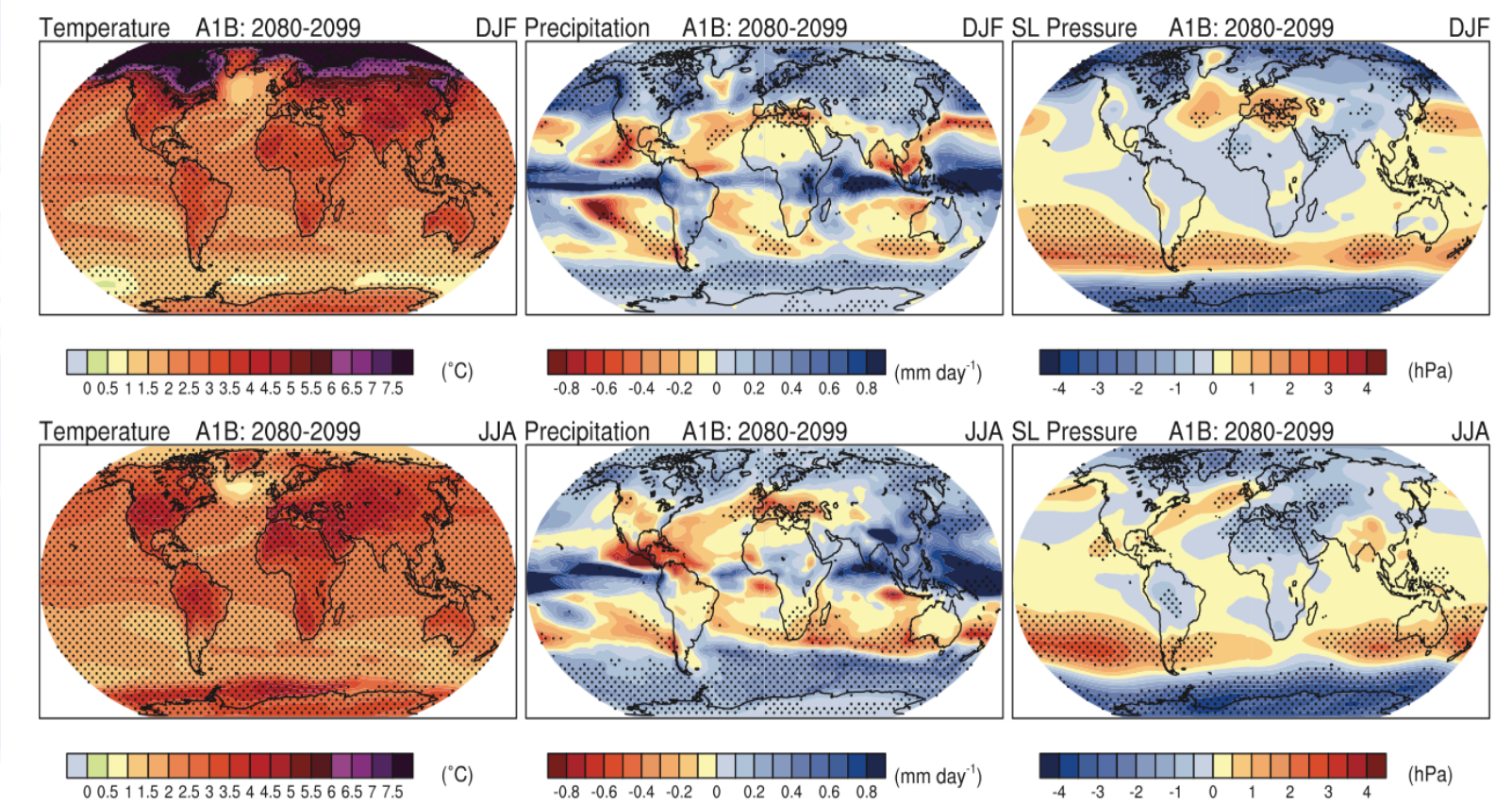
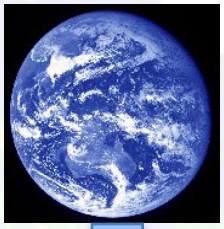


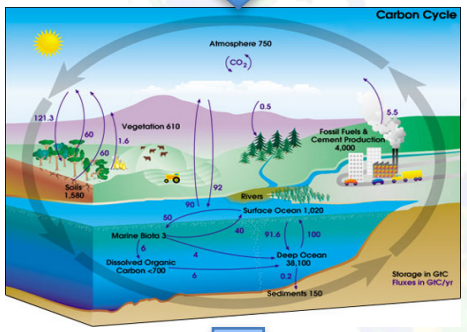

**Figure 10.9.** *Multi-model mean changes in surface air temperature (°C, left), precipitation (mm day$^{-1}$, middle) and sea level pressure (hPa, right) for boreal winter (DJF, top) and summer (JJA, bottom). Changes are given for the SRES A1B scenario, for the period 2080 to 2099 relative to 1980 to 1999. Stippling denotes areas where the magnitude of the multi-model ensemble mean exceeds the inter-model standard deviation. Results for individual models can be seen in the Supplementary Material for this chapter.*

Is agreement a sign of robustness?

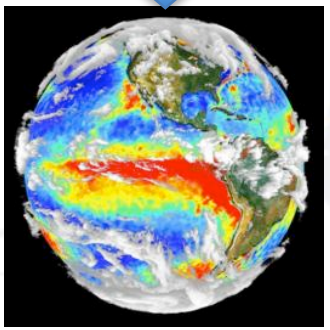
# Framing dependence: choices in model development

- PERCEPTUAL MODEL – identify features of the system
- CONCEPTUAL MODEL – identify relationships between features/processes in the perceptual model
- MATHEMATICAL/SYMBOLIC MODEL – identify equations that describe the conceptual model
- NUMERICAL MODEL – codification of equation solutions, spatial and temporal aggregation choices; implementation on a computer system.

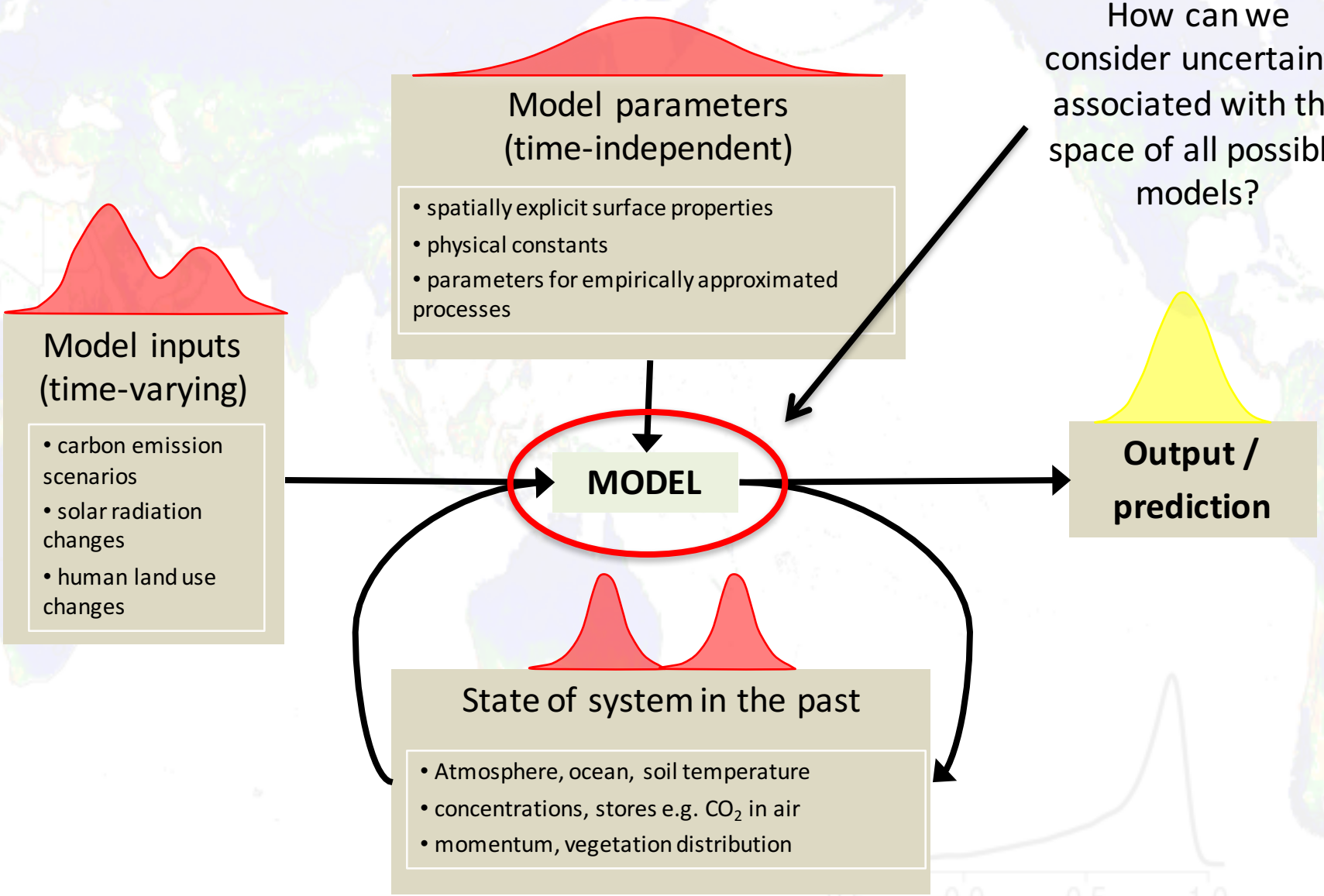





$$\frac{\partial(\eta_{Asat}\,\eta_{lf})}{\partial t} = \frac{\partial}{\partial z}\left( K_s \psi_s b\, \eta_{lf}^{b+2} \frac{\partial \eta_{lf}}{\partial z} - K_s \eta_{lf}^{2b+3} \right) + r(z).$$

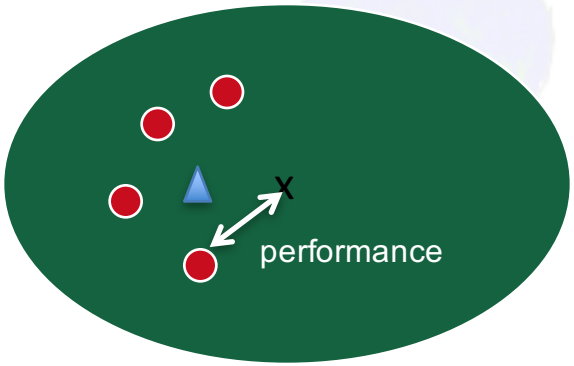


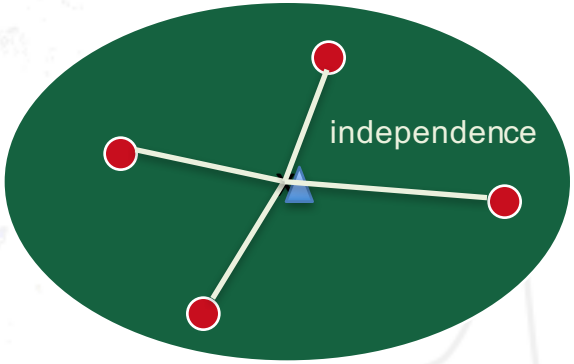
# Framing dependence: propagating uncertainty



# Framing dependence

- Modelling groups share literature, data sets, parametrisations, even model code – do different climate models constitute independent estimates of a prediction problem?
- Each RCP projection in CMIP5 has a different set of contributions – are we really just comparing differences between RCPs?
- There are many ways to define dependence
  - Different research group => independence? Model structure/shared parametrisation (evolutionary cladistics)
  - effect on ensemble performance (Linnaean taxonomy) – more like what we care about
  - Hilltop estimation analogy of mean estimate



# Structure

- Context
- Epistemic vs aleatory uncertainty
- Ensemble interpretation paradigms
- Why weighting / sub-selecting for dependence and/or performance is a calibration exercise
- Should calibration be application-specific or holistic?

# Dividing multi-model ensemble spread in two: epistemic and aleatory uncertainty

- **Epistemic:** uncertainty in *our* ability to model the system, a property of *us*, that we might hope to transcend as we learn more
  - Potentially predictable, uncertainty that could possibly be resolved
  - Ensemble spread that can be attributed to differences in models

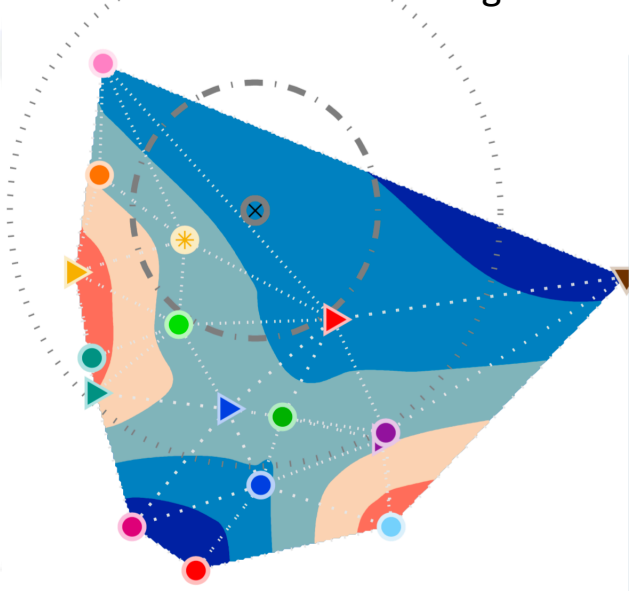
- **Aleatory:** either:
  - Uncertainty in the system itself, given the amount of information we have (e.g. observations of initial conditions states, at a particular spatial and temporal resolution), or
  - Uncertainty in the system itself (i.e. perfect measurement)
  - Distinction between the two is theoretical, practically it makes no difference
  - "internal variability", any evaluation that involves time series, or shorter term averages
  - Ensemble spread that is irreducible, fundamentally unpredictable

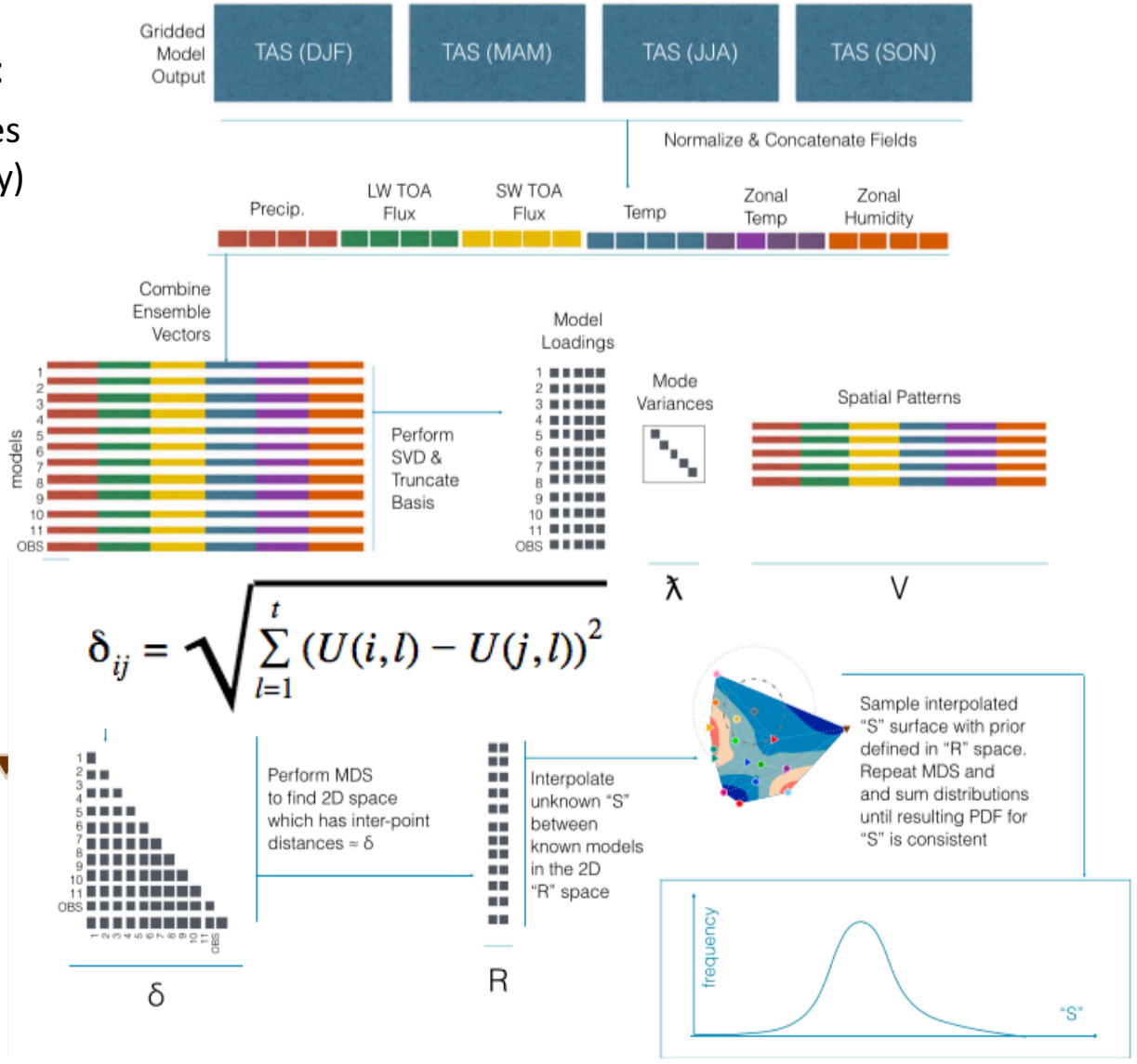
- We can consider dependence in estimates of either of these

# Dividing multi-model ensemble spread in two: epistemic and aleatory uncertainty

Sanderson et al (2015a,2015b) dealt with epistemic uncertainty:

- dependence in climatology biases (assumes no aleatory uncertainty)
- integrative metric (variables grouped together)
- Model distances projected to 2D space
- Weights intended to discount effect of model clustering

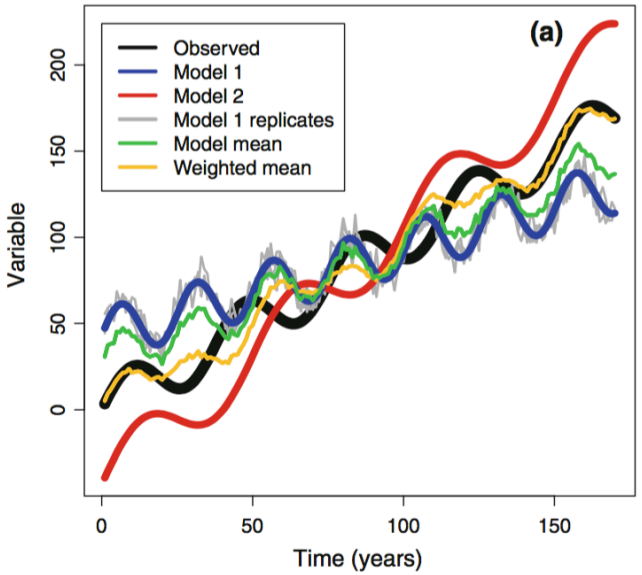


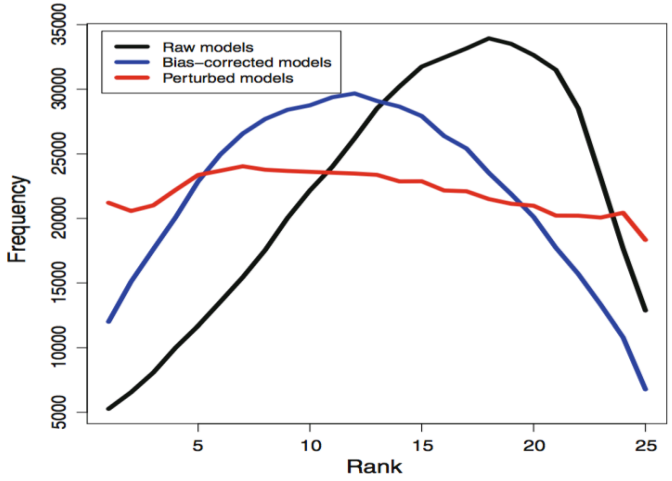


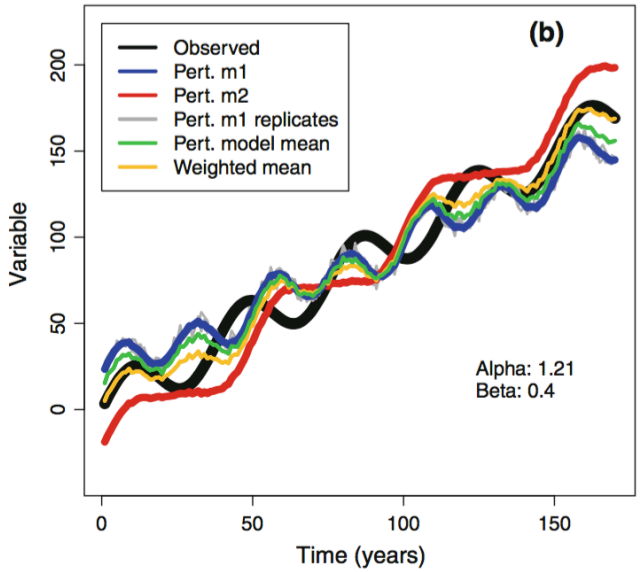
# Dividing multi-model ensemble spread in two: epistemic and aleatory uncertainty

Bishop & Abramowitz (2013) primarily dealt with aleatory uncertainty

- Attempted to calibrate ensemble to represent irreducible system uncertainty only
- Estimate forced climate signal and statistical properties of internal variability (without using I/C ensembles)
- Rescale CMIP ensemble about the forced climate signal estimate to replicate this internal variability estimate



# Structure

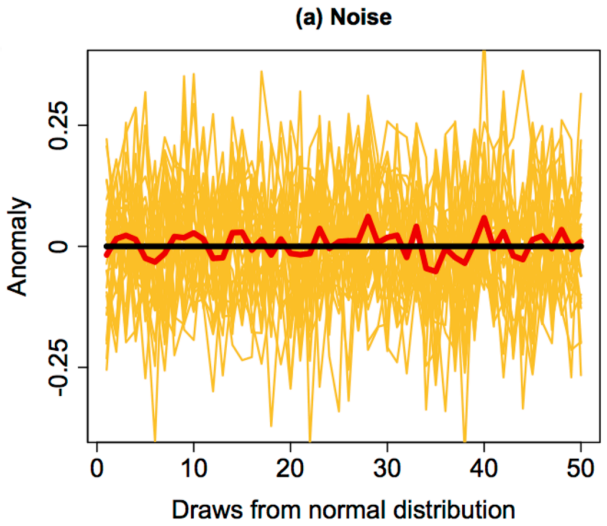
- Context
- Epistemic vs aleatory uncertainty
- Ensemble interpretation paradigms
- Why weighting / sub-selecting for dependence and/or performance is a calibration exercise
- Should calibration be application-specific or holistic?

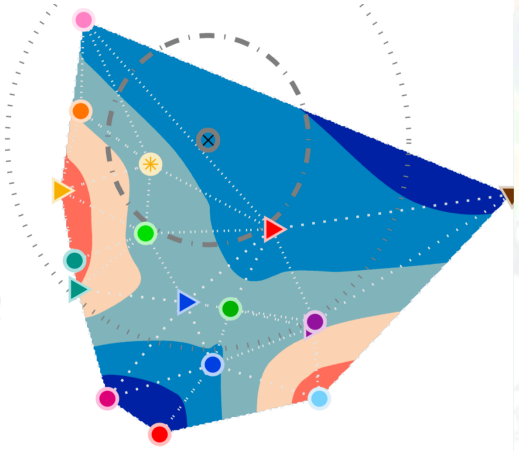
# What is an ensemble interpretation paradigm?

- Relationship between ensemble spread and observations
- The role of the ensemble mean
- What would an ensemble of ‘perfect’/independent models look like?

# Ensemble interpretation paradigms: "Truth-plus-error" paradigm

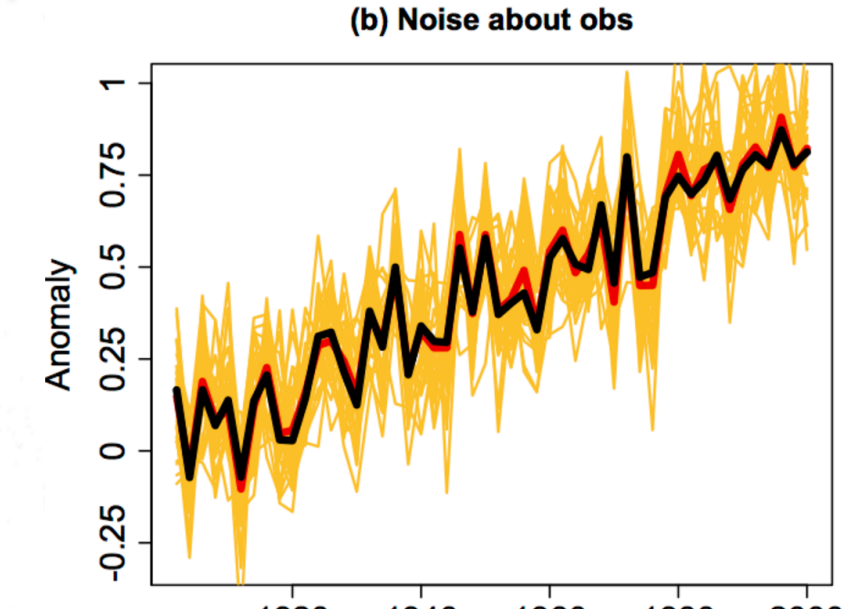
- A perfect model should essentially match observations
  - A perfect model's deviations from observations are noisy (model error as a random variable)
  - Observations are assumed to be at the centre of the distribution
- Makes some sense for climatology: assumes that climatology is deterministically predictable at timescales longer than observational record
- Why the multi-model mean works so well:
  - n independent random number fields (think model - observed), mean = 0, sd = 1
  - Standard deviation of their sum approximates $1/\sqrt{n}$
- Independence => pairwise error correlation = 0
- Implies ensemble mean should converge to observations as ensemble gets larger

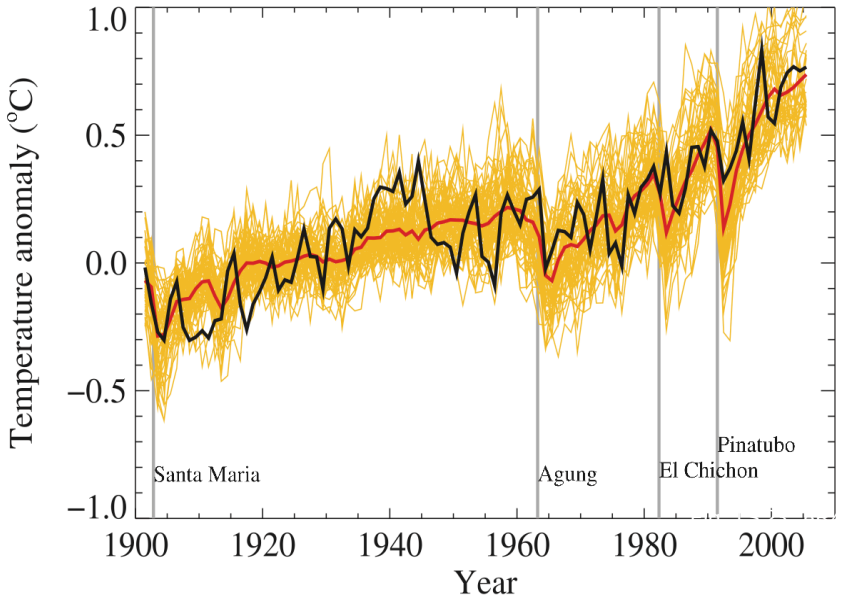


# Ensemble interpretation paradigms: "Truth-plus-error" paradigm

- Clearly not appropriate for time series
- Implies ensemble mean should converge to observations as ensemble gets larger
  - climate system is deterministic at all timescales – no internal variability
- Observations clearly look more like a model than the ensemble mean – mean is not a "true climate" and no good for variability estimates



# Ensemble interpretation paradigms: Indistinguishable paradigm (Annan and Hargreaves, 2010)

- Assumes models and observations are sampled from the same distribution in climatology space
- No statement of what spread represents (blog post: "*collective uncertainties about how best to represent the climate system*" Annan, 2010) – epistemic uncertainty?
- Implication that independence => pairwise error in spatial correlation of around 0.5, but independence is not explicitly discussed

# Ensemble interpretation paradigms: replicate Earth paradigm (Bishop and Abramowitz, 2013)

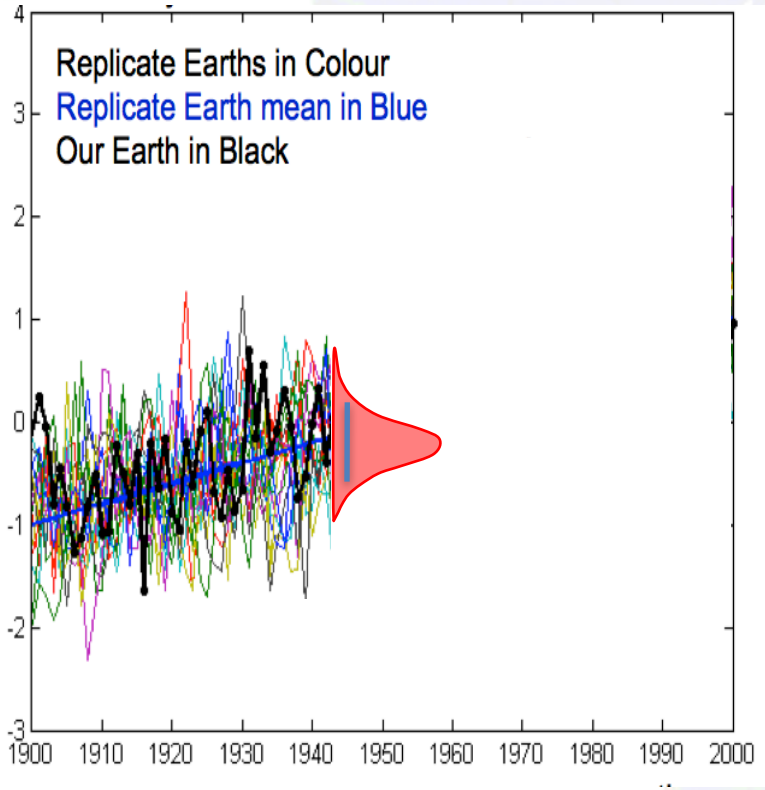
- Considers evolution of ensemble distribution in time
- Suggests that perfectly independent models and observations ('replicate Earths') would be draws from the same instantaneous distribution ('Climate PDF')
- Evolving CPDF mean represents the forced response over time
- Error correlation between replicate Earths would be 0.5
- Climate models viewed as (imperfect) attempts to create replicate Earths
  - A perfect model is a replicate Earth – one sample of inherent *system* uncertainty / internal variability
  - Model issues – our inability to create a replicate Earth – are separate source of uncertainty
  - Define replicate Earth ensemble statistical properties and show CMIPx doesn't have them
  - Transform CMIPx ensemble so it does have these properties

# 3. Replicate Earth paradigm

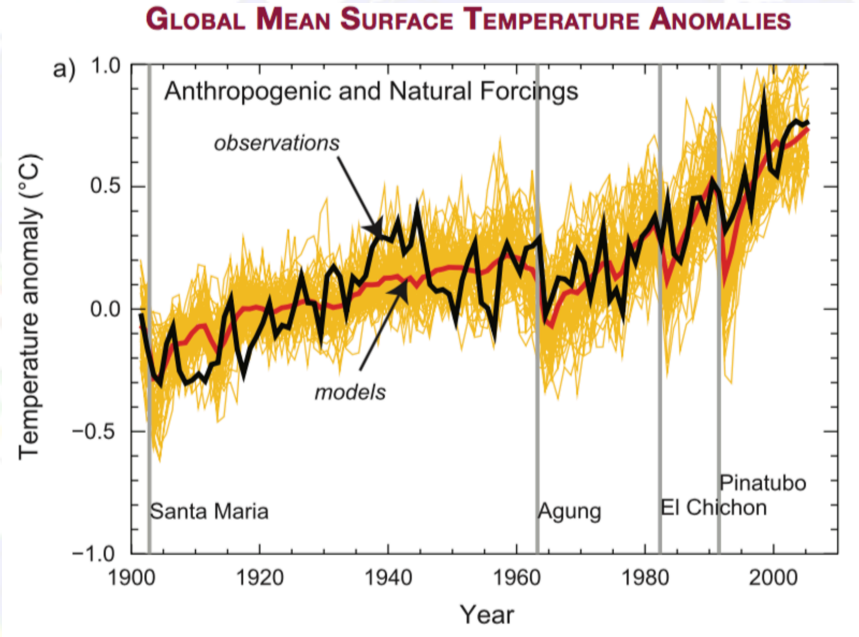
(Bishop and Abramowitz, 2013)

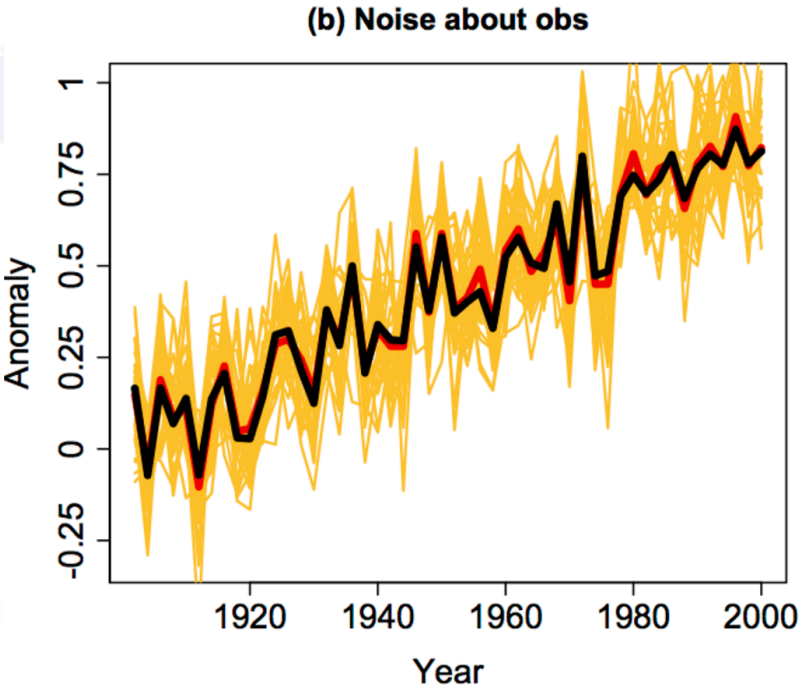
Replicate Earth ensembles (NOT models) have two key properties:

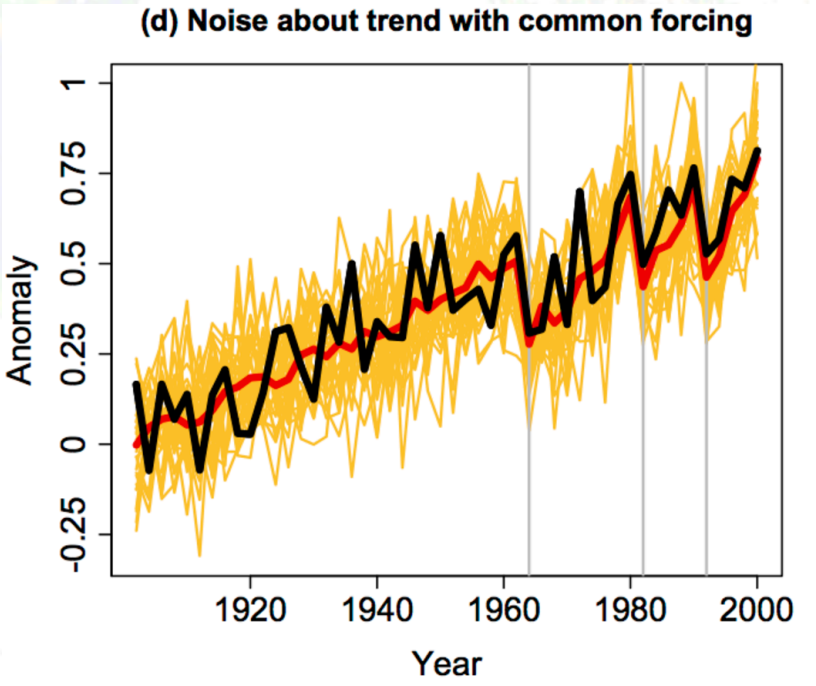
1. Mean of the distribution of replicate Earths (blue line) is ***the*** linear combination of replicate Earths that minimises distance from our Earth's observations.
2. Time average of the instantaneous variance of replicate Earths is approximately equal to the variance of observations about the CPDF mean over time

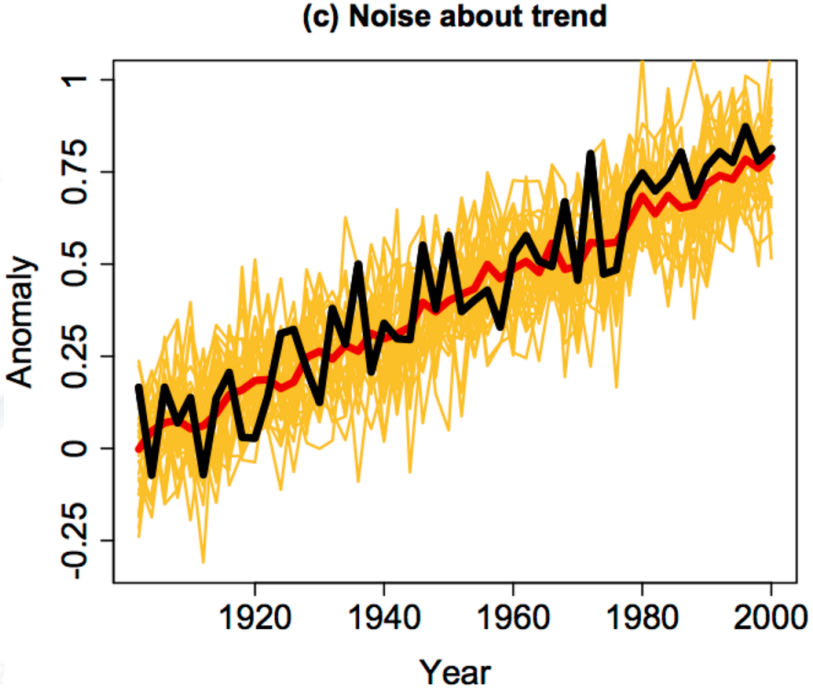



The CMIP3/5 ensembles don't have these properties, but we can transform them so they do…


GLOBAL MEAN SURFACE TEMPERATURE ANOMALIES
a)
Anthropogenic and Natural Forcings
observations
models
Santa Maria
Agung
El Chichon
Pinatubo
Temperature anomaly (°C)
Year


(b) Noise about obs
Anomaly
Year


(d) Noise about trend with common forcing
Anomaly
Year


(c) Noise about trend
Anomaly
Year

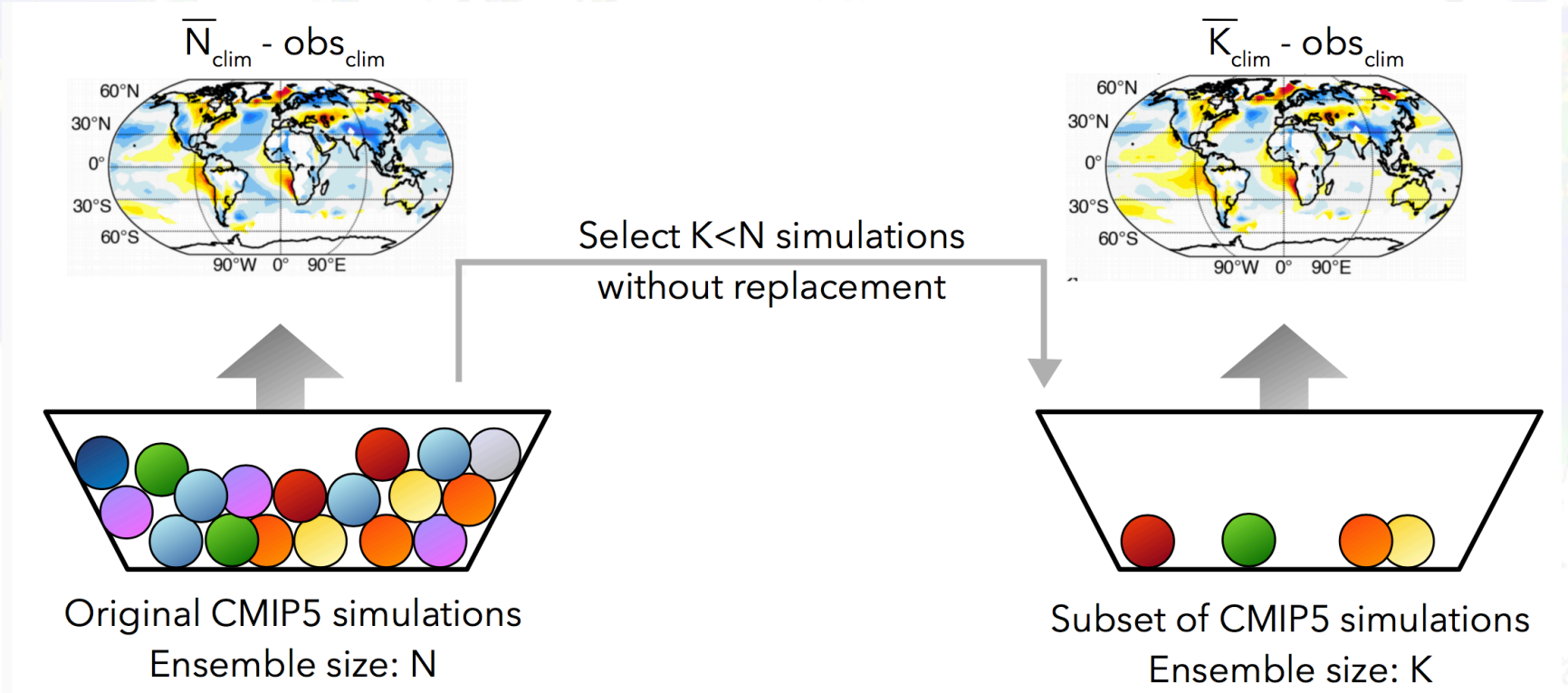
# Structure

- Context
- Epistemic vs aleatory uncertainty
- Ensemble interpretation paradigms
- Why weighting / sub-selecting for dependence and/or performance is a calibration exercise
- Should calibration be application-specific or holistic?

# Ensemble sub-sampling to account for dependence

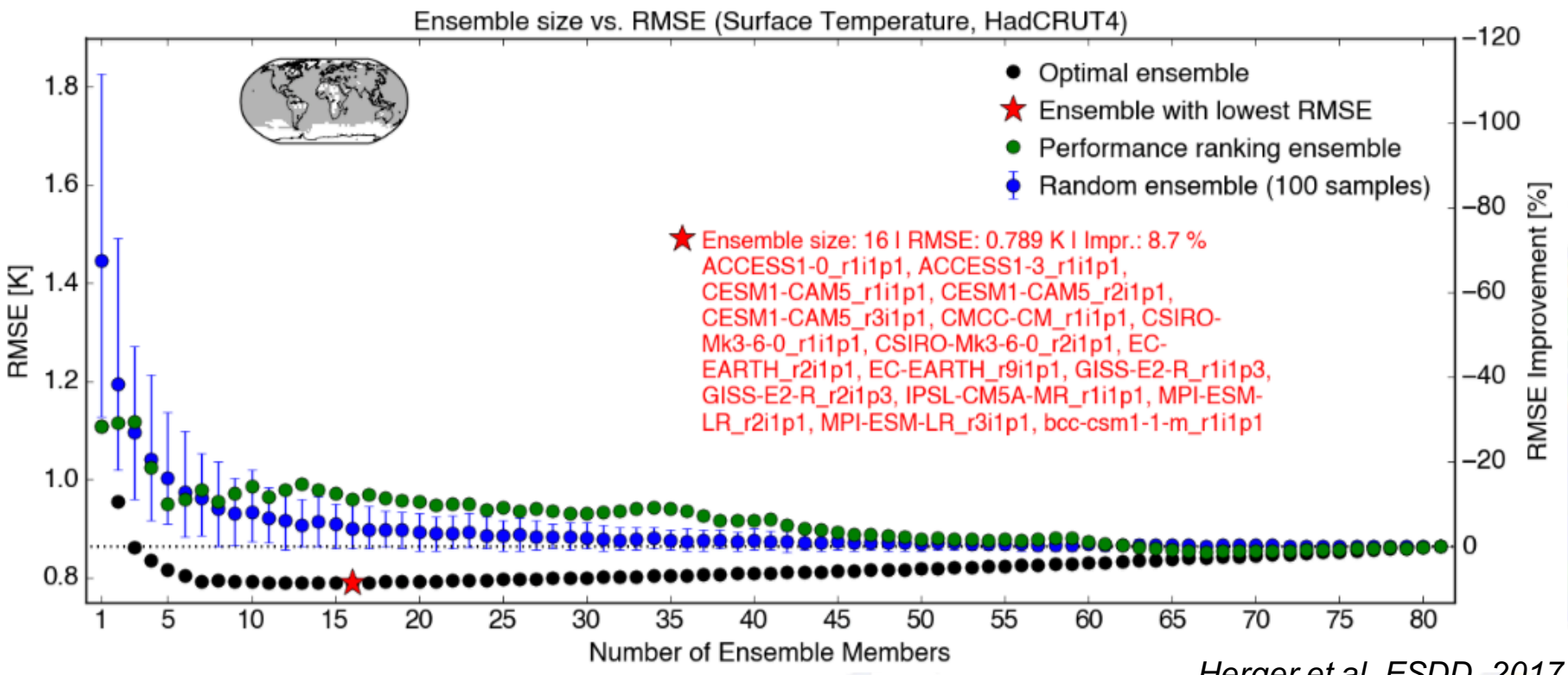
Three ensemble sub sampling approaches:

1. Random sampling of *K* simulations from a pool of *N* (100 times)
2. Choose the best performing *K* simulations (in terms of climatology)
3. Choose the *K* simulations whose mean has minimum RMSE in climatology against obs – account for dependence in regional biases



# Ensemble size vs RMSE

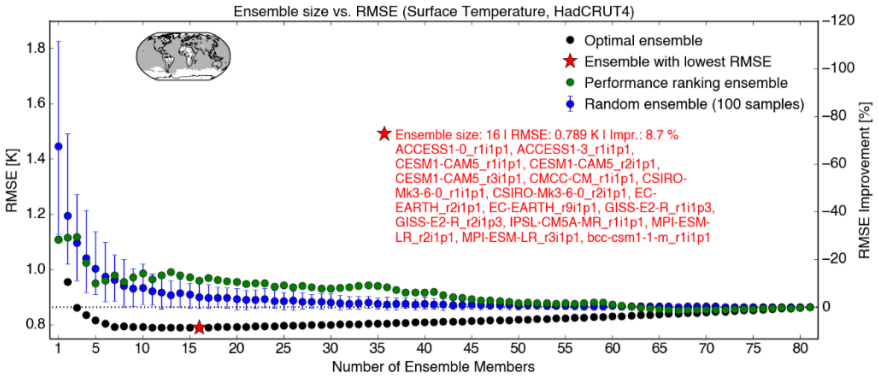
- Choosing the optimal ensemble is non-trivial – choosing K=40 (of N=81) means there are 212,392,290,424,395,860,814,420 possible ensembles

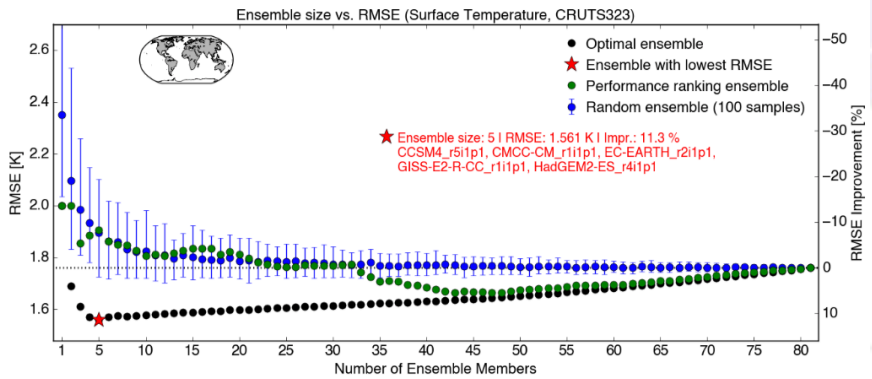


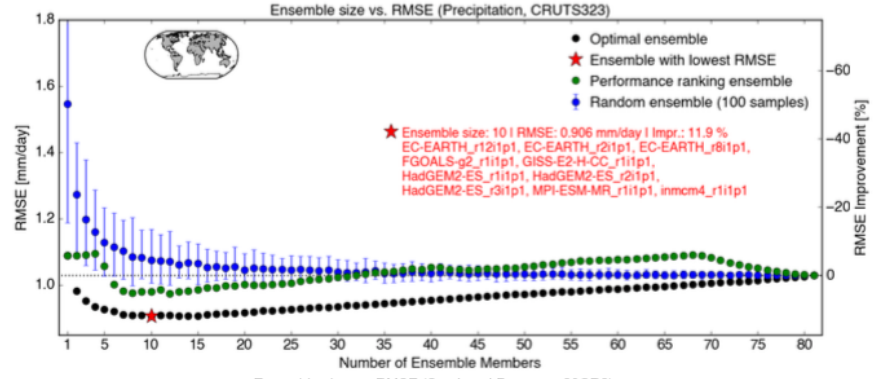

*Herger et al, ESDD, 2017*

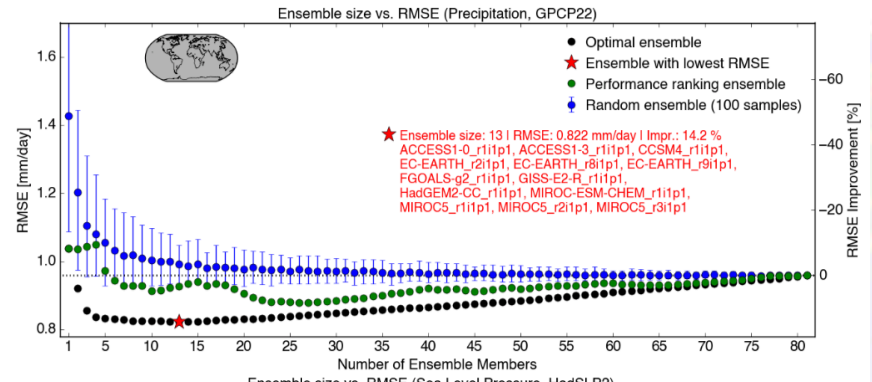
- Choosing the best performing models does not imply the best performing ensemble mean – dependence degrades the mean

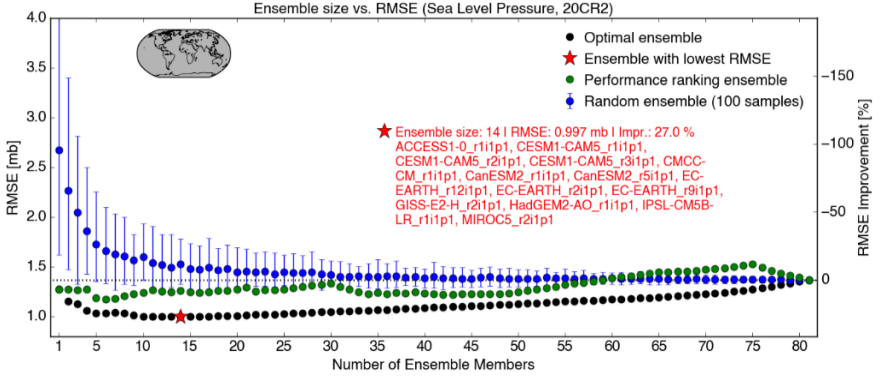
# Results differ across variables and obs products

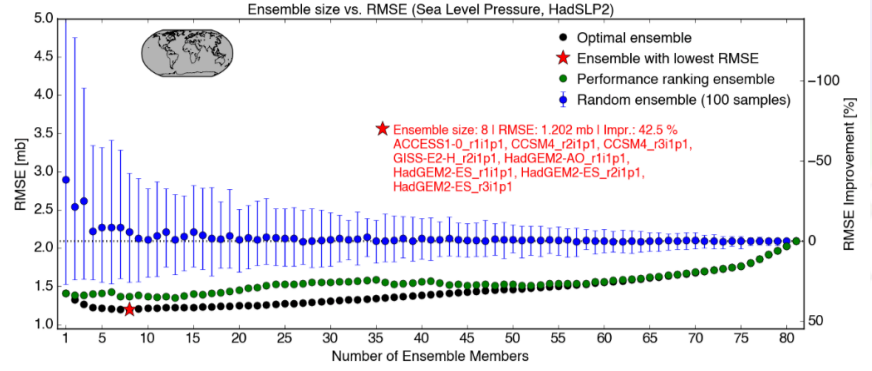


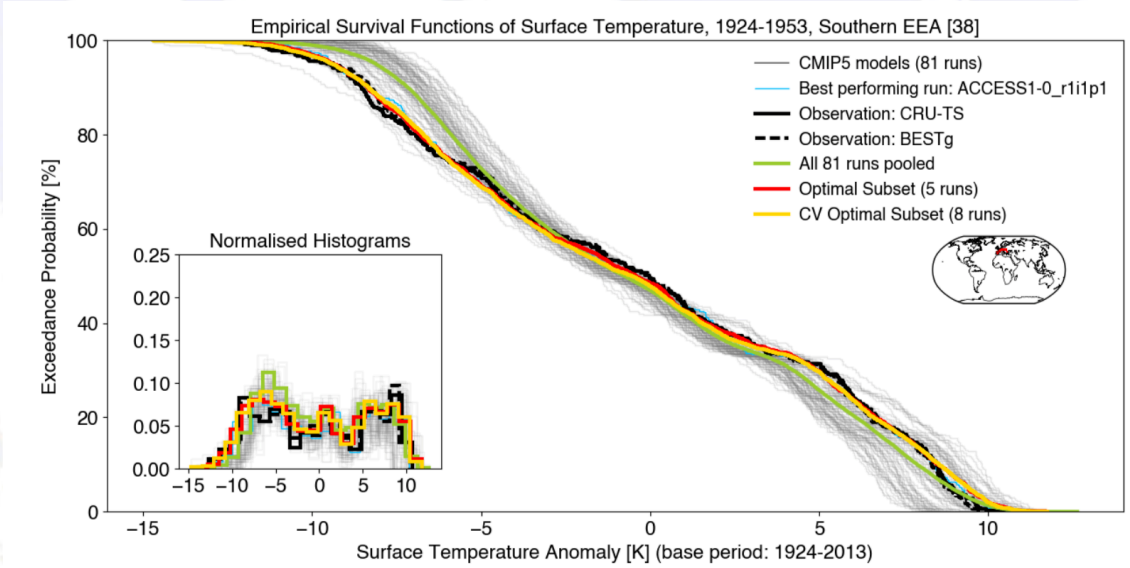



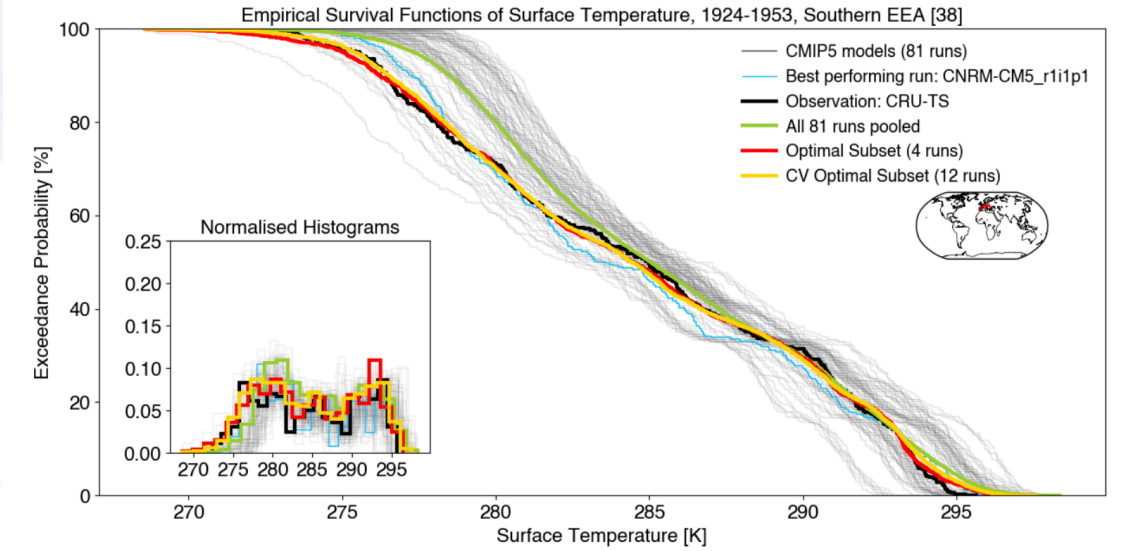

*Herger et al, ESDD, 2017*

# Results differ across different metrics and regions

- This optimisation approach can be applied to a range of definitions of dependence by adjusting the cost function used in the optimisation

- Select ensemble subset that minimises K-S test
- With or without prior bias correction
- Results vary markedly by region




*Herger et al, in prep*

# Structure

- Context
- Epistemic vs aleatory uncertainty
- Ensemble interpretation paradigms
- Why weighting / sub-selecting for dependence and/or performance is a calibration exercise
- Should calibration be application-specific or holistic?

# Should calibration be application-specific or holistic?

- Optimal weights / ensemble selection are specific to time period, climatology vs time varying, time step size, resolution, region, metric, variable, data set
  - i.e. specific to each application
- In most cases, this calibration process is robust when tested out of sample
  - Cal/val on different time periods, model-as-truth / perfect model experiments
  - Dependence exists and can be accounted for
- Is this satisfactory?
- Should we just use the whole ensemble? Isn't an entirely uncalibrated ensemble underutilising the information we have?
- How would we implement universal/holistic calibration for model dependence for a given ensemble?

# Towards holistic calibration - is a single integrated cost function the best solution?

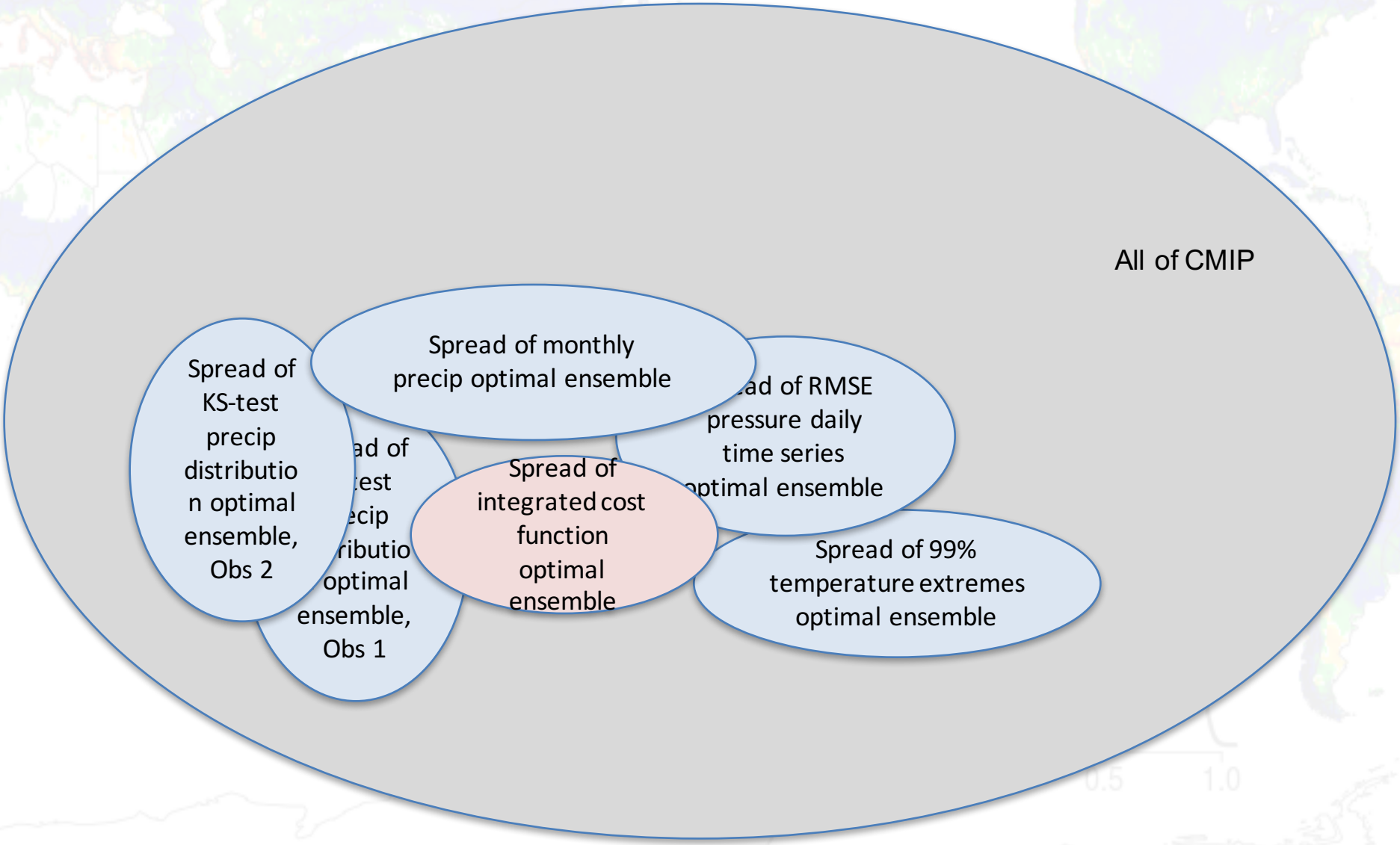
- E.g. should we find weights, or choose sub-ensemble that optimises:

$$\text{mean}(ta_{obs1}) \,\&\, \text{cor}(ta_{obs1}) \,\&\, \text{sd}(ta_{obs1}) \,\&\, \text{Kstest}(ta_{obs1}) \,\&$$
$$\text{mean}(ta_{obs2}) \,\&\, \text{cor}(ta_{obs2}) \,\&\, \text{sd}(ta_{obs2}) \,\&\, \text{Kstest}(ta_{obs2}) \,\&$$
$$\text{mean}(pr_{obs1}) \,\&\, \text{cor}(pr_{obs1}) \,\&\, \text{sd}(pr_{obs1}) \,\&\, \text{Kstest}(pr_{obs1}) \,\&$$
$$\text{mean}(pr_{obs2}) \,\&\, \text{cor}(pr_{obs2}) \,\&\, \text{sd}(pr_{obs2}) \,\&\, \text{Kstest}(pr_{obs2}) \,\&$$

……. (with normalising weights)....

- Weights non-commensurable variables & metrics against each other
- Does not fairly represent the uncertainty involved

# Towards holistic calibration - is a single integrated cost function the best solution?



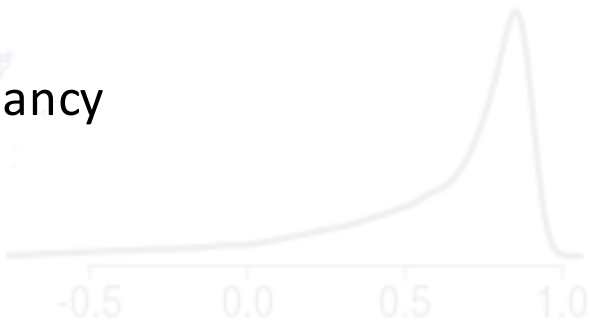
# Towards holistic calibration – why an integrative cost function seems like a bad idea

- We have a chronic lack of constraint, which results in confirmation holism:
  - ‘‘If the predicted phenomenon is not produced, not only is the questioned proposition put into doubt, but also the whole theoretical scaffolding used by the physicist’’ (Duhem 1954).
  - If we only have enough observational data to test the end result of a long chain of hypotheses / parametrisations strung together, and not every step in the process:
  - If we don’t get the right answer, it’s very hard to understand why
  - If we get the right answer, we don’t know whether it’s for the right reasons

- dim(world) >> dim(model) > dim(obs)

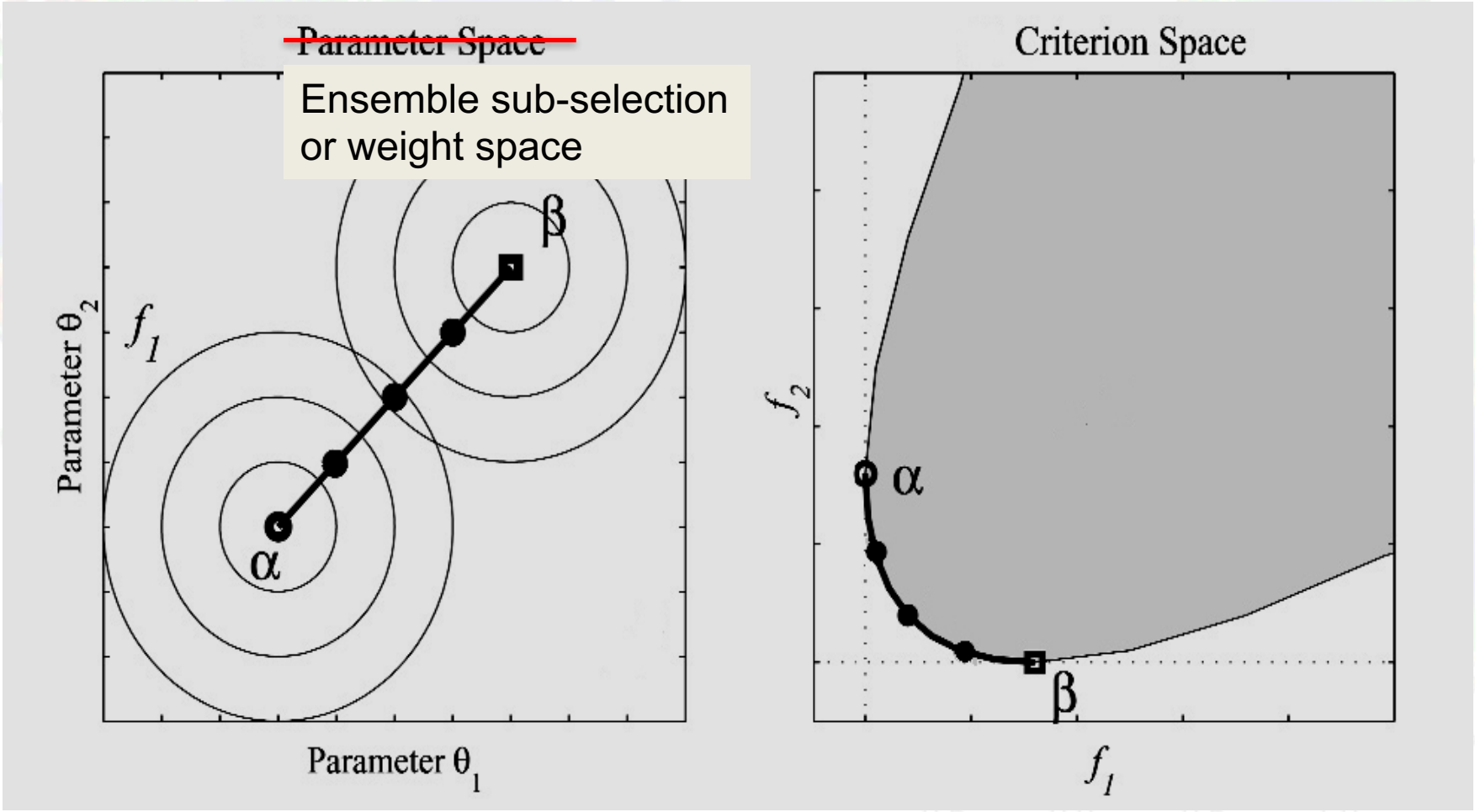
# Towards holistic calibration – why an integrative cost function seems like a bad idea

- The discrepancy between solutions that are optimal for different variables, metrics, obs products etc, is meaningful!
- Conservation of crap
- If our ensemble (say CMIP) were perfectly calibrated, each of these optimisations would give the same result – *the* optimal ensemble
- The inability of our ensemble to deliver this is an indication of its inability to cover the degrees of freedom needed to simultaneously answer the problems that were interested in (? Over-calibration?)
- We could attempt to measure the discrepancy

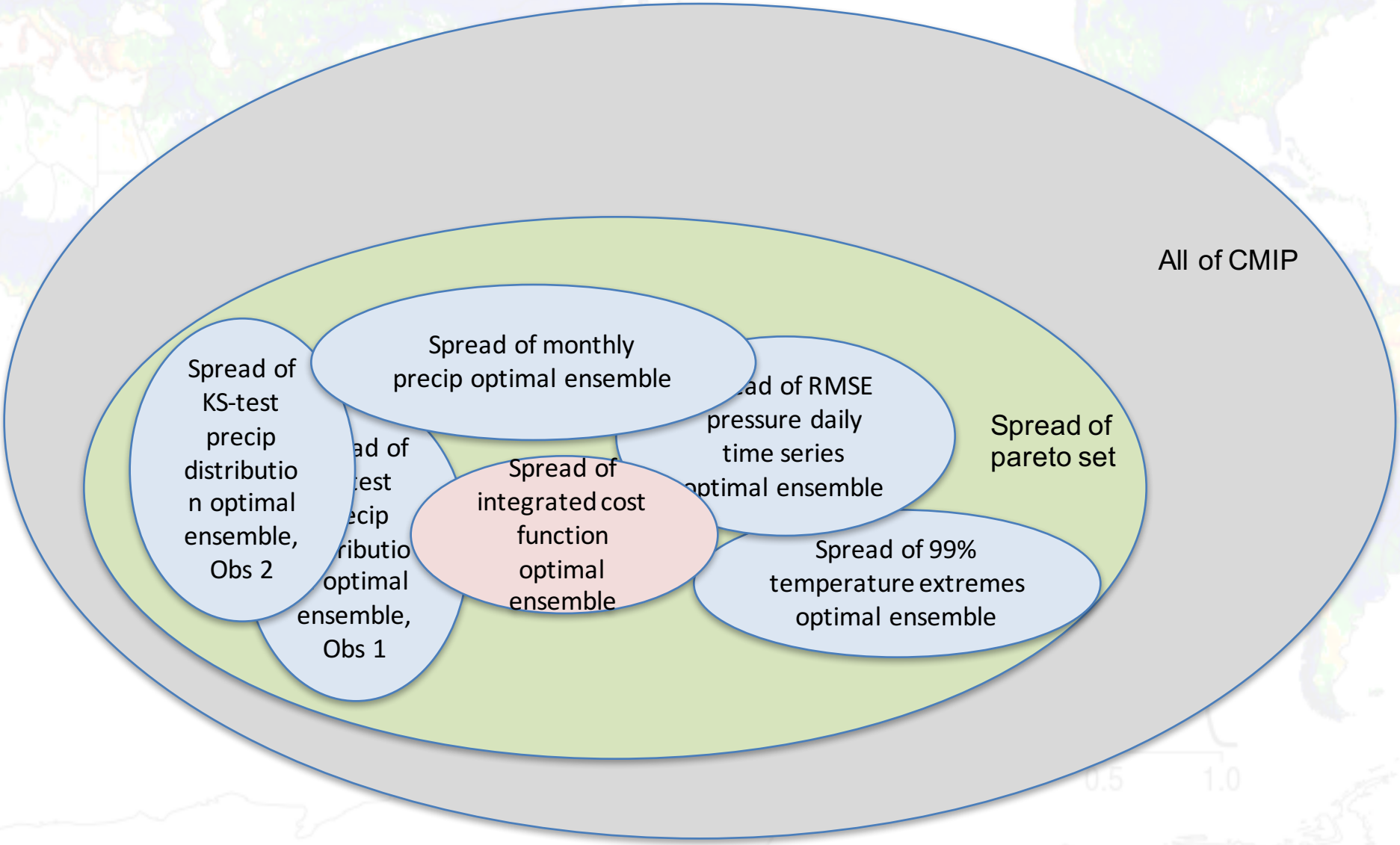


# Towards holistic calibration – a better idea

- A pareto set – an ensemble of ensembles – seems a better approach?



# Towards holistic calibration – why an integrative cost function is a bad idea



# Conclusions

- Weighting or sub-selecting ensemble members for independence / performance is essentially a calibration exercise

- If we are interested in a best estimate of a single projection property / quantity then calibrating for it likely to improve estimates
  - Several techniques are available, each appropriate for different applications
  - Needs to be tested out of sample, in a way commensurate with the application
  - Does this help with meaningful uncertainty estimates?

- If we want/need a holistic accounting for model dependence with more meaningful uncertainty estimates, an ensemble of ensembles (pareto set) might be a more reasonable approach