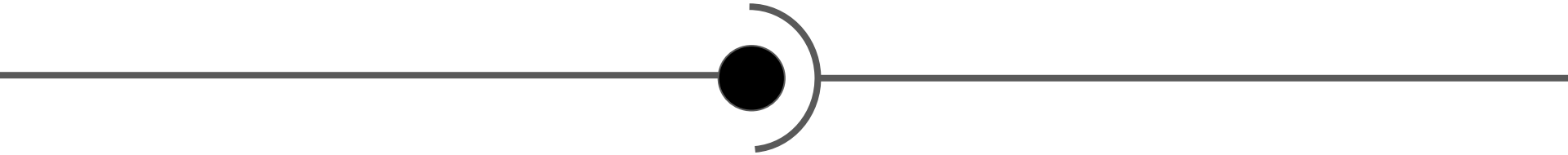
Reconciling approaches to address model performance and interdependency in CMIP6

Ben Sanderson, NCAR

Earth System Model Evaluation to Improve Process Understanding, Aspen, 2017



Bayesian Worldview

IPCC AR5
Tebaldi et al (2005,2009)

Replicate Earth Paradigm

Bishop & Abramowitz (2013)

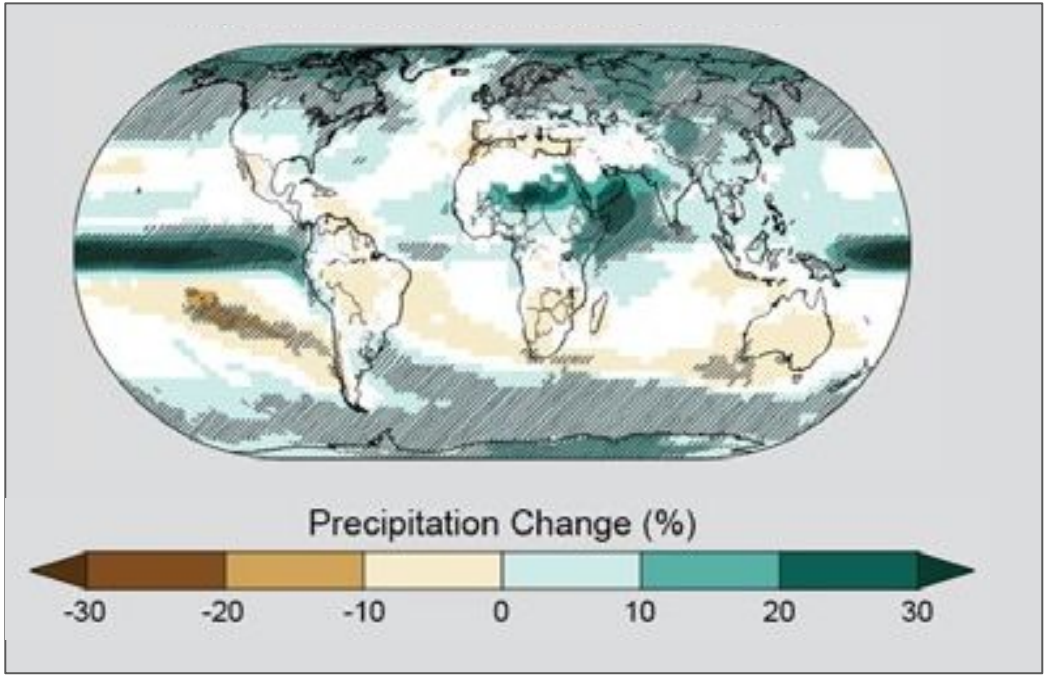
Geometric Worldview

Sanderson (2015a,b, 2017)
Knutti (2017)

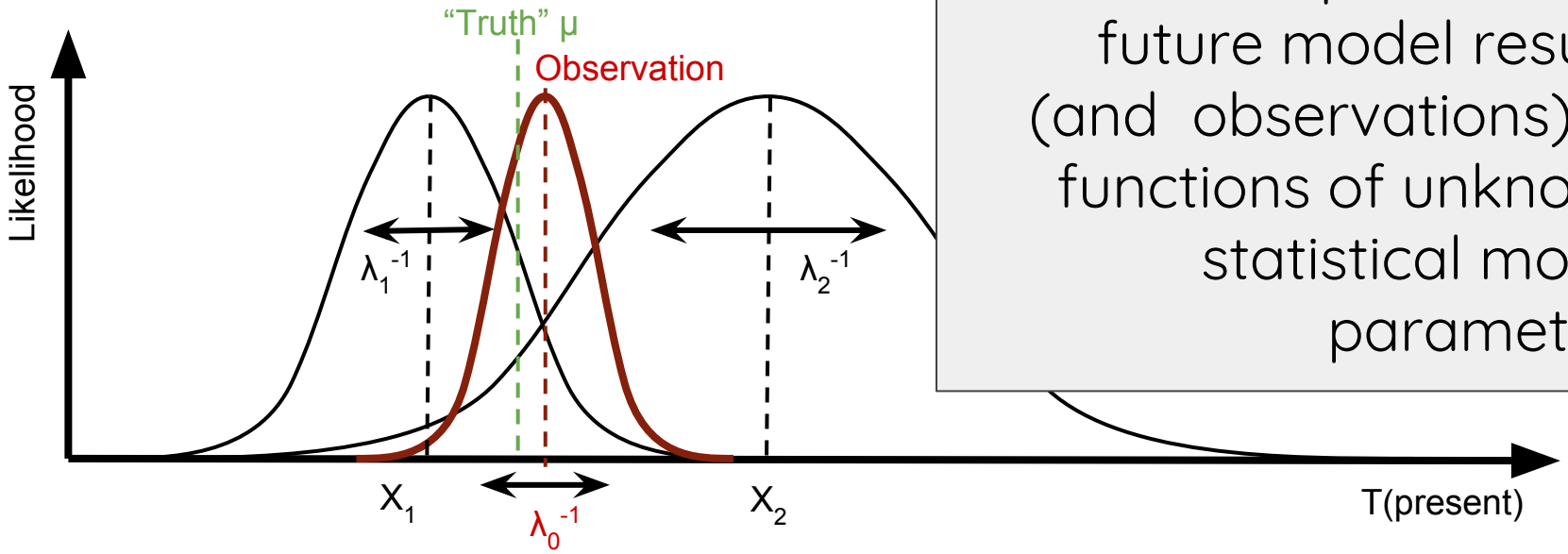
Bayesian Worldview

Ensemble members are observations or samples of the true state of the climate system. Additional models in agreement will improve confidence in the most likely future climate.

IPCC AR5
Tebaldi et al (2005,2009)



Accounting for model skill: the Bayesian approach



Framework establishes likelihood of present and future model results (and observations) as functions of unknown statistical model parameters

In the fashion of Tebaldi et al (2005)

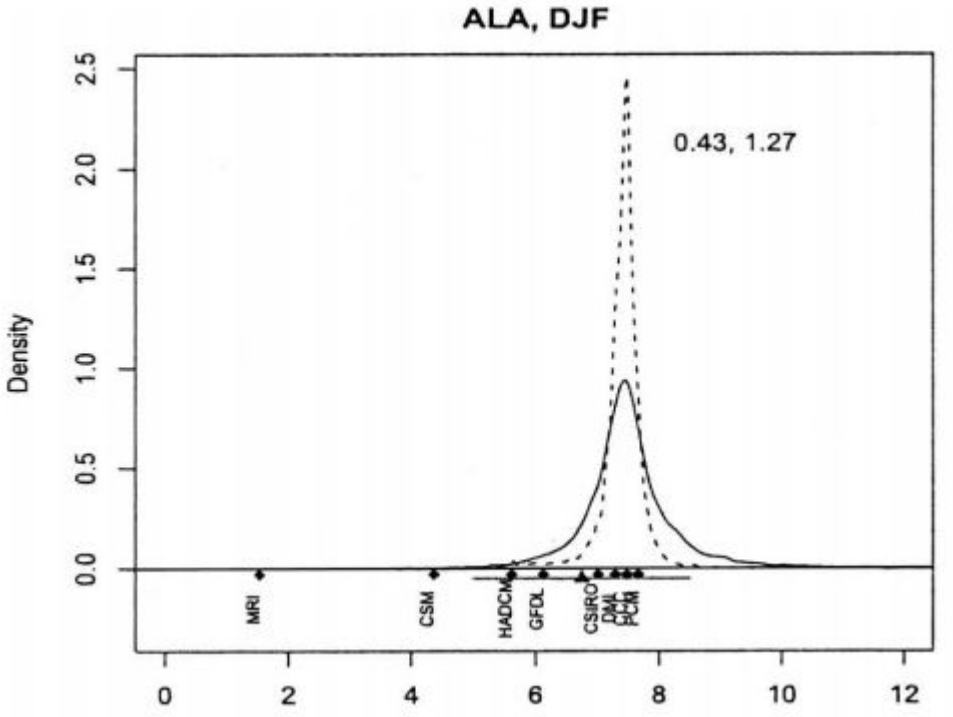
Assumptions

Bayesian Worldview

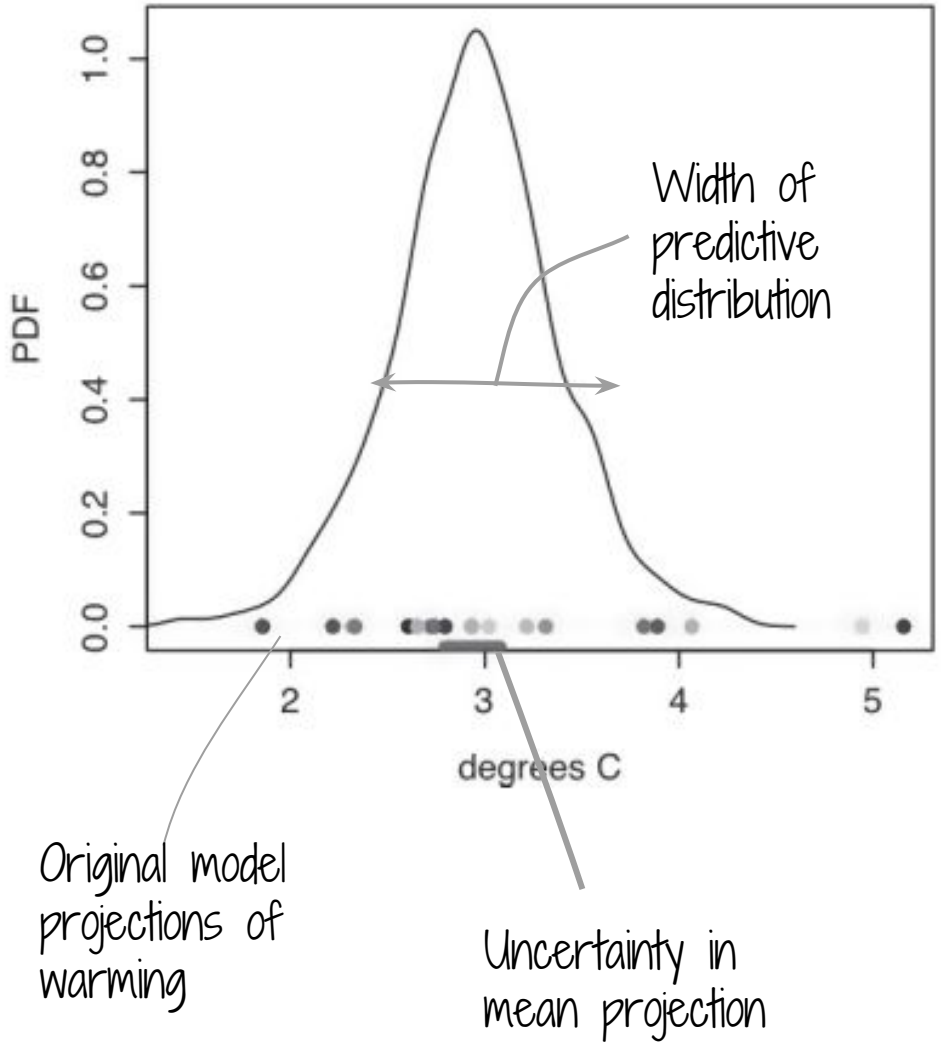
IPCC AR5
Tebaldi et al (2005,2009)

1) More models in agreement imply more confidence in projection
2) Each model is an equally plausible representation the true system
3) Present-day performance can be informative about future behavior

Monte-Carlo Sampling over self-consistent parameter sets in posterior distribution yields PDFs for model parameters. Large number of models -> very low spread in projection

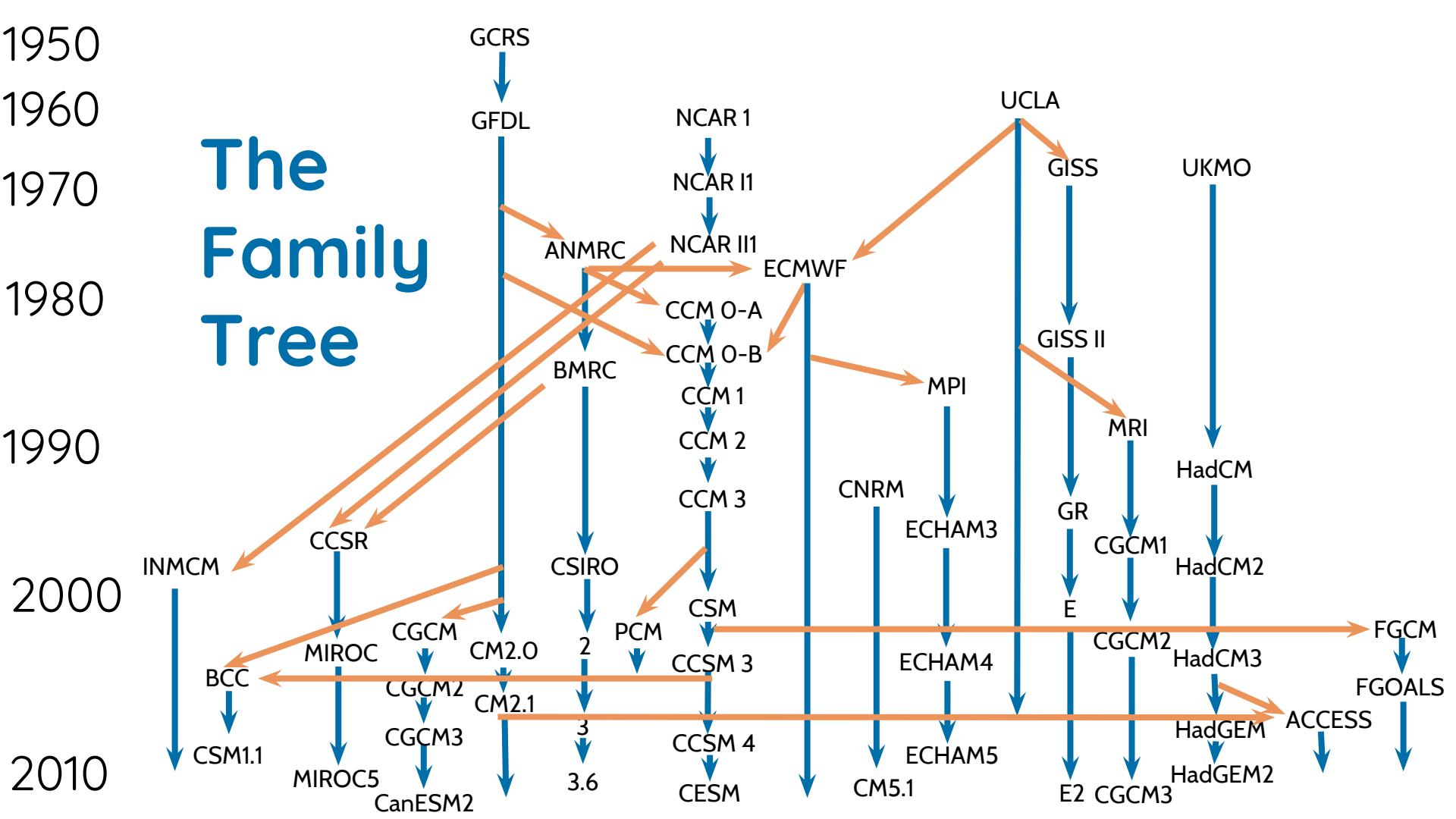


Tebaldi et al (2005)

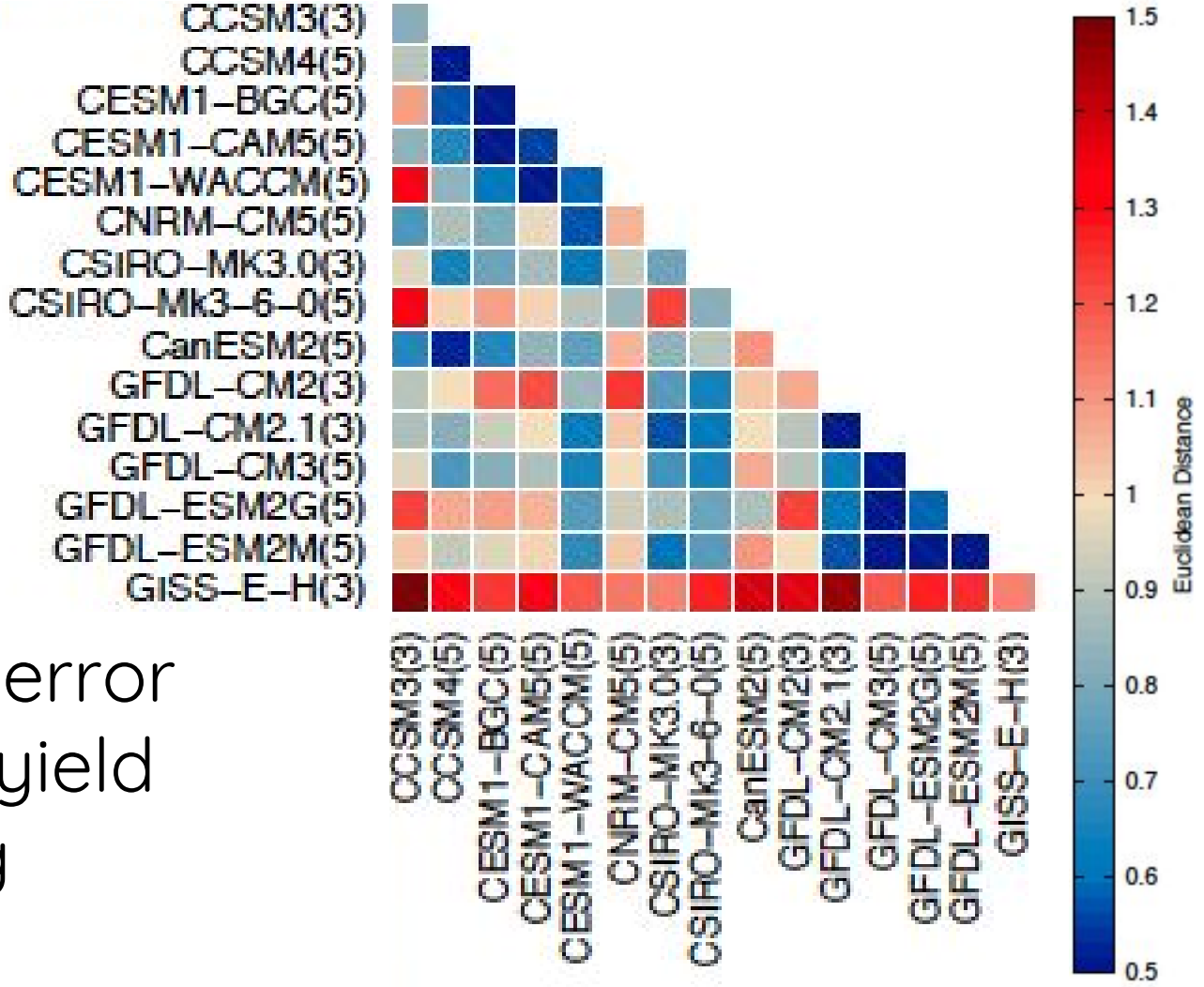


A predictive distribution considers possible plausible states of the real system, informed by the prior of CMIP model precision

Tebaldi *et al* (2009)

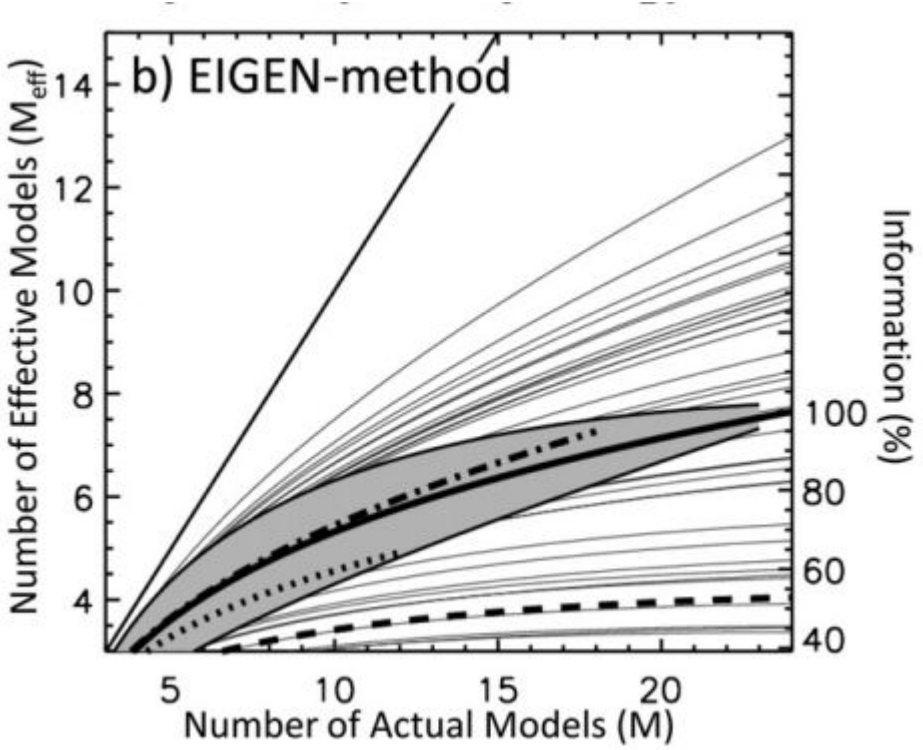

1950
1960
1970
1980
1990
2000
2010
The Family Tree
GCRS
GFDL
NCAR 1
NCAR I1
NCAR II1
ANMRC
ECMWF
UCLA
GISS
UKMO
CCM 0-A
CCM 0-B
BMRC
GISS II
CCM 1
CCM 2
CCM 3
MPI
MRI
HadCM
CNRM
ECHAM3
GR
CGCM1
HadCM2
INMCM
CCSR
CSIRO
CSM
E
FGCM
CGCM
PCM
CGCM2
MIROC
CM2.0
2
CCSM 3
ECHAM4
HadCM3
FGOALS
BCC
CGCM2
CM2.1
ACCESS
CGCM3
3
CCSM 4
HadGEM
CSM1.1
ECHAM5
MIROC5
3.6
HadGEM2
CanESM2
CESM
CM5.1
E2
CGCM3

From Sanderson *et al* (2015a)



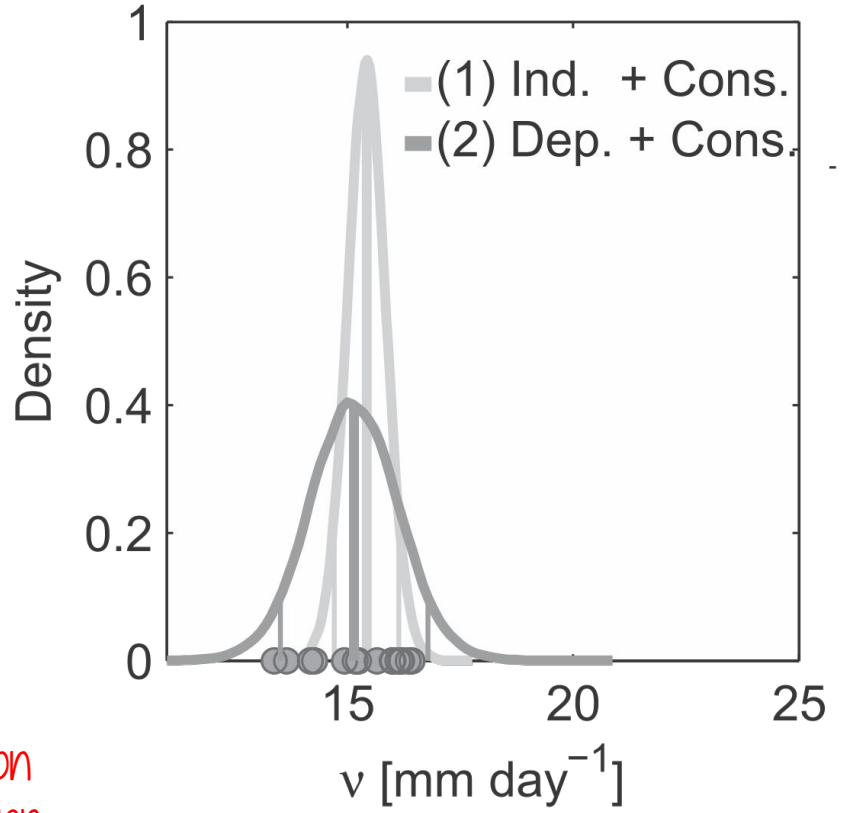
Inter-model distances (or error correlations) yield an underlying structure…

Eigen-analysis of this matrix suggests that the number of effective models may be smaller than the number of actual models…



Pennell and Reichler (2011)

Solution 1: assume models are **correlated,** drawn from a multivariate distribution



Multi-Model temperature vector

"True" temperature

Common bias

Correlation matrix

Common precision

$$\mathbf{X} \sim N(\mu+\beta, \mathbf{R}\,\lambda^{-1})$$

Sunyer *et al* (2014)

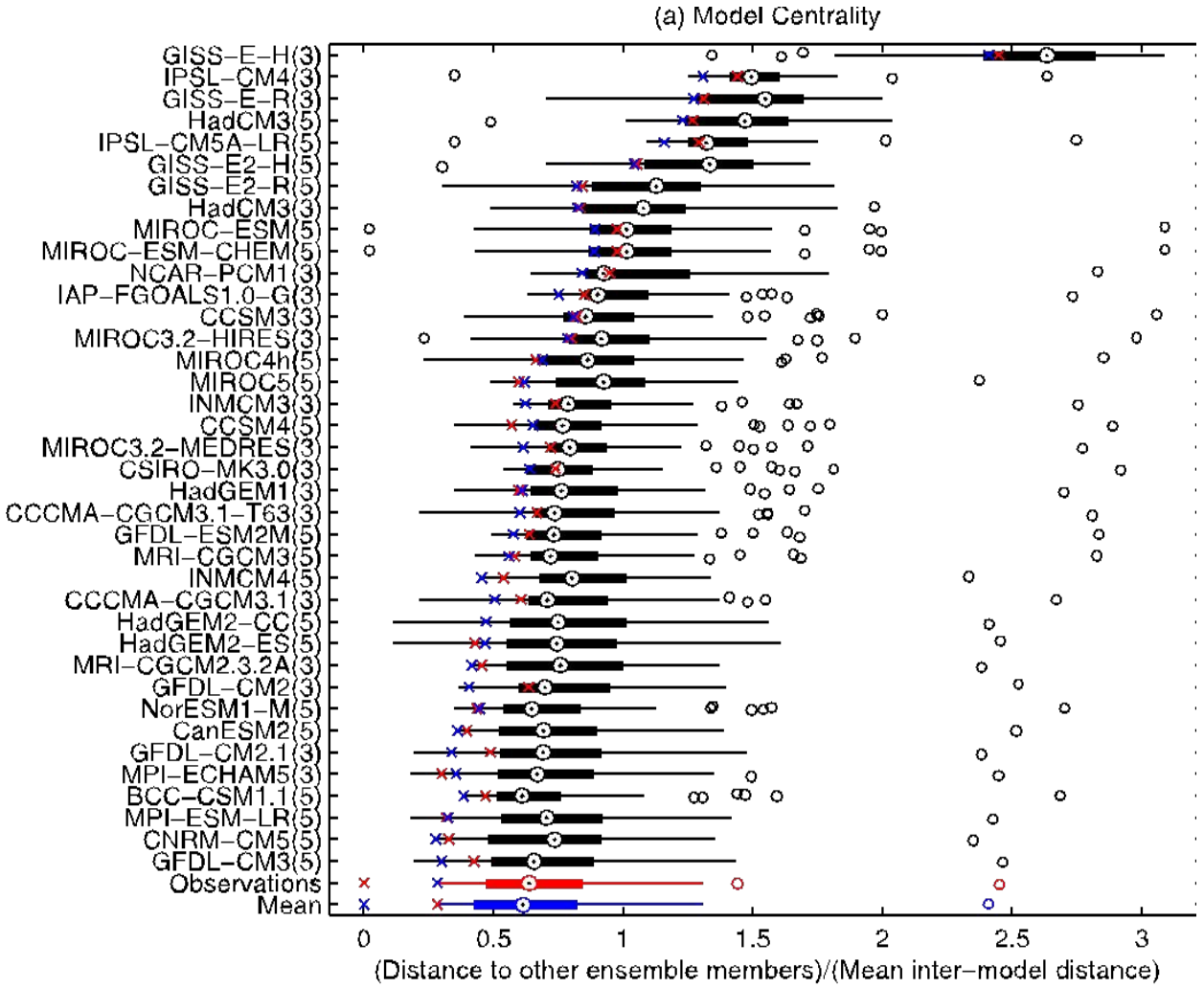
Bayesian Worldview

IPCC AR5
Tebaldi et al (2005,2009)

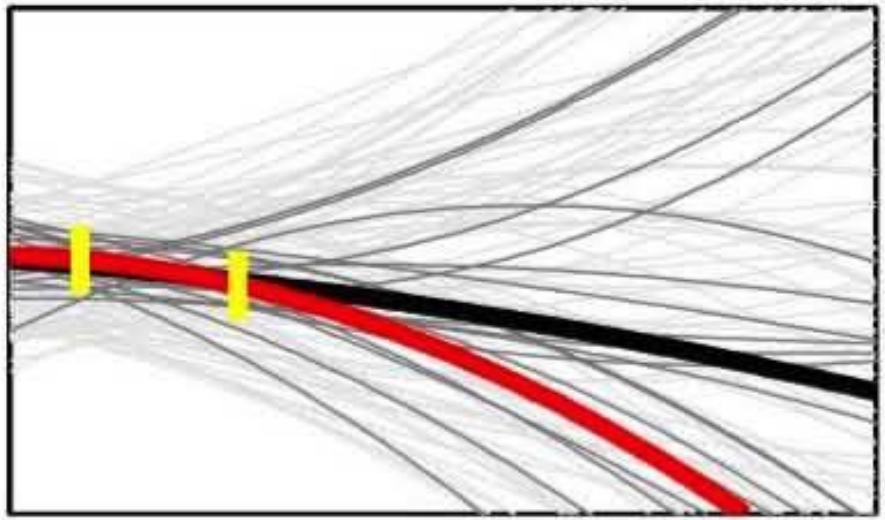
1) More **independent** models in agreement implies more confidence in **mean** projection
2) Each model is an equally plausible representation of the true system
3) Present-day performance can be informative about future behavior

Models are in general closer to obs than **any** other model…

Sanderson and Knutti (2012)

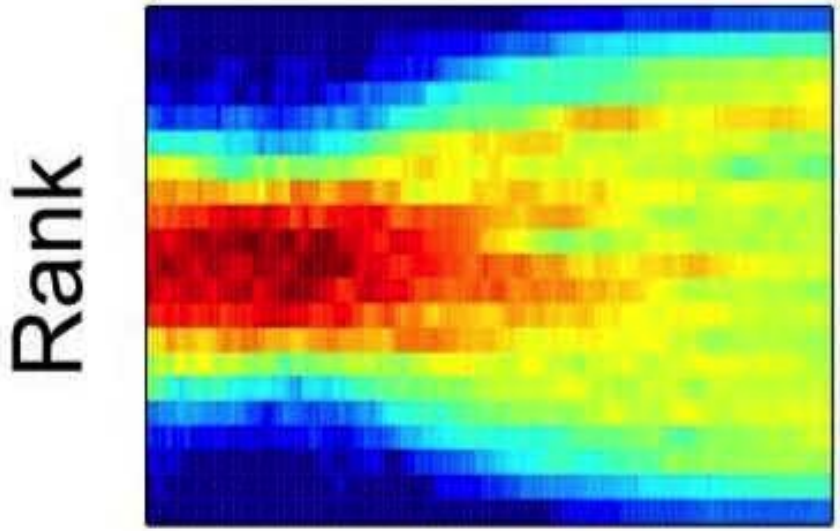


Present day apparent truth-centeredness can be an artifact of **tuning**



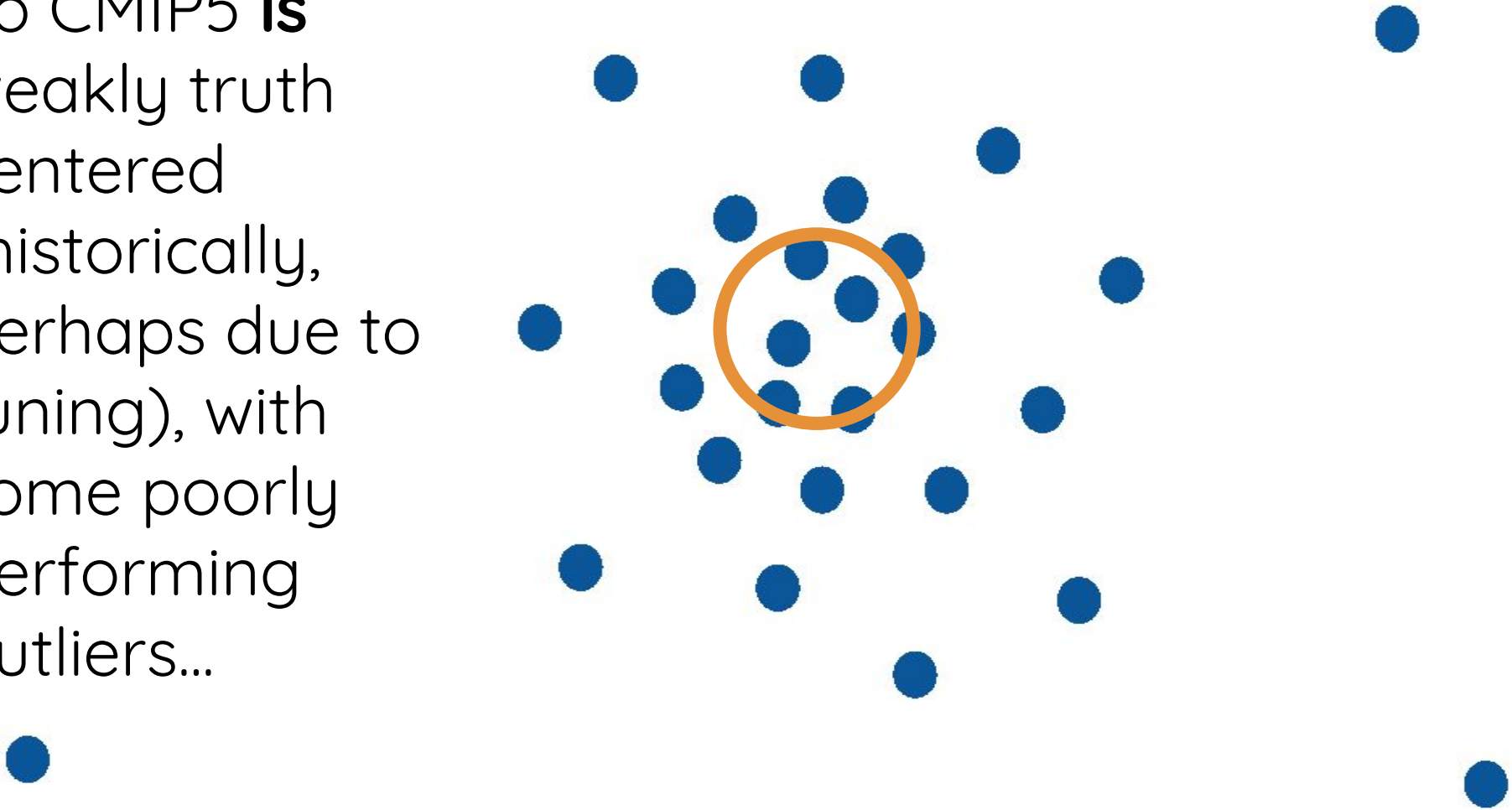
Time

Sanderson and Knutti (2012)

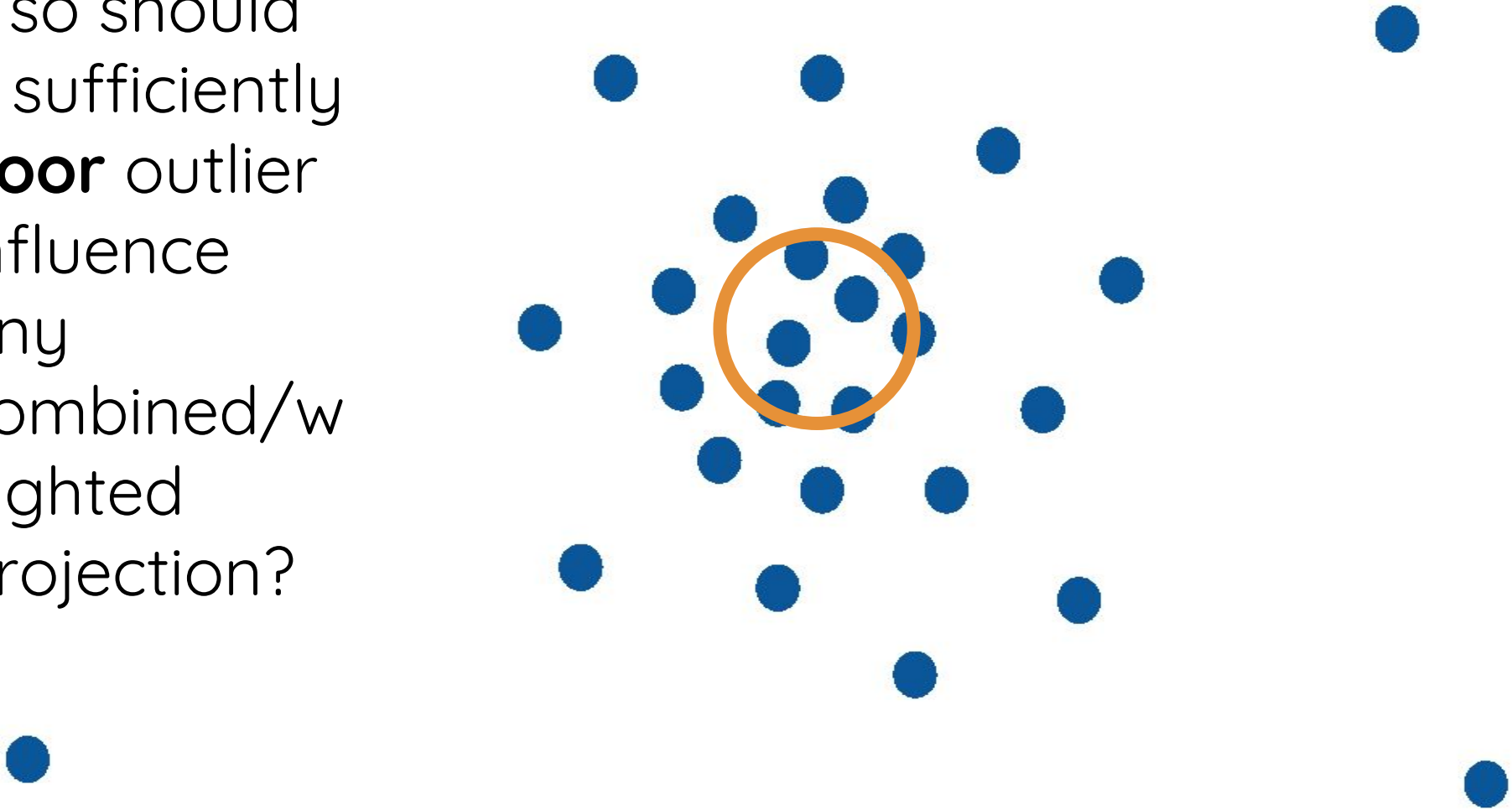


Time

So CMIP5 **is** weakly truth centered (historically, perhaps due to tuning), with some poorly performing outliers…



… so should a sufficiently **poor** outlier influence any combined/weighted projection?



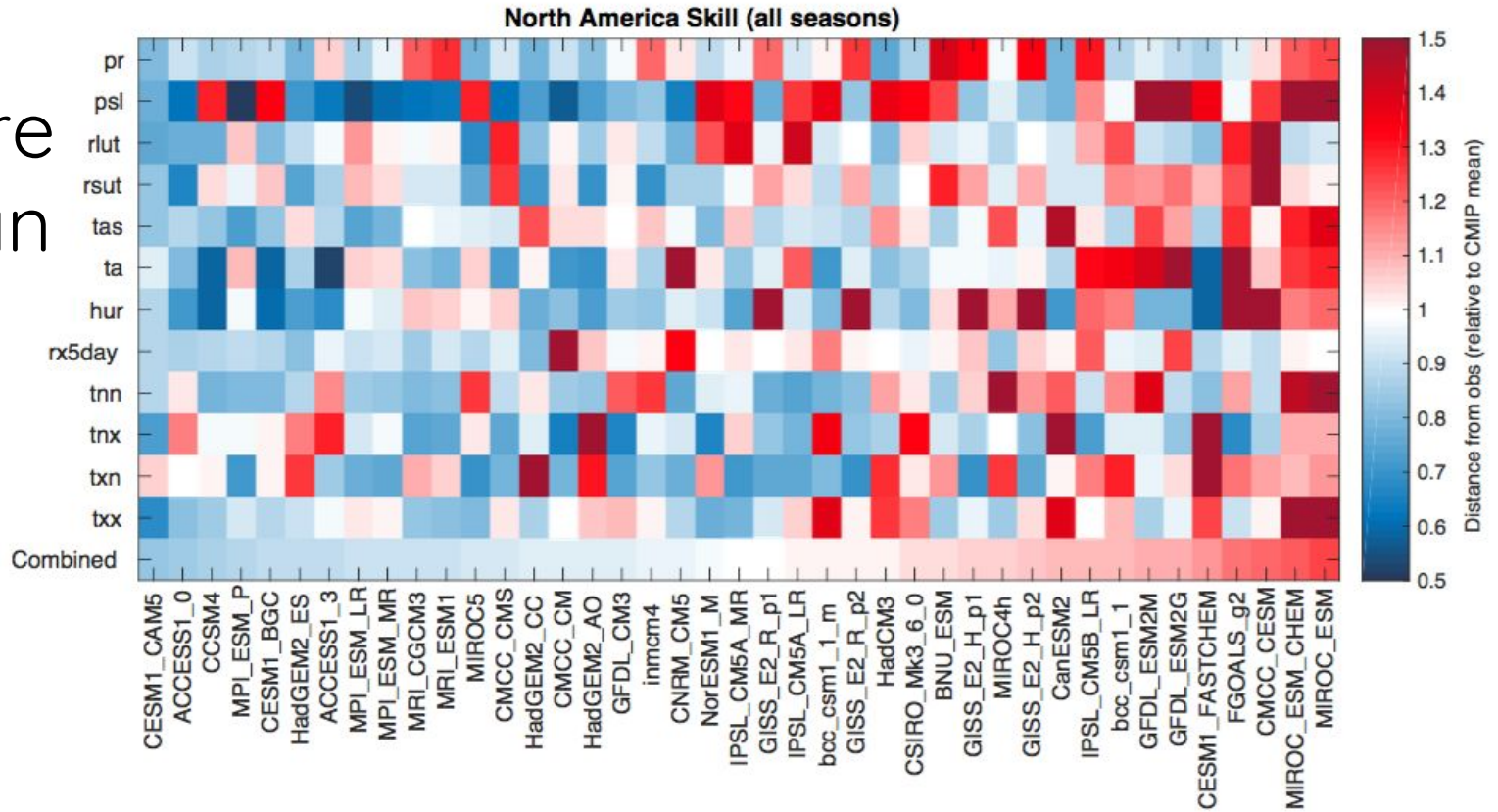
Assumptions

Bayesian Worldview

IPCC AR5
Tebaldi et al (2005,2009)

1) More **independent** models in agreement implies more confidence in mean projection (fixable)
2) Each model is an equally plausible representation of the true system (not generally true, tuning overstates agreement, very poor models would overinflate uncertainty)
3) Present-day performance can be informative about future behavior

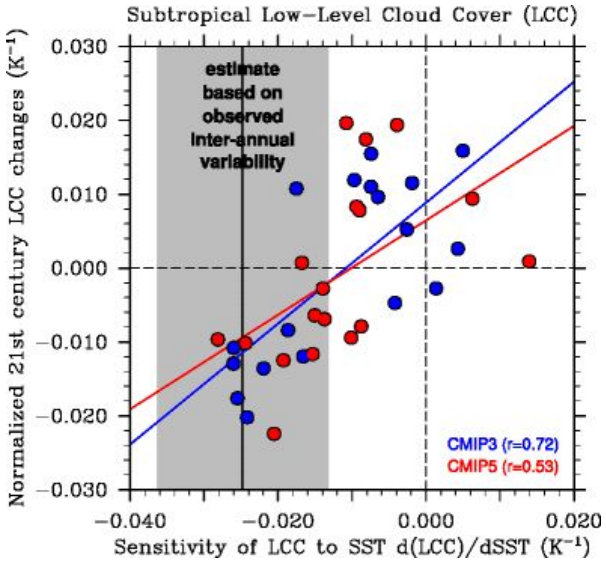
Some models are better than others (but the winner depends on what you look at…)



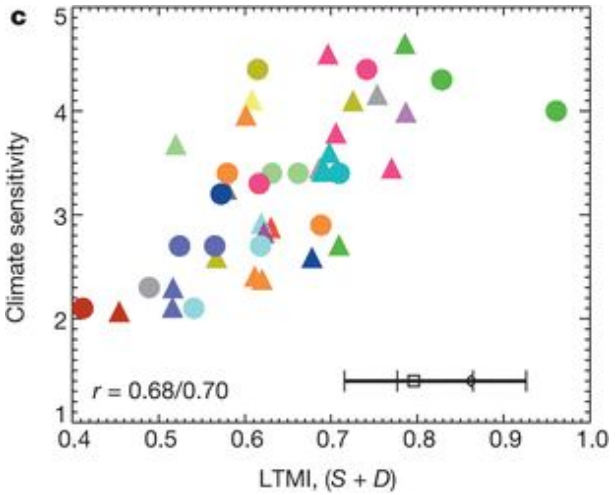
Sanderson *et al* (2017)

Emergent constraints have yet to be posed in a comprehensive, multivariate probabilistic framework.

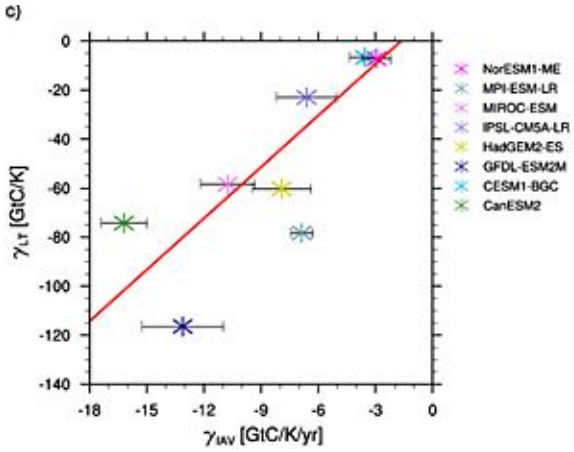
Qu et al (2014)



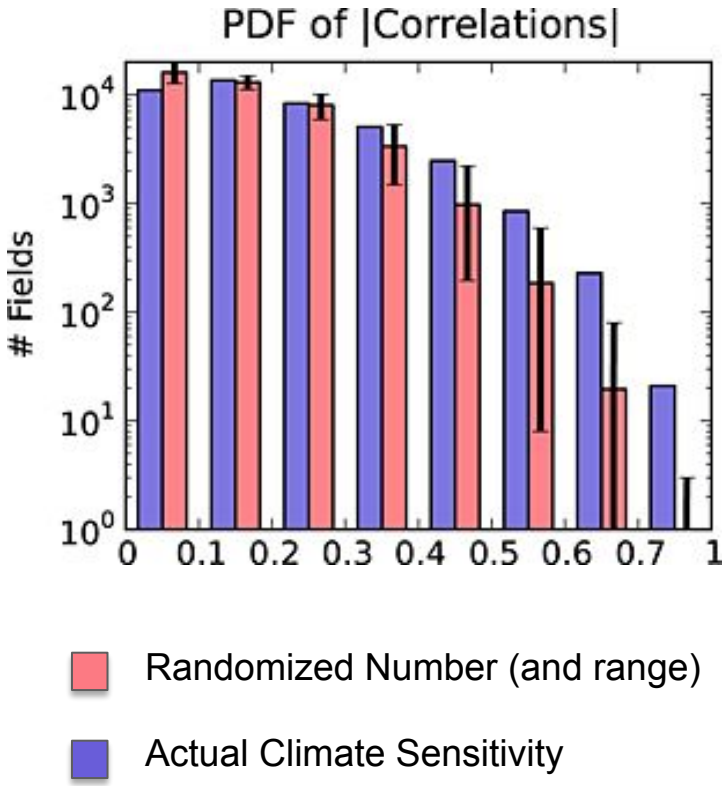
Sherwood et al (2014)



Wenzel et al (2014)

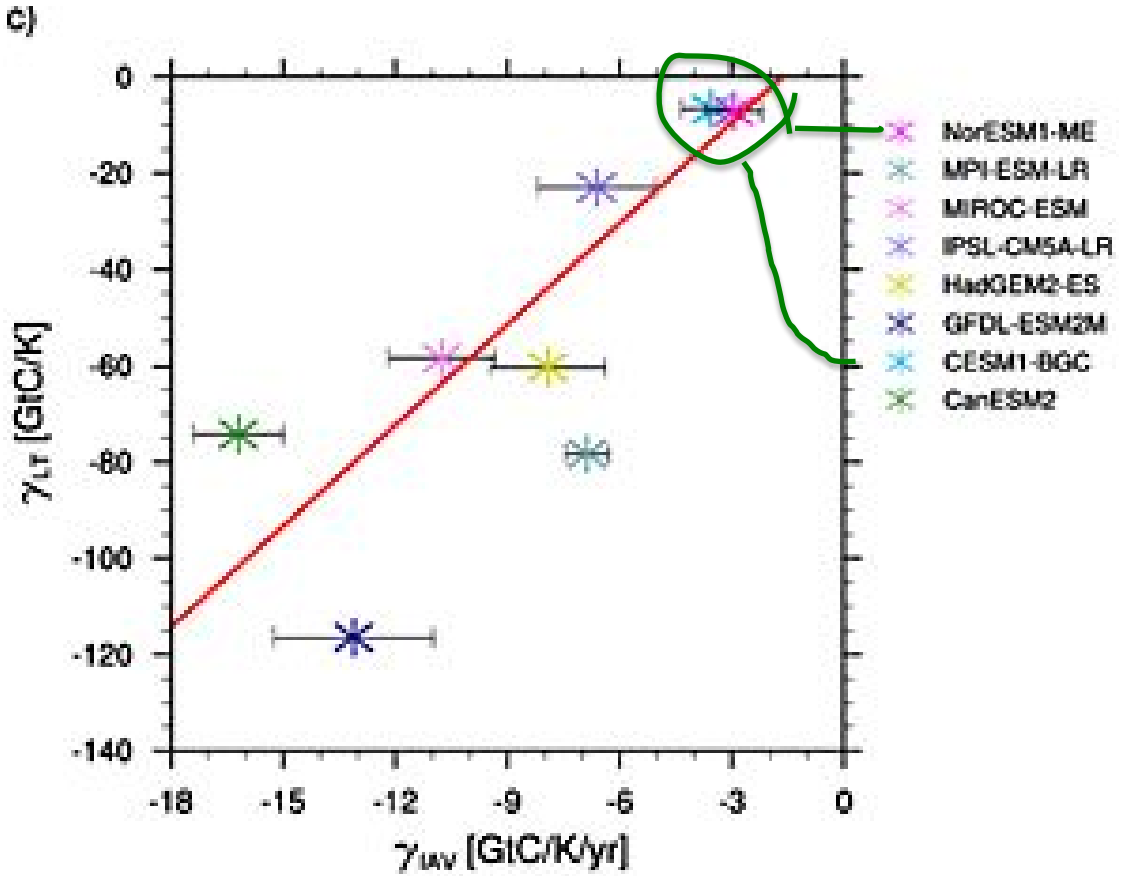


But data-mining can produce spurious constraints in a small sample



Caldwell *et al* (2014)

… and constraints can be artificially enhanced due to replication



Wenzel et al (2014)

Assumptions

Bayesian Worldview

IPCC AR5
Tebaldi et al (2005,2009)

1) More **independent** models in agreement implies more confidence in **mean** projection (reinflation and method to assess intermodel correlation needed)
2) Each model is an equally plausible representation of the true system (not generally true, tuning overstates agreement, poor models would overinflate uncertainty)
3) Present-day performance can be informative about future behavior (not generally true, general skill metrics are inconsistent, possible use of process-based constraints?)

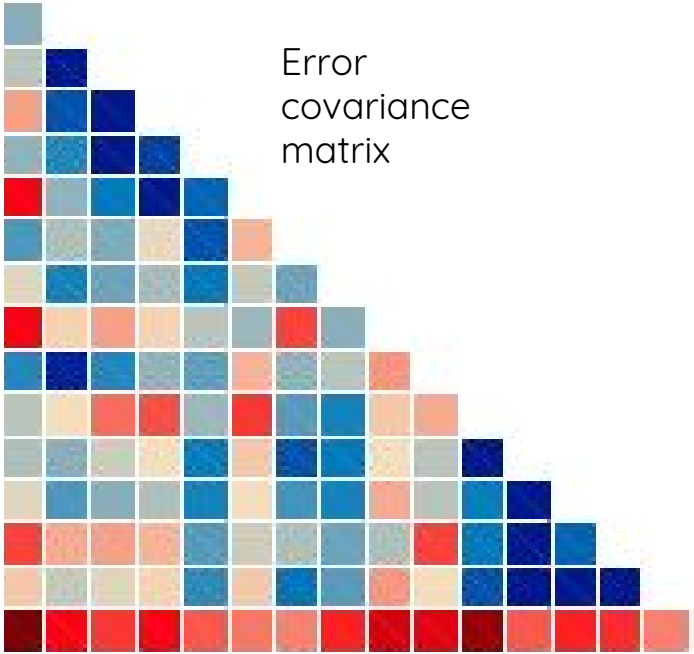
Worldviews

Replicate Earth

The ensemble as a whole is a biased representation of reality, which can be transformed to better represent the set of possible states of the true climate system.

Bishop & Abramowitz (2013)

A



Error covariance matrix

$$\mathbf{w} = \frac{\mathbf{A}^{-1}\mathbf{1}}{\mathbf{1}^{T}\mathbf{A}^{-1}\mathbf{1}}$$

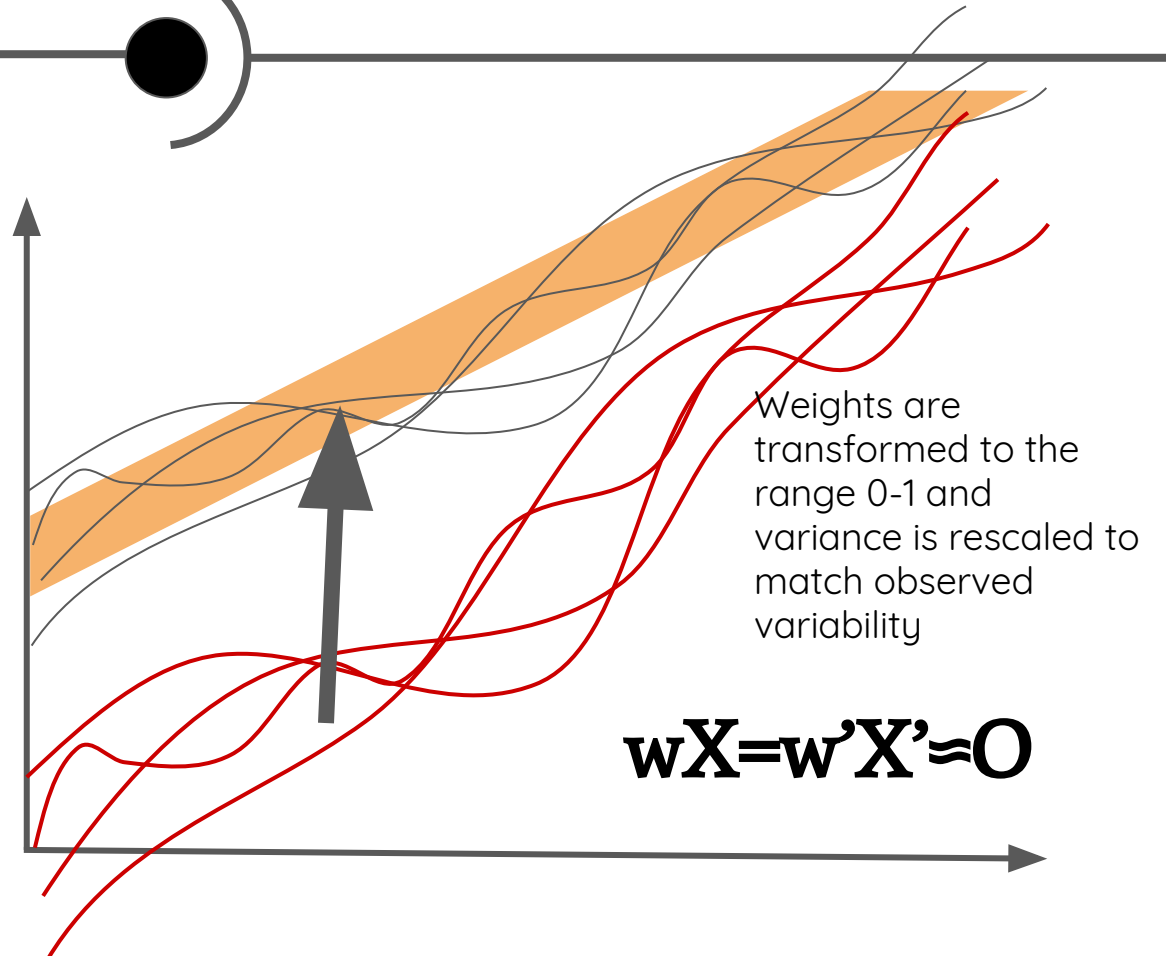
Linear estimator for optimal combination of models to fit observed trajectory (automatically accounts for interdependency)

Worldviews

Replicate Earth

The ensemble as a whole is a biased representation of reality, which can be transformed to better represent the set of possible states of the true climate system.

Bishop & Abramowitz (2013)



Assumptions

Replicate Earth Worldview

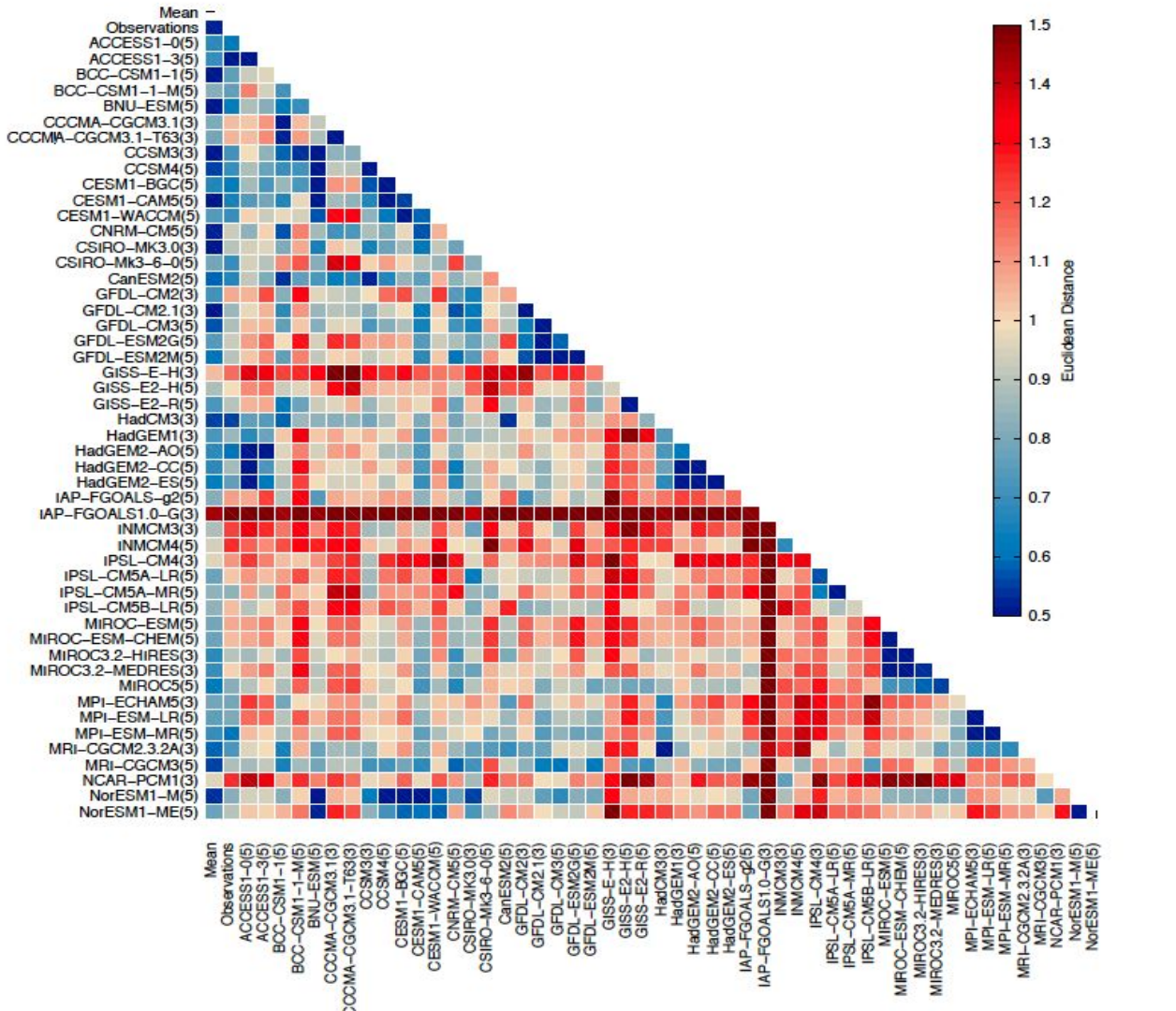
Bishop & Abramowitz (2013)

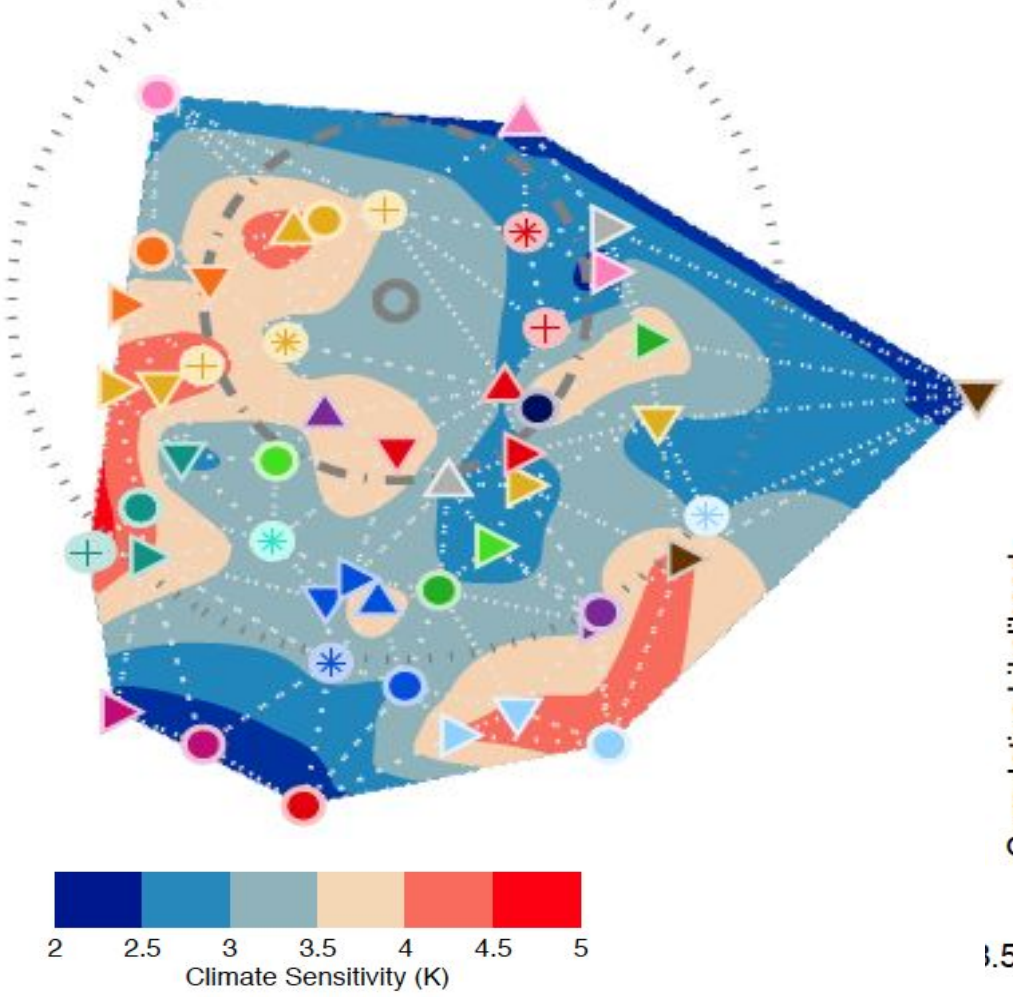
1) Models can be linearly combined without constraint to produce an optimal estimate of the true system
2) Systematic differences in simulated climate can be rescaled to represent real-world internal variability
3) Model interdependency can be accounted for through error correlation.

Geometric

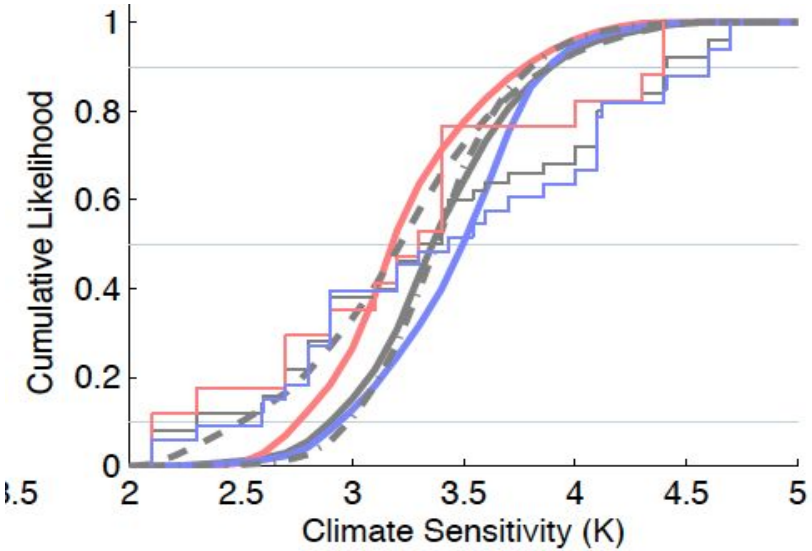
The ensemble is a partially replicated subset of a infinite set of models of varying quality, members of which can be geometrically represented, weighted and interpolated to approximate a distribution of plausible, independent models

Sanderson (2015a,b, 2017)
Knutti (2017)

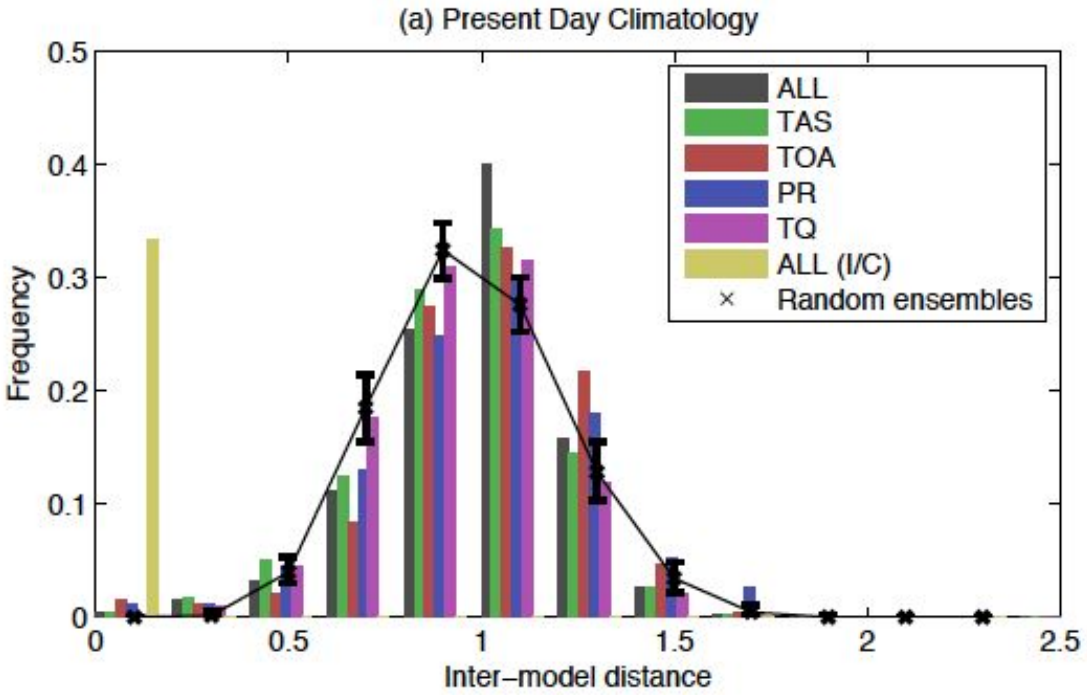


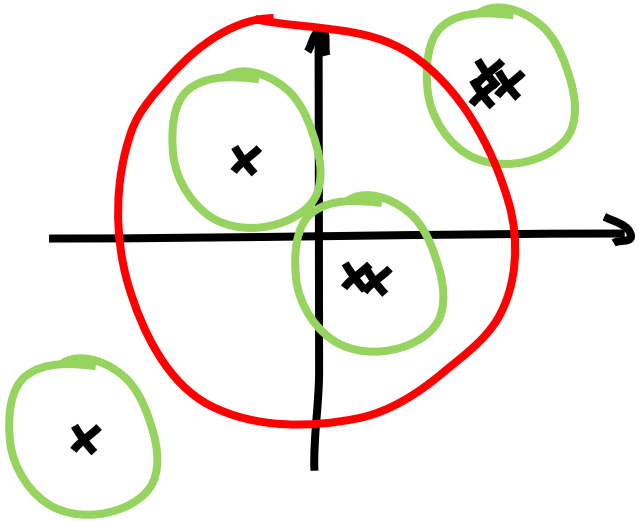


Models are represented in a similarity space, which can be resampled to produce distributions for model properties



Inter-model: Distances between CMIP mean states are **much** greater than those from initial conditions

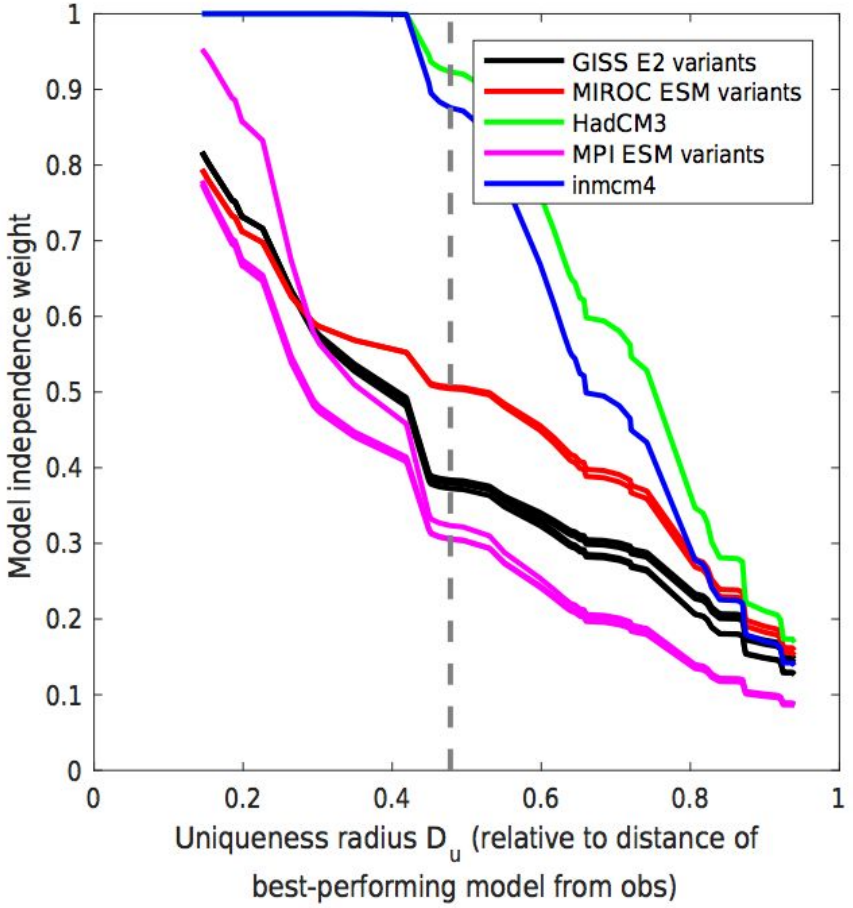


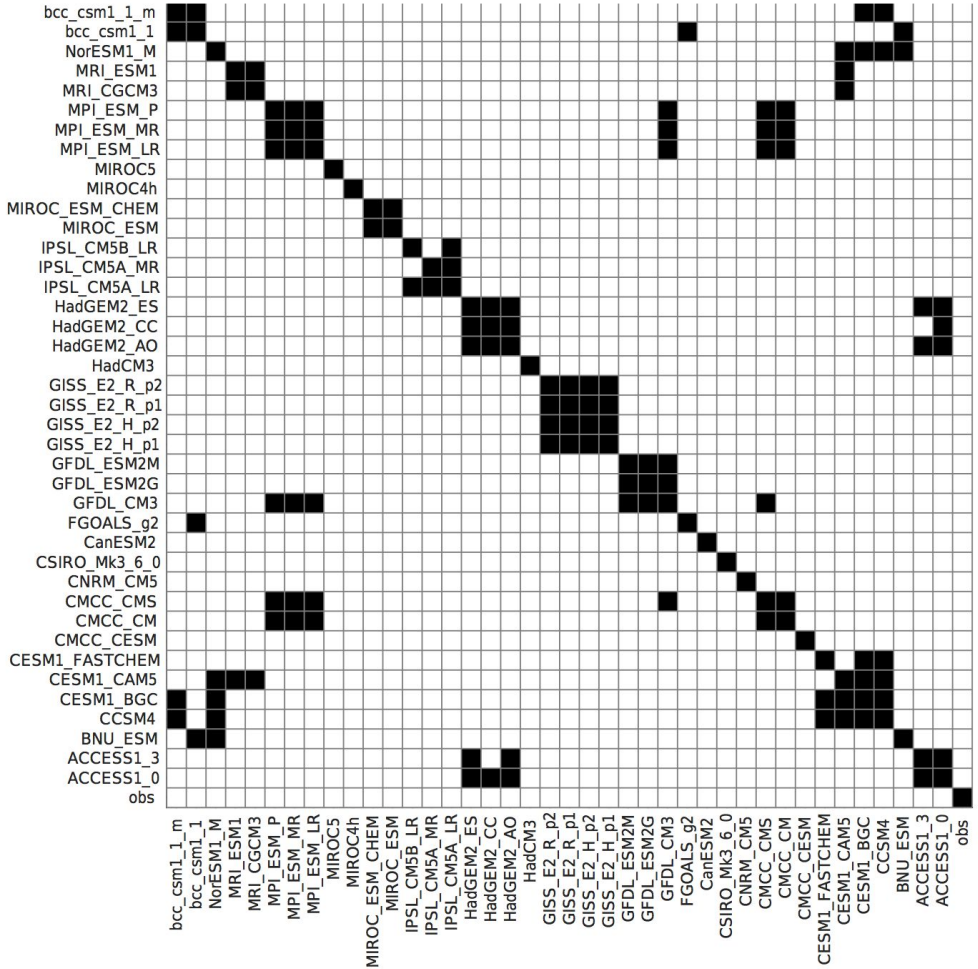


Models in immediate vicinity of other models can be downweighted. Skill is defined by RMSE.

Uniqueness weighting coefficients make intuitive sense

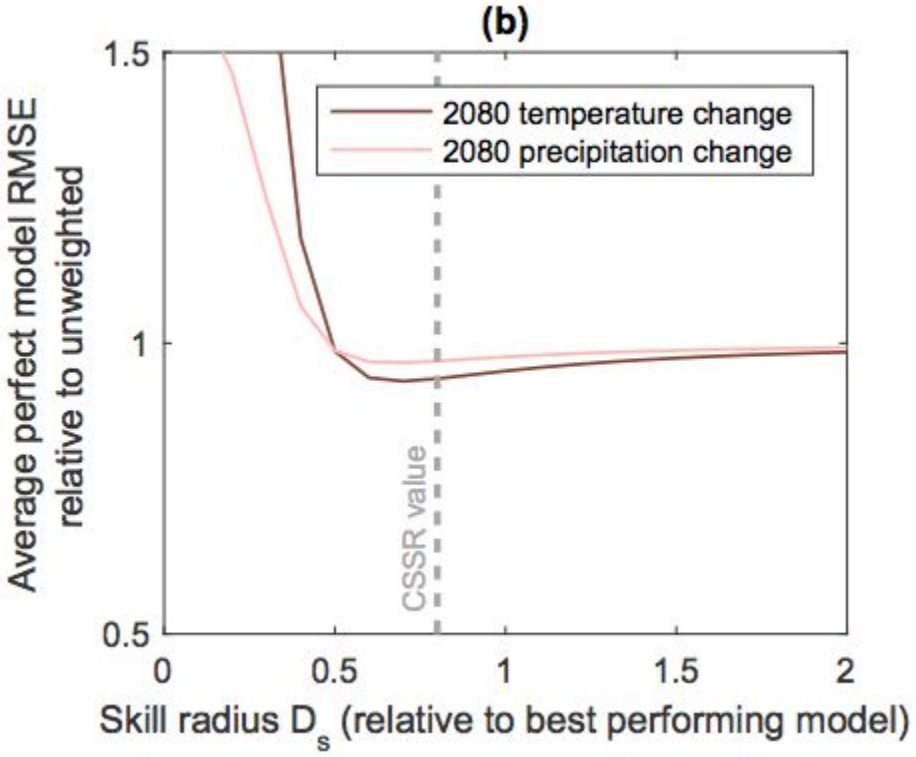
$$w(i) = \frac{e^{-\left(\frac{\delta^{20c}_{i(obs)}}{D_q}\right)^2}}{1 + \sum_{j \neq i}^{m} e^{-\left(\frac{\delta^{20c}_{ij}}{D_u}\right)^2}}$$

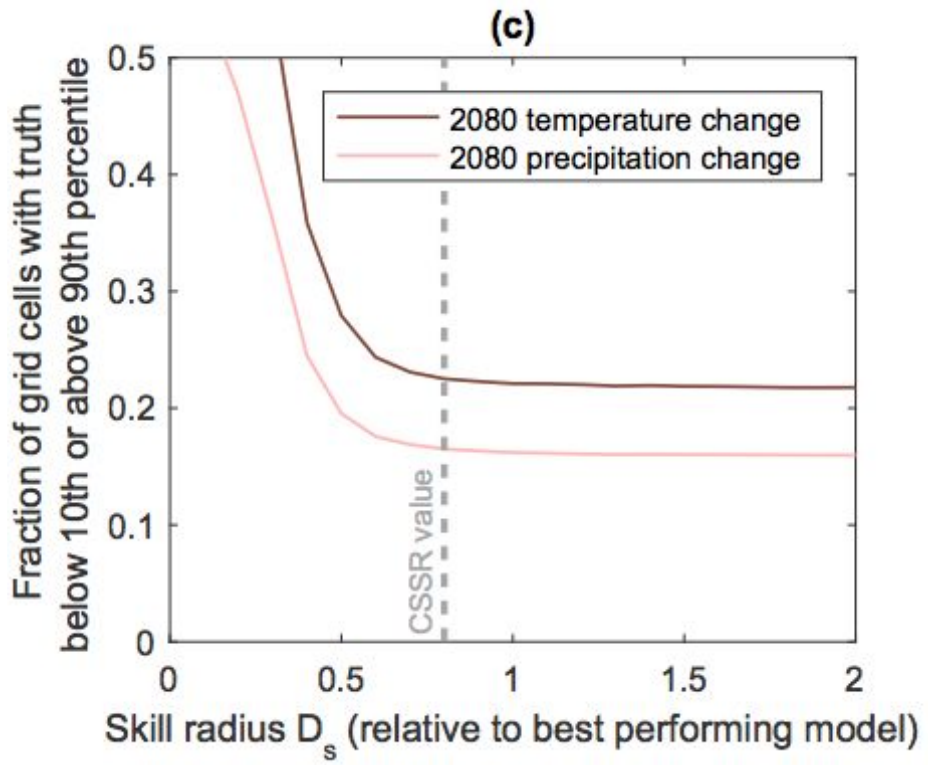


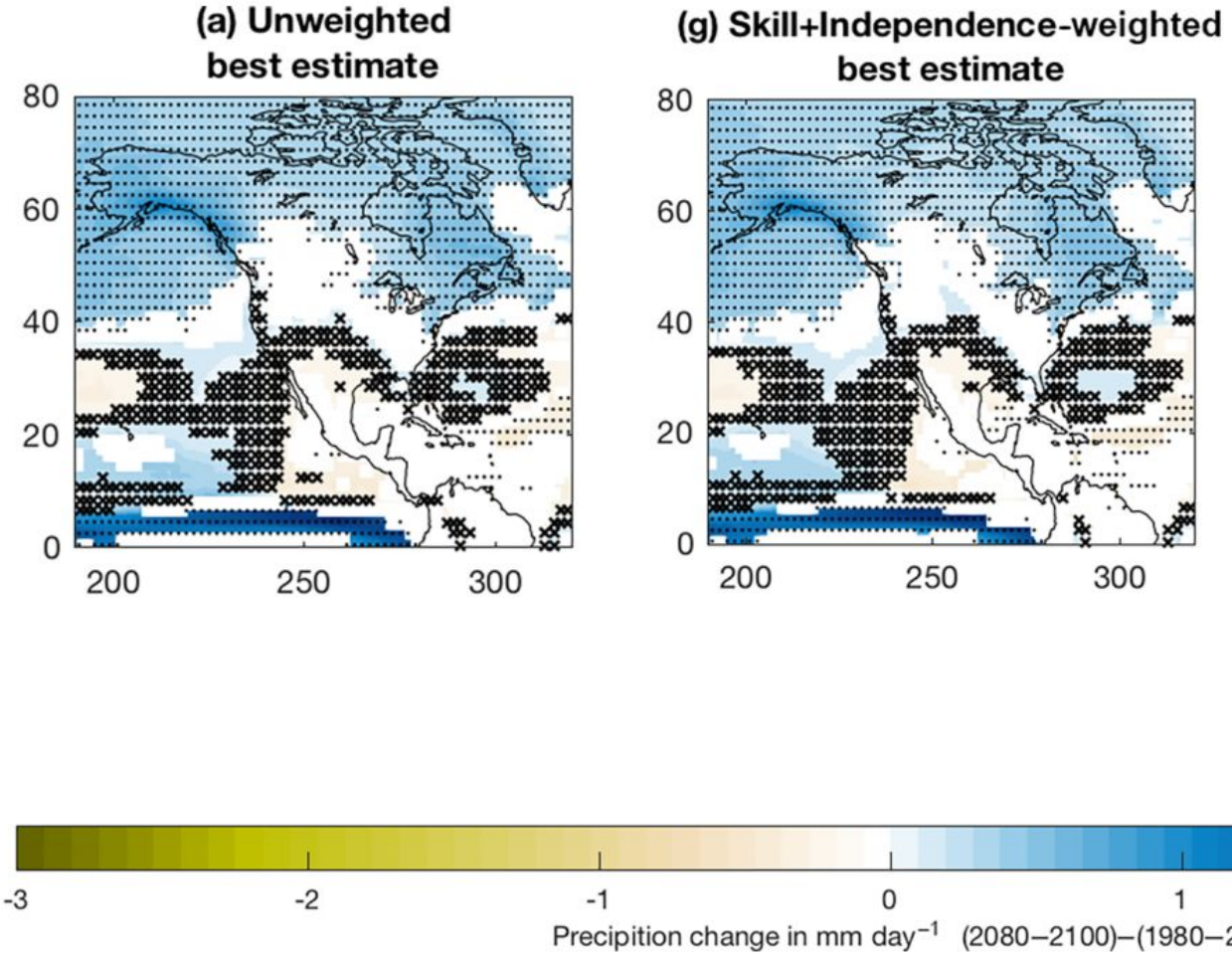


Which provides guidance for identifying strongly interdependent models which should be excluded from out of sample tests.

Skill weighting has the potential to improve out-of-sample performance, with risks…

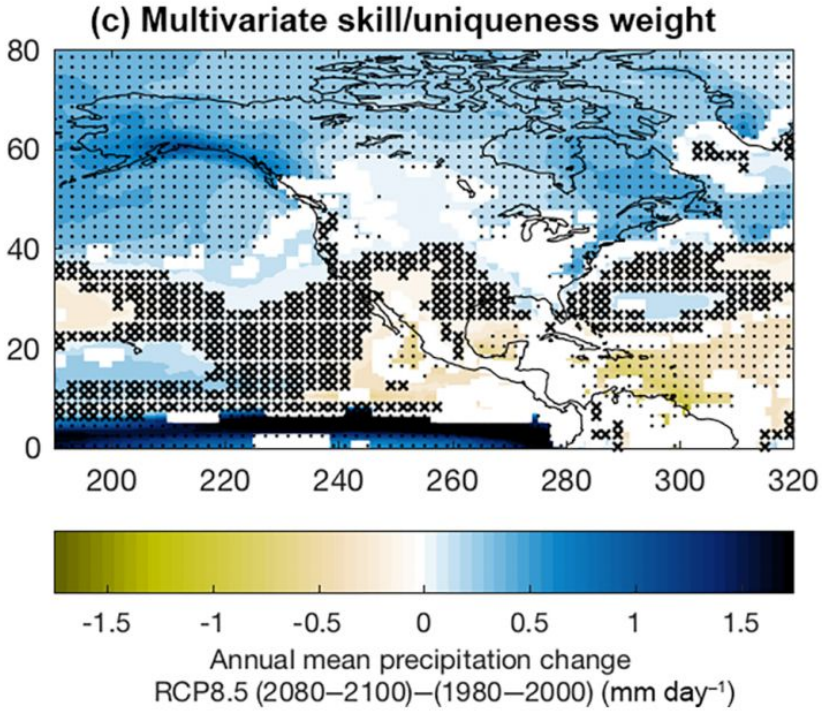


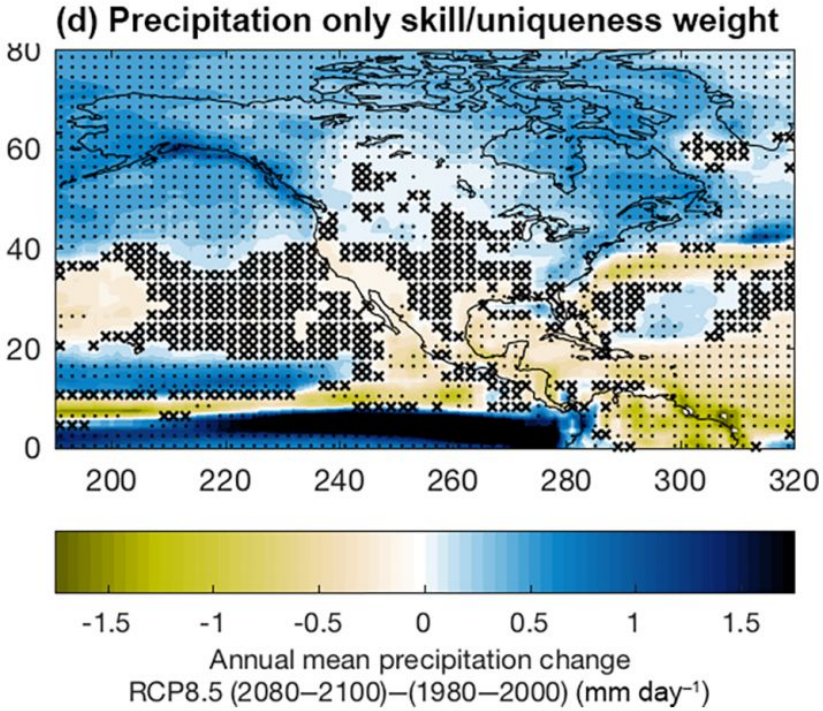




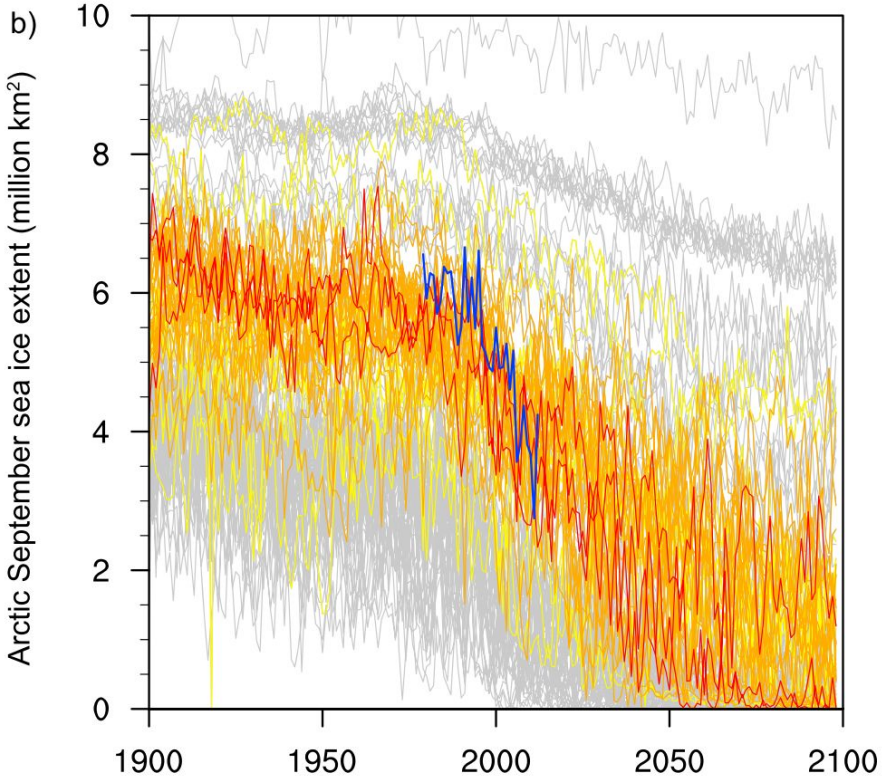
Conservative weighting with multi-variate metrics has little effect on projections

But univariate metrics have a more noticeable effect





Low dimensional metric spaces provide powerful constraints, but lose the ability to identify interdependency through error correlation



Knutti et al (2017)

Assumptions

Geometric Worldview

Sanderson (2015a,b, 2017)
Knutti (2017)

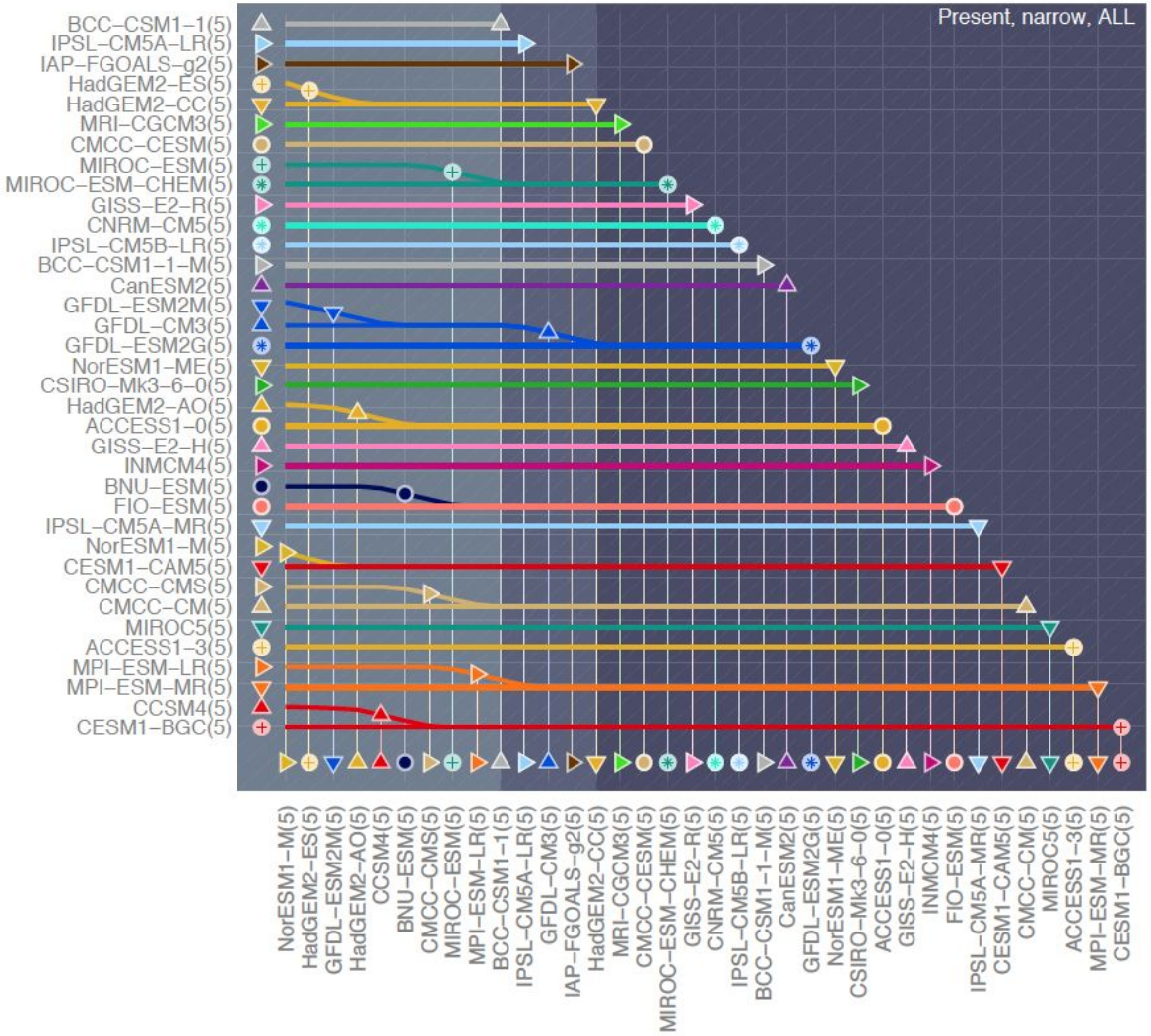
1) CMIP is a subset of possible models of the Earth system. True uncertainty would only represented by the parent distribution.
2) The distribution of inter-model and model-obs distances can provide meaningful weighting information on model interdependency and skill
3) Models and observations can be represented in a low dimensional space constructed from diagnostic output and projections can be interpolated to approximate parent distribution

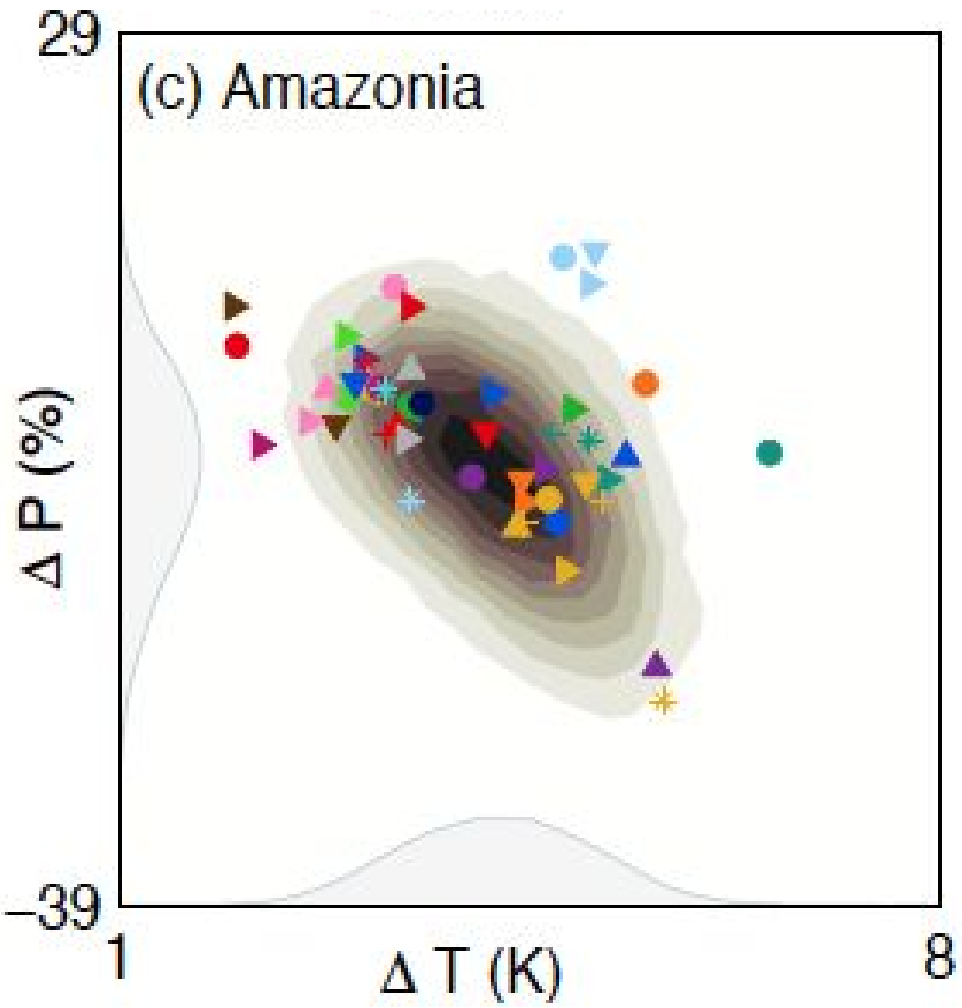
	Bayesian Worldview	Replicate Earth Paradigm	Geometric Worldview
Key strength	Established statistical framework	Remap to earth-like mean and variance	Prediction of unbiased ensemble
Addresses Interdependency?	With modification / error covariance info	Native	Native
Behavior as $N_{models} \to \infty$	Posterior converges	Underconstrained regression	Stable
Ad-hoc assumptions	Prior, precision distribution, metrics	Training period, metrics	Metrics, truncation, MDS, sampling prior
Projection within original model range	Yes (for mean)	Not necessarily	Yes
Limit of interpretations	Constraint relevance, over-confidence with common bias	variability/structure, unphysical metamodel, overconfidence	Constraint relevance, scaling/interpolation assumptions
Constraints	Anything	Same as projection	Anything

Next Steps

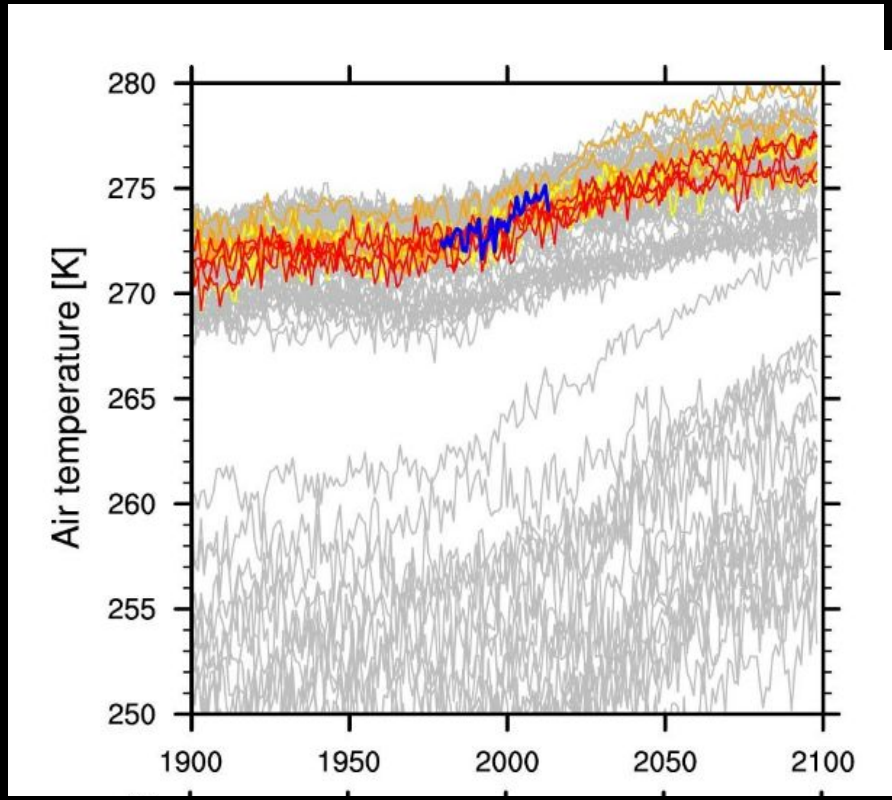
- Greater connection between the emergent constraint / process-based constraint literature and statistical literature
- Better treatment/consideration of internal variability in statistical approaches
- Out of sample testing / validation
- Assess robustness to poor models

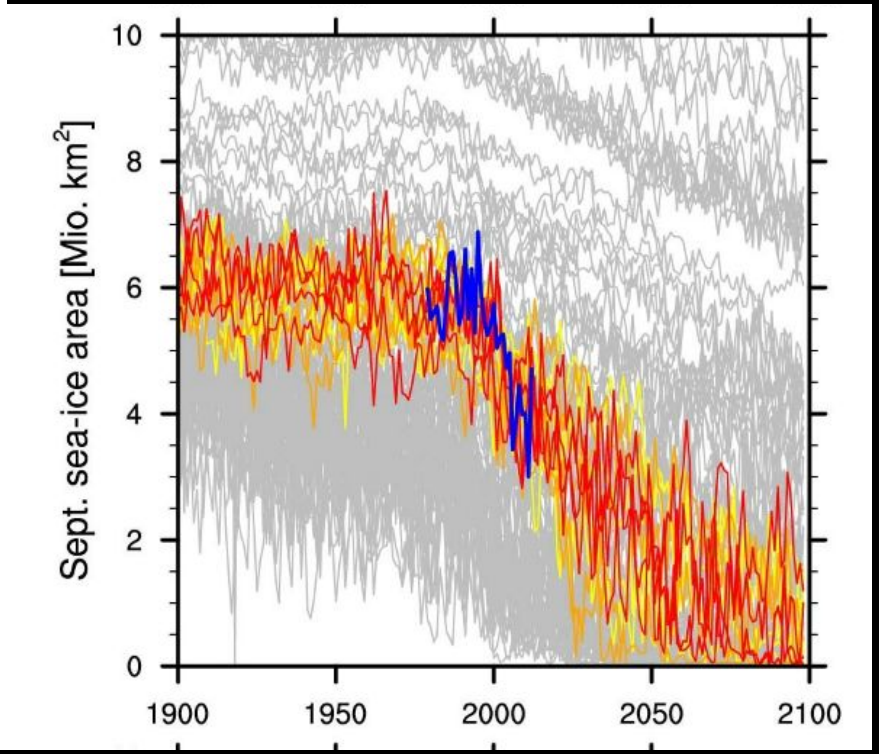
Or subsets can be created, eliminating clear duplicates first, and keeping highly performing models until last



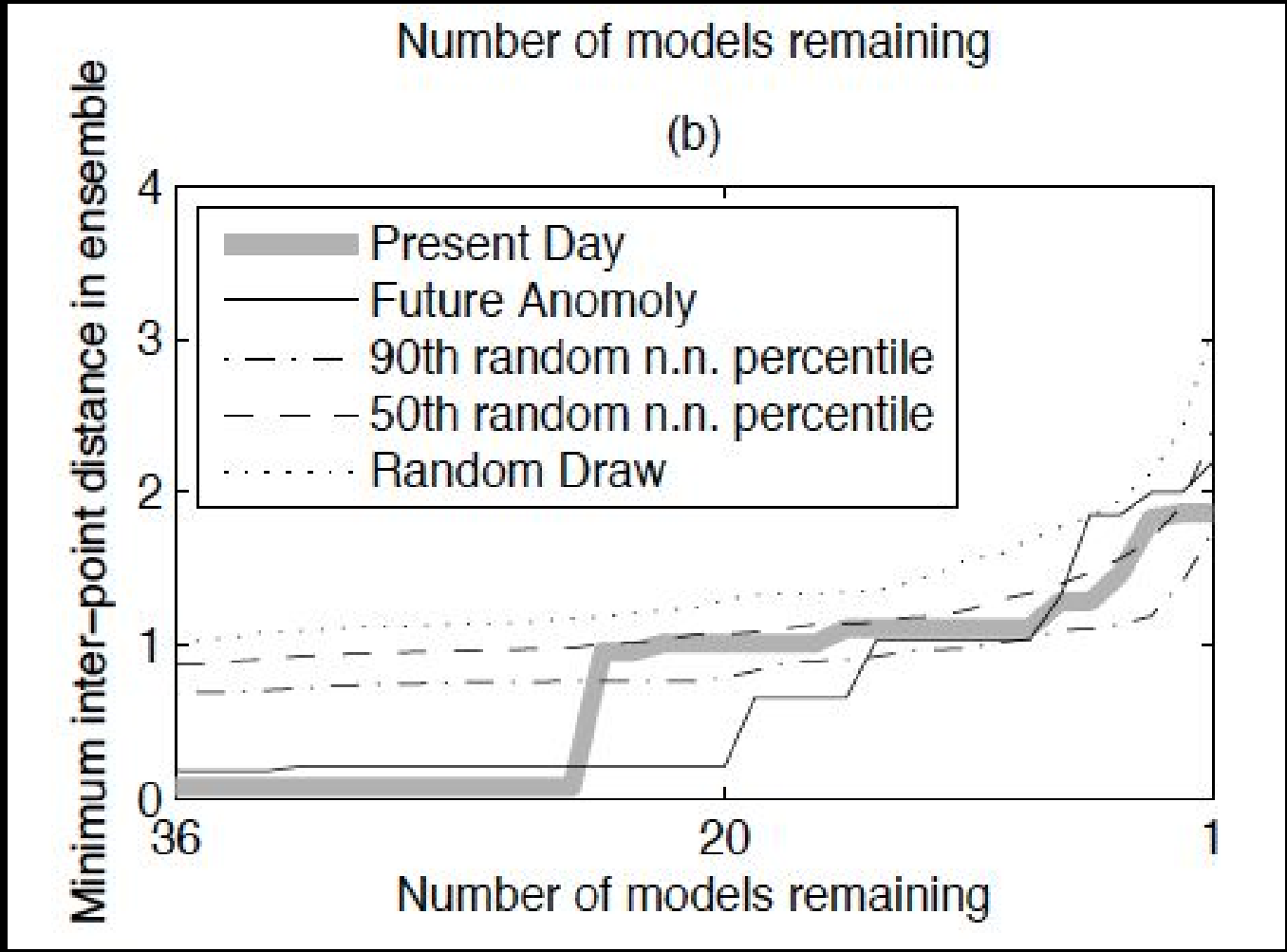


- NorESM1–ME(5)
- NorESM1–M(5)
- NCAR–PCM1(3)
- MRI–CGCM3(5)
- MRI–CGCM2.3.2A(3)
- MPI–ESM–MR(5)
- MPI–ESM–LR(5)
- MPI–ECHAM5(3)
- MIROC5(5)
- MIROC3.2–MEDRES(3)
- MIROC3.2–HIRES(3)
- MIROC–ESM–CHEM(5)
- MIROC–ESM(5)
- IPSL–CM5B–LR(5)
- IPSL–CM5A–MR(5)
- IPSL–CM5A–LR(5)
- IPSL–CM4(3)
- INMCM4(5)
- INMCM3(3)
- IAP–FGOALS1.0–G(3)
- IAP–FGOALS–g2(5)
- HadGEM2–ES(5)
- HadGEM2–CC(5)
- HadGEM2–AO(5)
- HadGEM1(3)
- HadCM3(3)
- GISS–E2–R(5)
- GISS–E2–H(5)
- GISS–E–H(3)
- GFDL–ESM2M(5)
- GFDL–ESM2G(5)
- GFDL–CM3(5)
- GFDL–CM2.1(3)
- GFDL–CM2(3)
- CanESM2(5)
- CSIRO–Mk3–6–0(5)
- CSIRO–MK3.0(3)
- CNRM–CM5(5)
- CESM1–WACCM(5)
- CESM1–CAM5(5)
- CESM1–BGC(5)
- CCSM4(5)
- CCSM3(3)
- CCCMA–CGCM3.1–T63(3)
- CCCMA–CGCM3.1(3)
- BNU–ESM(5)
- BCC–CSM1–1–M(5)
- BCC–CSM1–1(5)
- ACCESS1–3(5)
- ACCESS1–0(5)
- Narrow Std.Dev,
- Gaussian Std.Dev
- I/C Variability

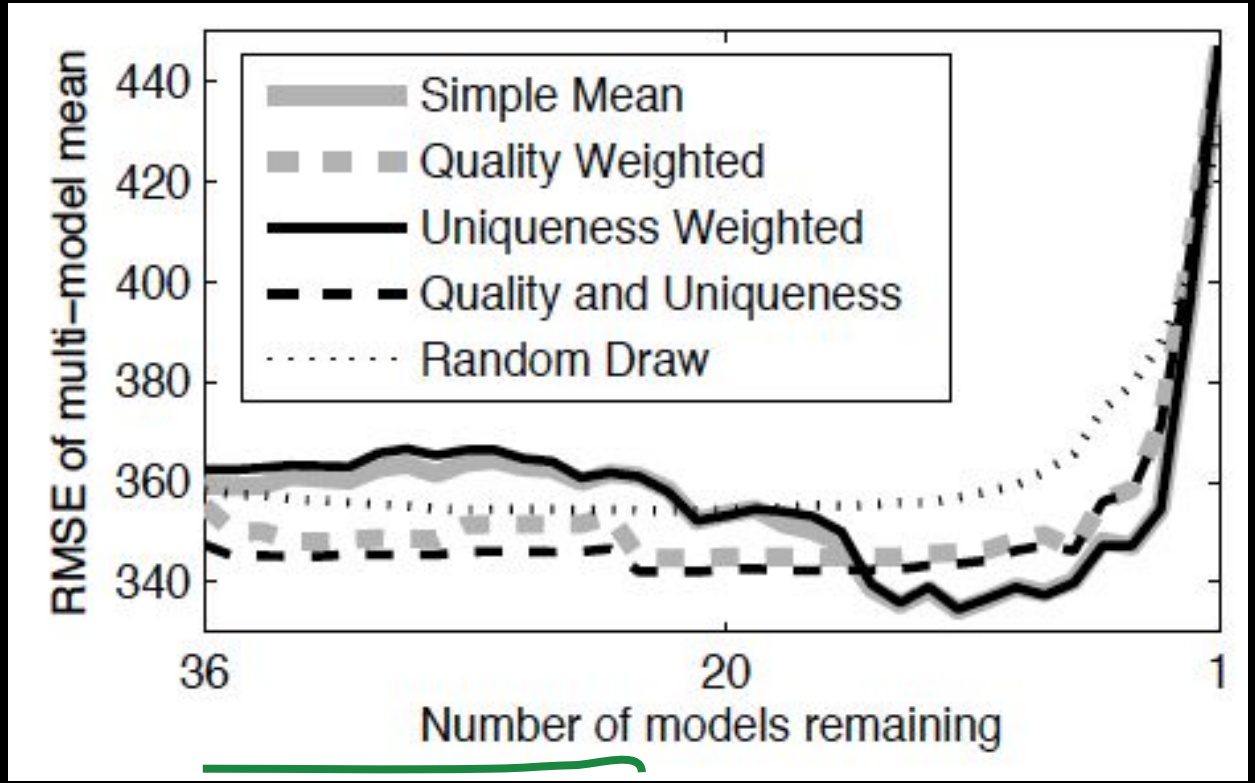

Air temperature [K]
280
275
270
265
260
255
250
1900
1950
2000
2050
2100


Sept. sea-ice area [Mio. km²]
10
8
6
4
2
0
1900
1950
2000
2050
2100

But when to stop?



But when to stop?



Removing replicates

Removing poor performers

Removing better models