Physical systematic biases

Aspen Model Evaluation Workshop

Greg Flato
Canadian Centre for Climate Modelling and Analysis

August, 2017

Issues

- Many large-scale errors/biases persist from one generation of models to another; if they were well understood and easy to fix, they wouldn't persist.
- Our ability to quantify these biases depends on observational quality and on the confounding effects of internal variability.
- Will more systematic evaluation help? How much do errors/biases affect model fitness for purpose? Can we focus on the 'essential' ones? Is weighting/selection really a good idea? Is post-processing (e.g. bias correction) helpful?

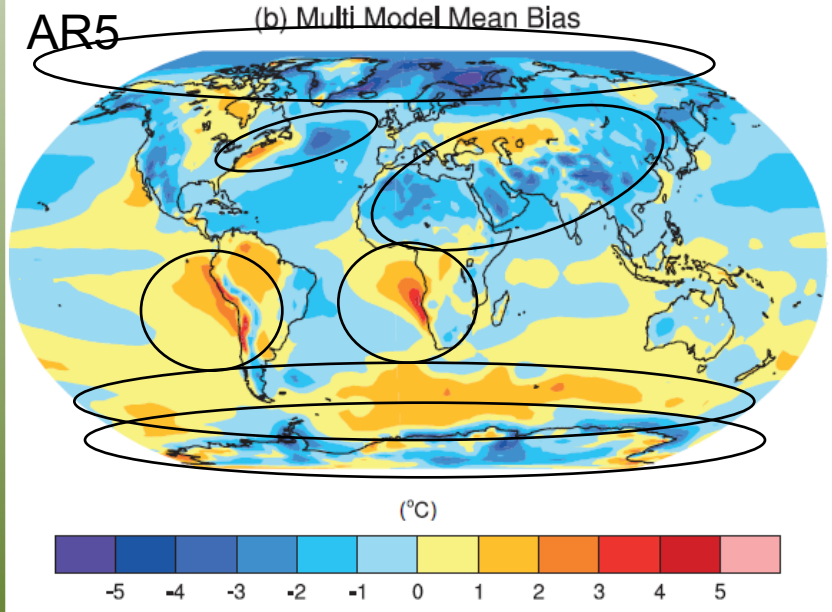
IPCC Assessments …

- … have shown some aspects of persistent, systematic errors, and temperature is indicative.

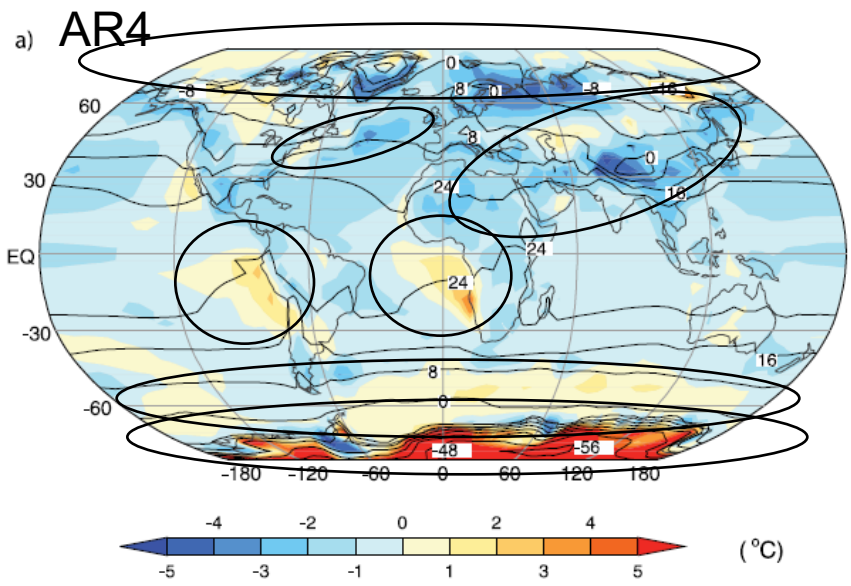
Not all systematic errors are persistent

Annual mean surface air temperature

AR5



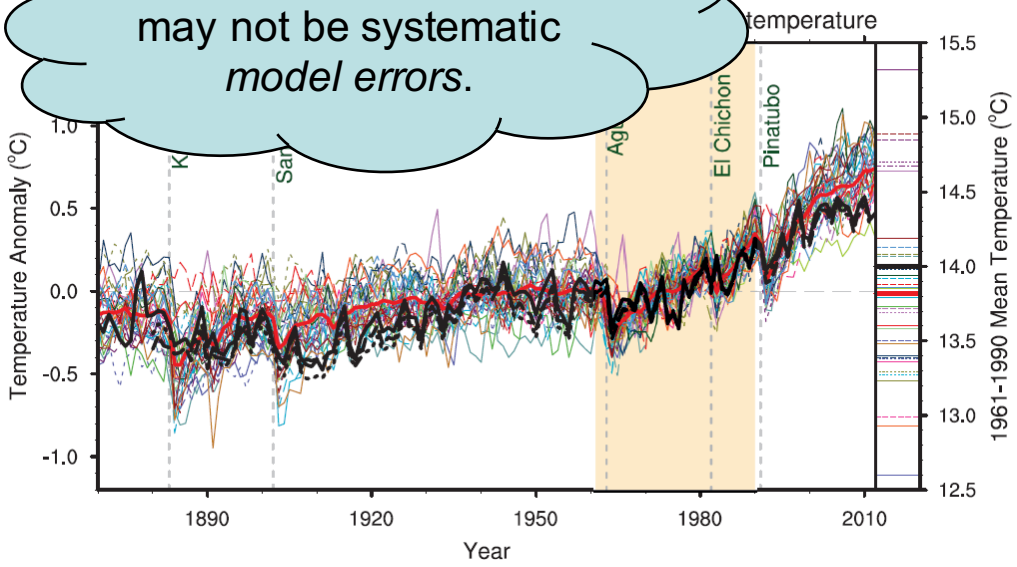
AR4

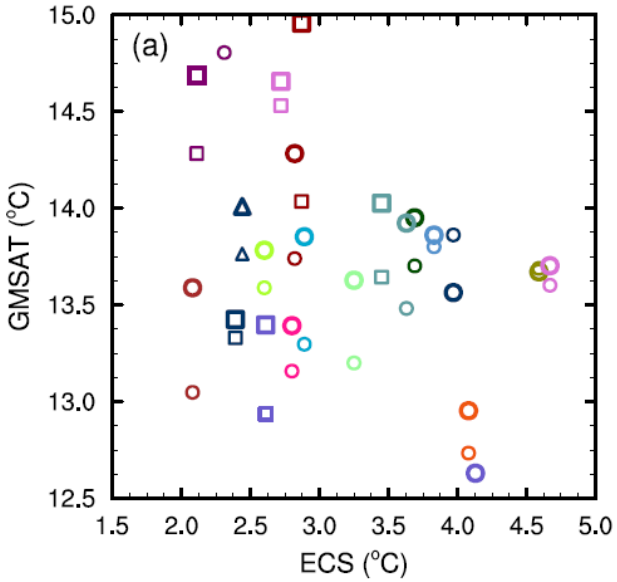


Not all errors or biases are of equal importance – different applications will place different demands on model quality.

- How important are these biases?
- Some you think would be important, don't seem to be.

Systematic *differences* may not be systematic *model errors*.



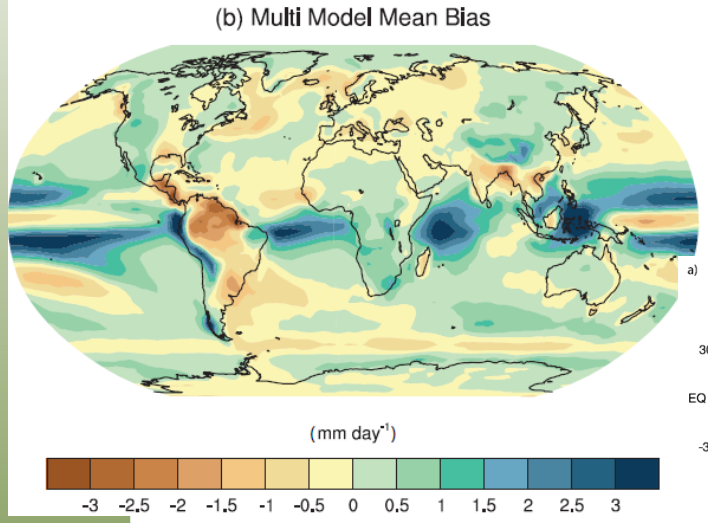


IPCC AR5, Ch. 9

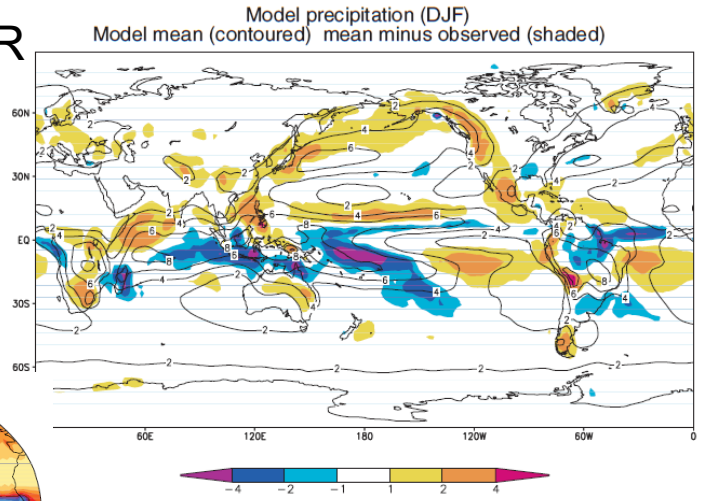
For some quantities, like precipitation, changing presentation format in IPCC reports makes tracking systematic errors difficult

The famous 'double ITCZ'

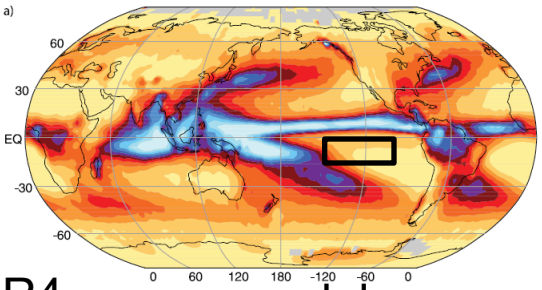
AR5



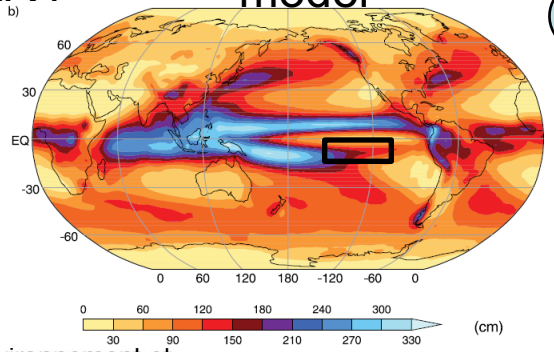
TAR



obs



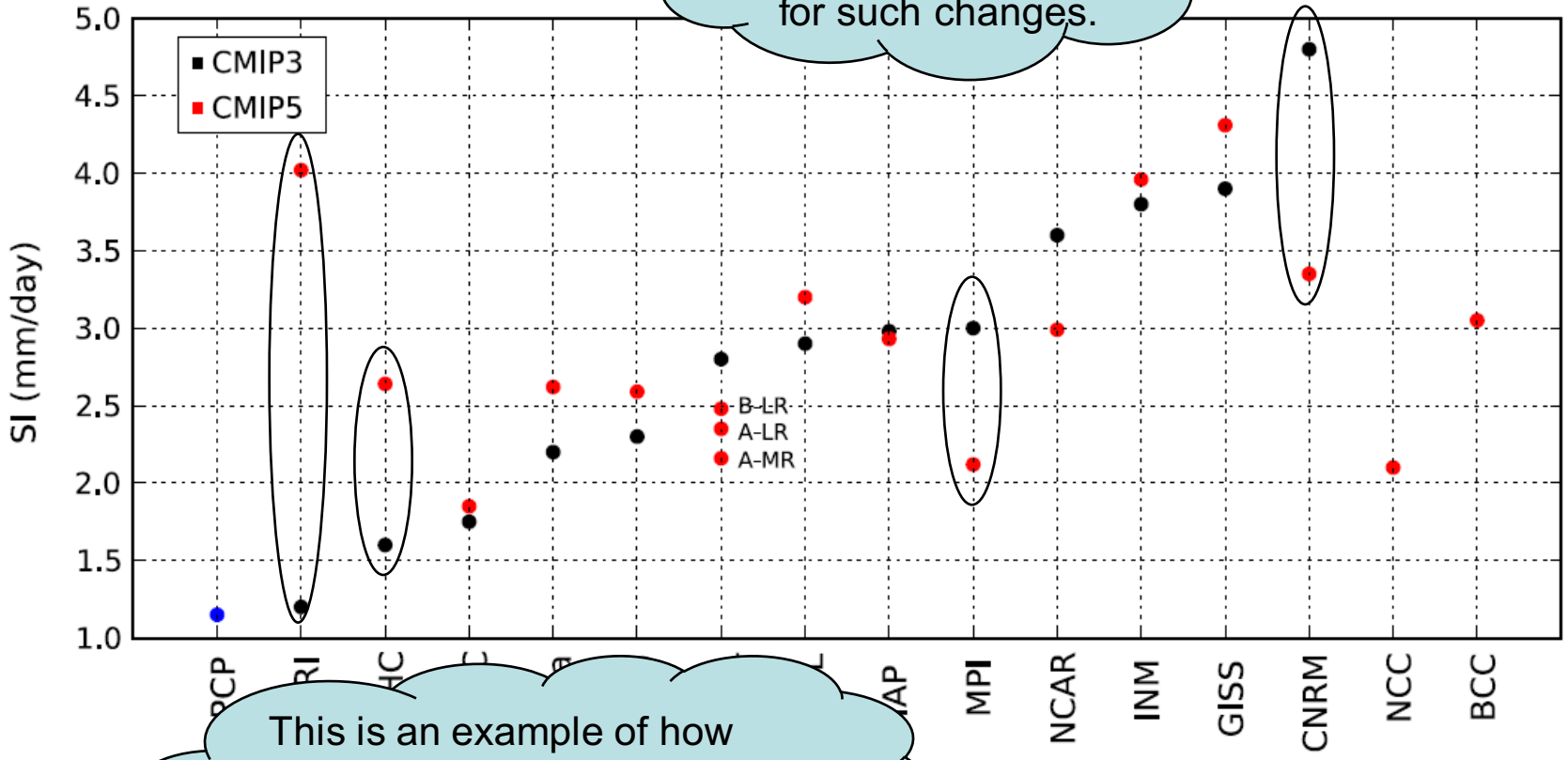
AR4 model



More systematic evaluation tools (applied retrospectively where possible) would improve this.

Little evidence for improvement in this particular bias

Better model documentation might help identify reasons for such changes.



…eslati and Bellon, Clim. Dyn., 2015

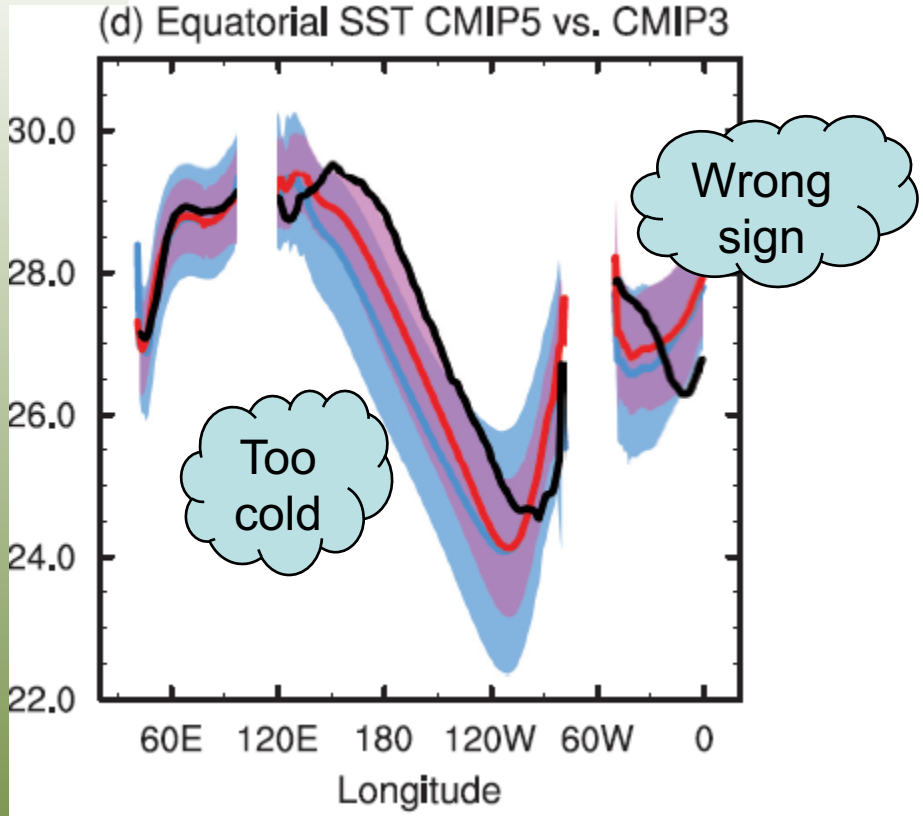
This is an example of how more 'automated' diagnostics could benefit modelling groups

Other systematic errors (list courtesy of Ron Stouffer):

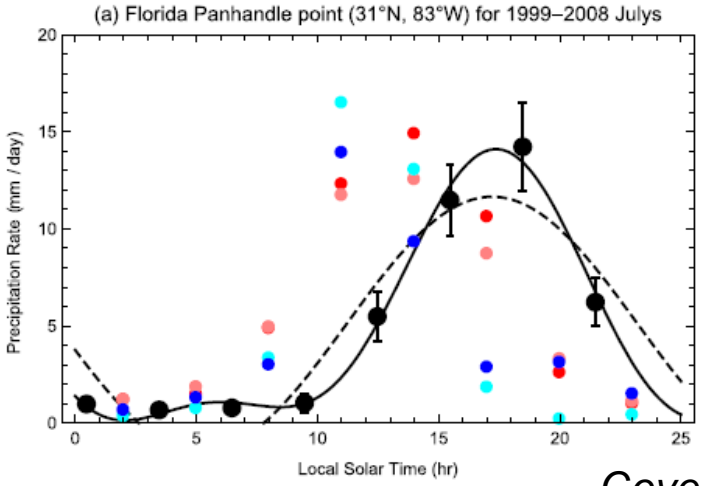
1. Double ITCZ
2. Clouds in general,
    a. low clouds over low latitude oceans
    b. radiative properties
3. Oceanic heat uptake – Southern Ocean in particular
4. Sea ice – both hemispheres (coverage and thickness and seasonal cycle)
5. Precipitation over land, river flow
6. Diurnal cycle – T, P
7. Oceanic mixing both near surface and in regions of topography;
    a. horizontal and vertical => eddies, tides and etc.
8. Magnitude of oceanic boundary flows
9. AMOC strength and poleward heat transport (model mismatch with obs)
    a. Model typically have too large volume transport and too little heat transport
    b. High resolution may improve things

Diurnal cycle of Precipitation

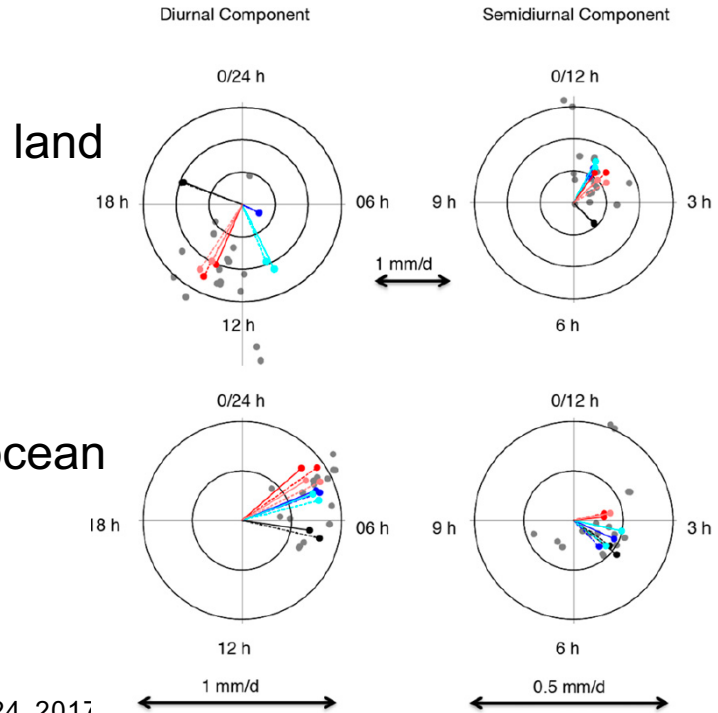
Sea-Surface Temperature



IPCC AR5, Ch. 9



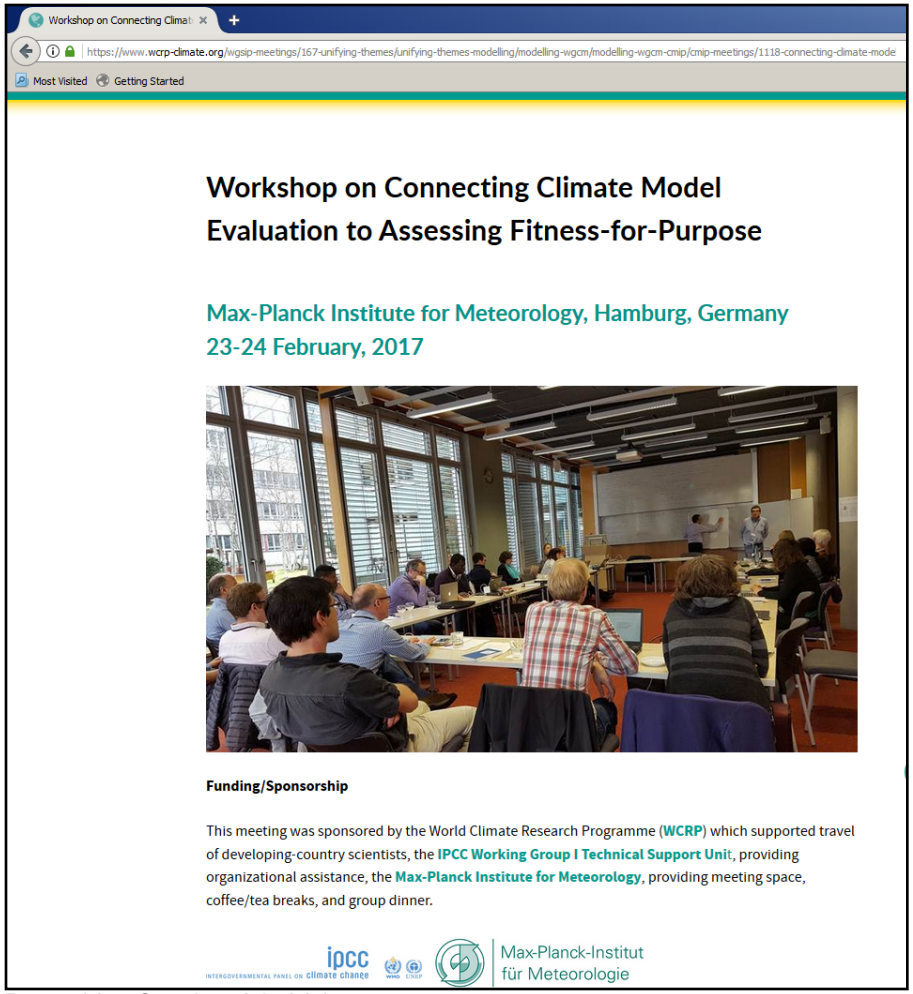
Covey et al., 2016



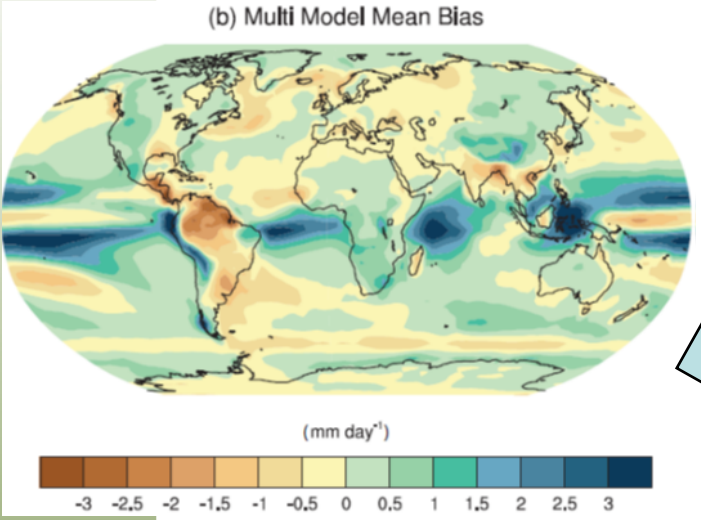
Lots of new diagnostic methods, new observations, new errors identified ....

We need to go beyond 'documenting' model errors/biases.

In the IPCC 5th Assessment we struggled to make a connection between the model evaluation chapter and the later chapters where models were used (in D&A, in future projections, in driving RCMs, etc.)

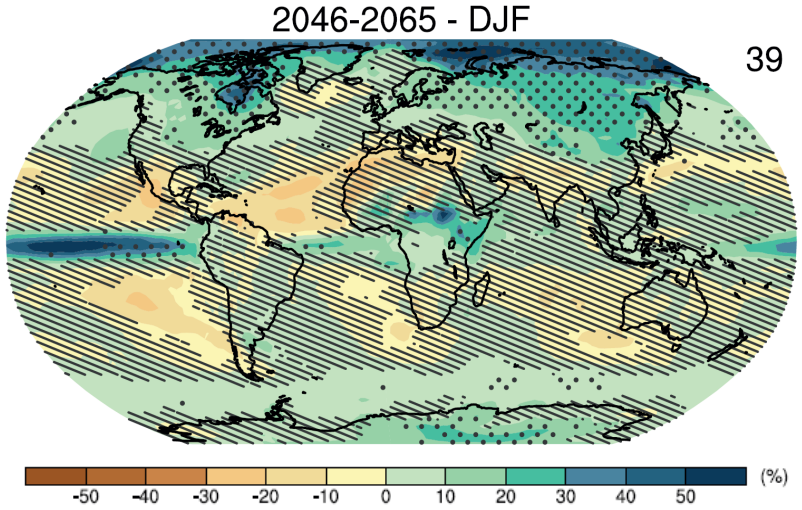


In other words … we have a growing suite of performance measures, but what we'd like to know is how errors/biases affect, e.g., a model projection of future change.



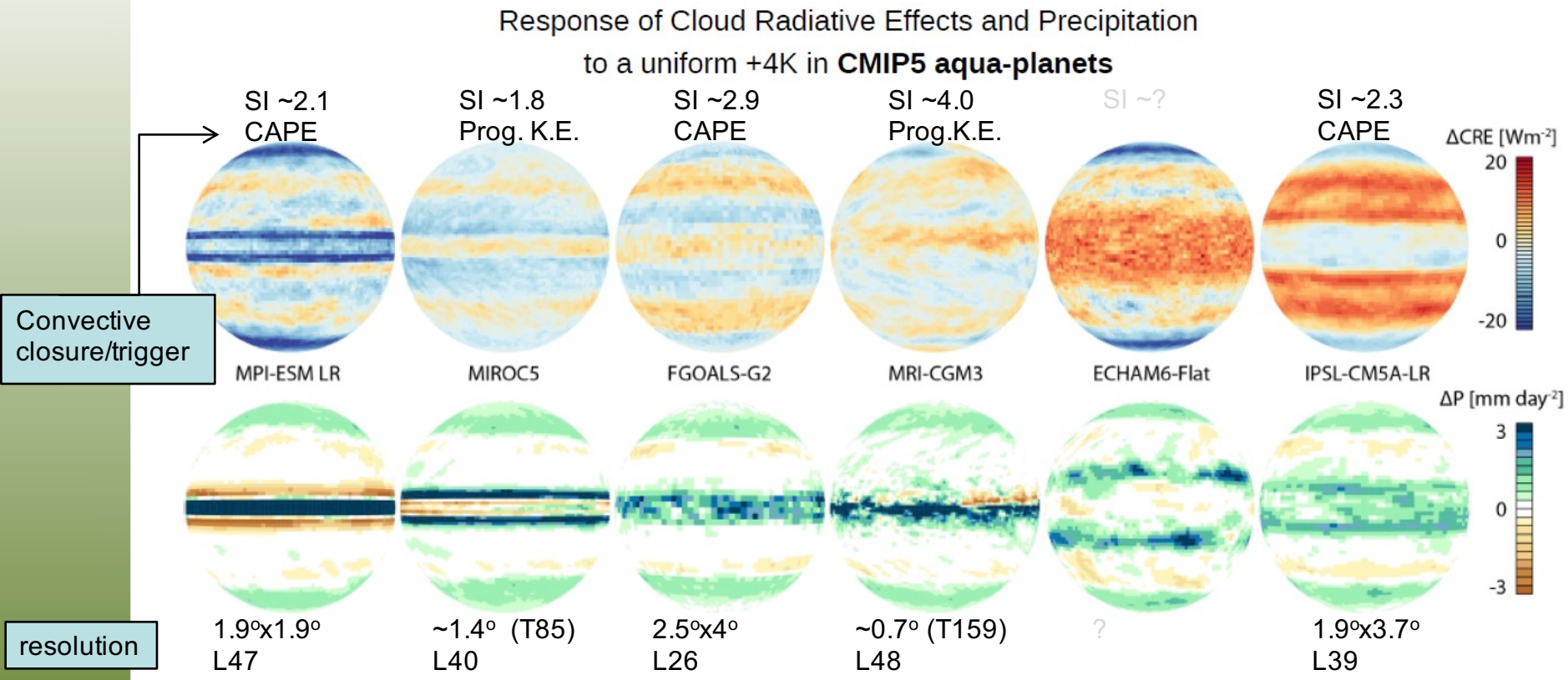
?

Multi-model mean precipitation change, relative to 1985-2005 (RCP8.5).



IPCC AR5, Ch 12

Can we get some insight from idealized experiments?

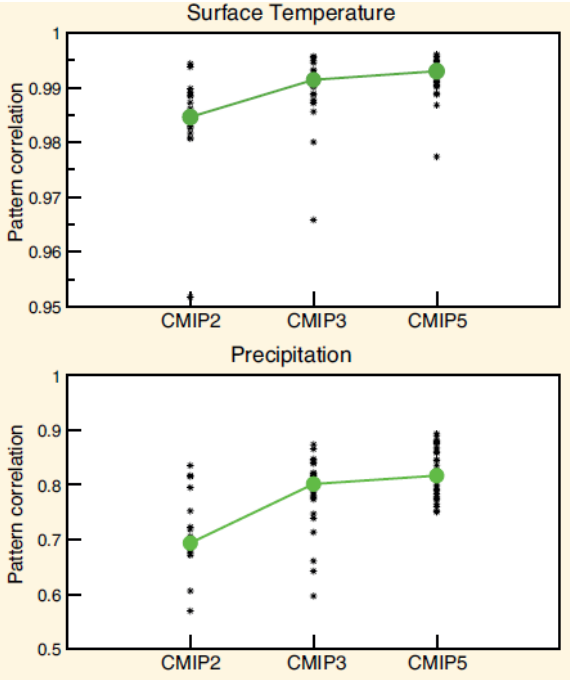


Presentation by Bony and Stevens, 2013.

- So, we have systematic biases that have persisted from one generation of model to the next. They represent *hard* problems and compelling model development challenges.

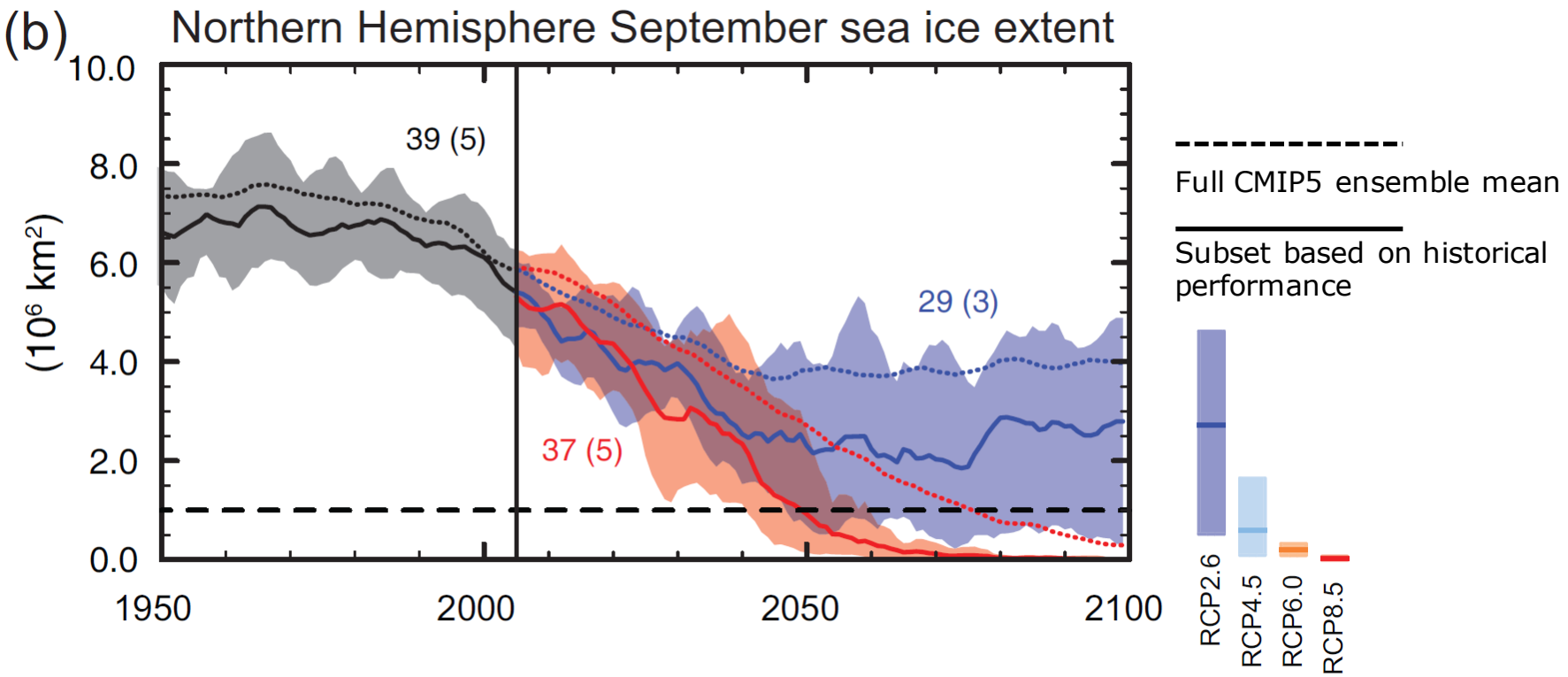
- But, many of the diagnostics we have to not clearly identify the root cause or even the processes involved, and so model improvement is slow (though visible).

- So we have to explore some other options in parallel.

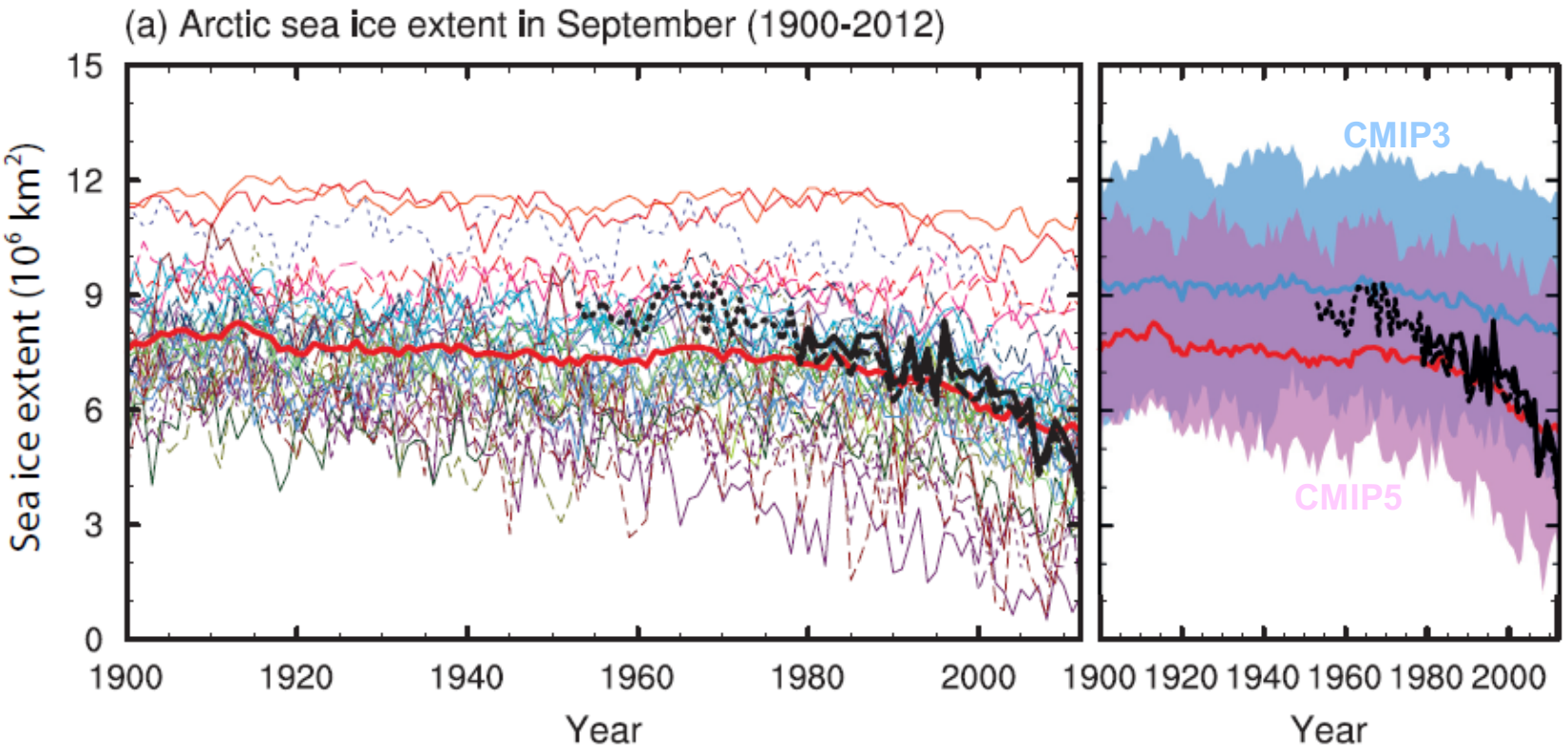


IPCC AR5, Ch. 9

Model sub-setting was used in the IPCC AR5 for Arctic sea-ice extent, based on a paper by Massonnet et al. (2012).

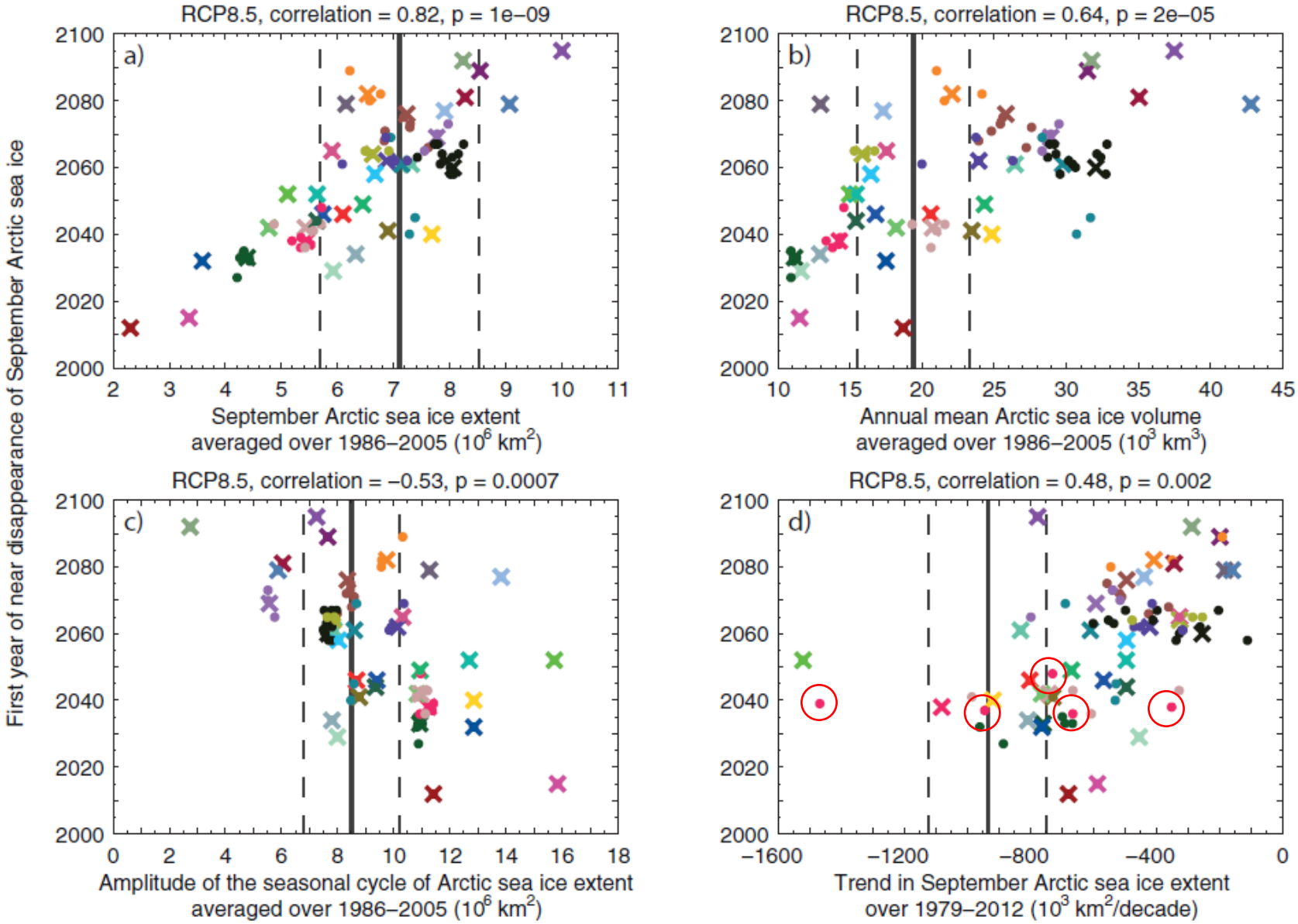


IPCC AR5, SPM (based on Ch.12)


(a) Arctic sea ice extent in September (1900-2012)
Sea ice extent (10^6 km^2)
15
12
9
6
3
0
1900
1920
1940
1960
1980
2000
Year
CMIP3
CMIP5
1900 1920 1940 1960 1980 2000
Year

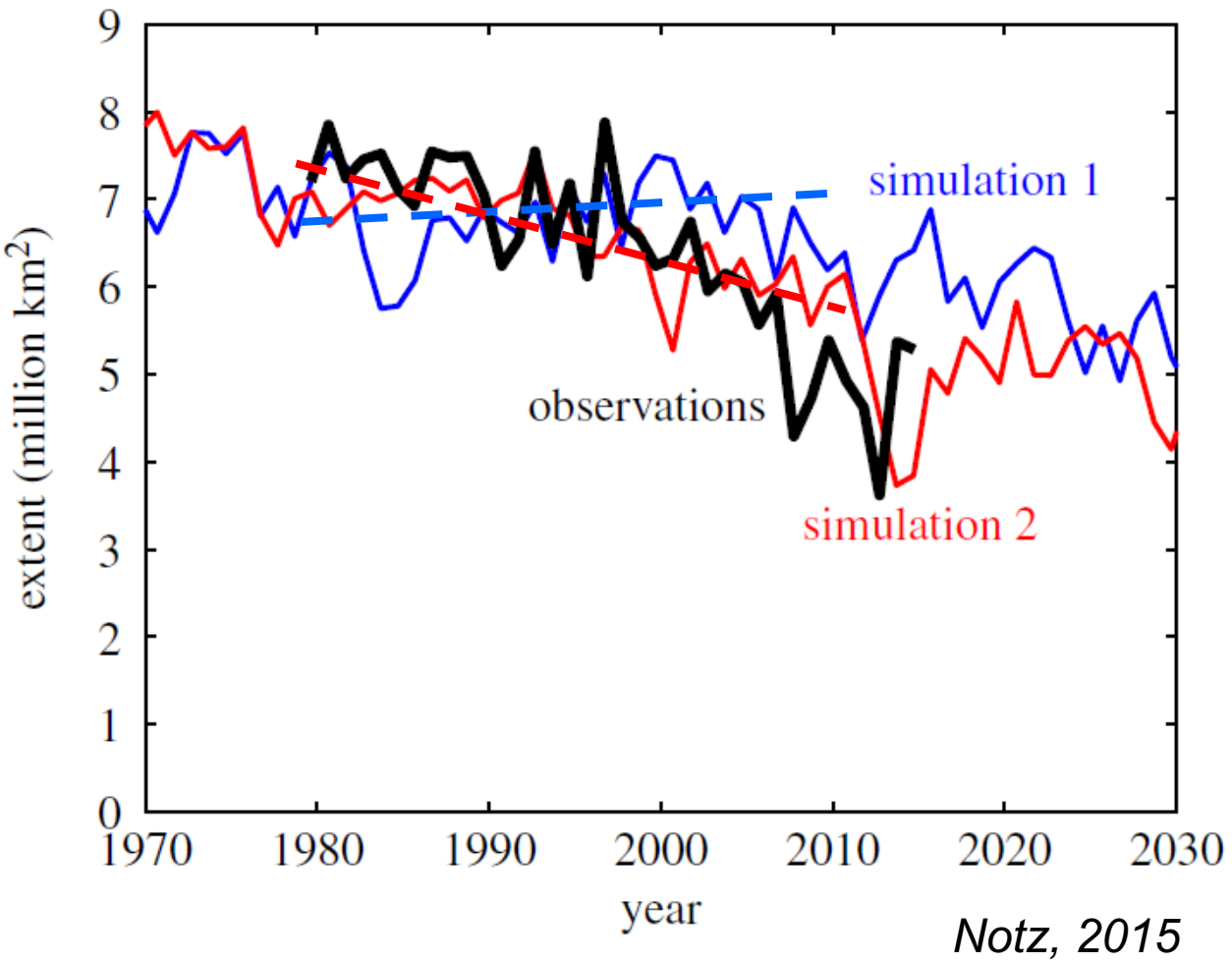
IPCC AR5, Ch. 9

- Systematic bias in *mean* and *trend* improved from CMIP3 to CMIP5.
- But large spread remains.
- And that directly impacts our ability to use model output to estimate time at which Arctic will become essentially ‘ice free’ (an important, impact and policy relevant issue).



IPCC AR5, Ch 12

An example of internal variability can completely confound a performance metric.



Notz, 2015

Large ensembles make this very clear …

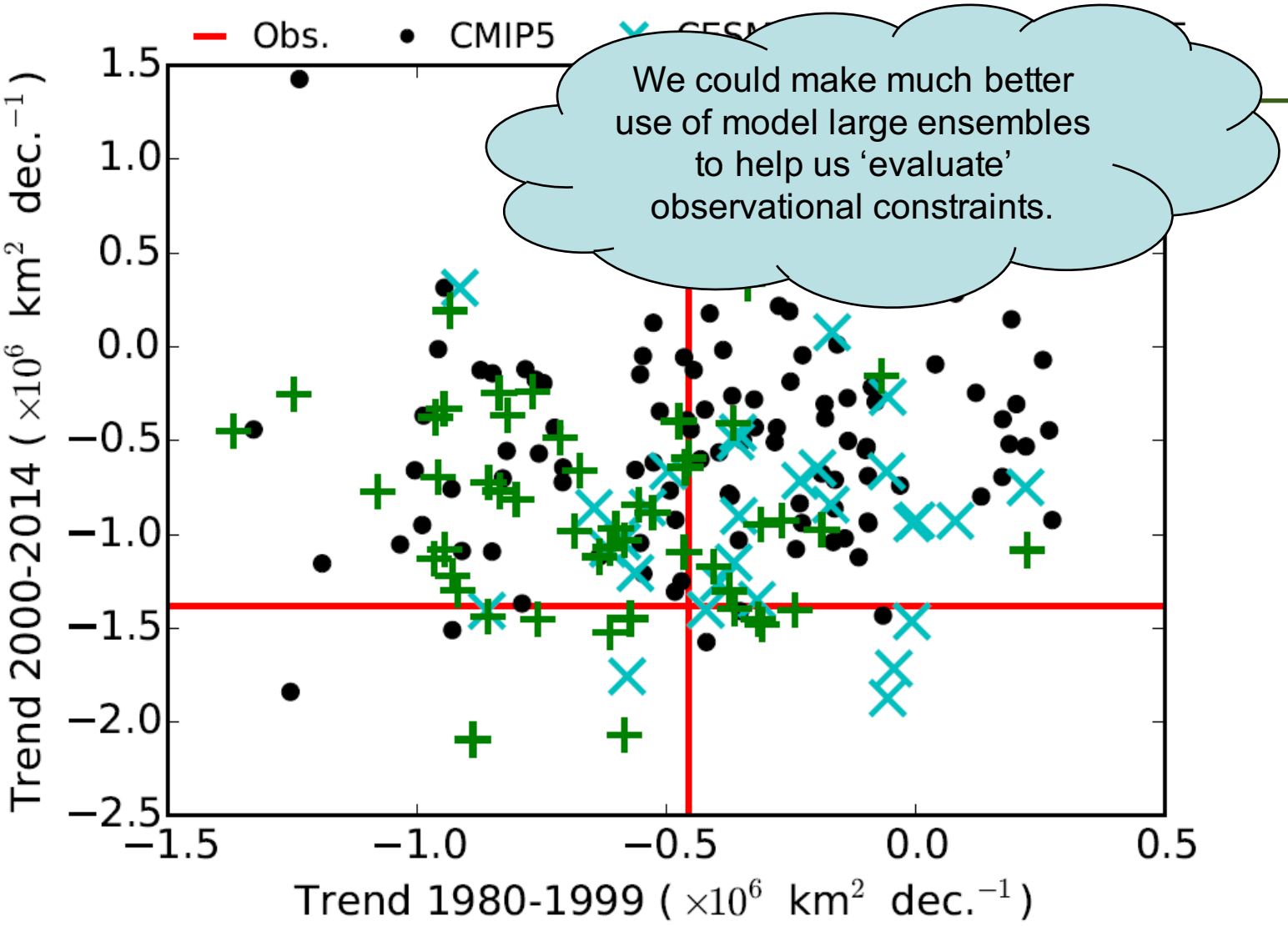
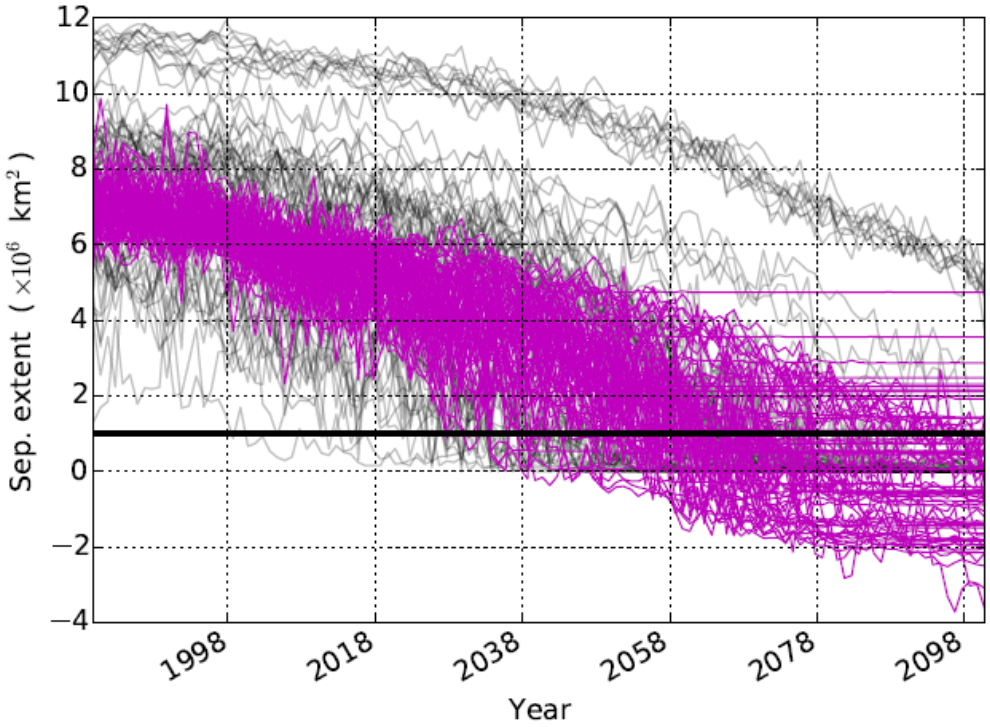


Figure prepared by Neil Swart, CCCma

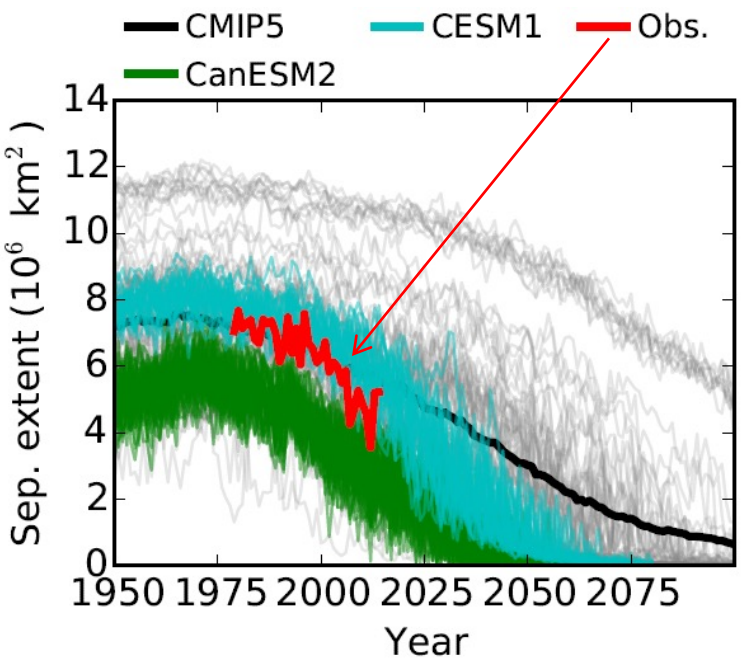
Alternatives to selection/weighting?

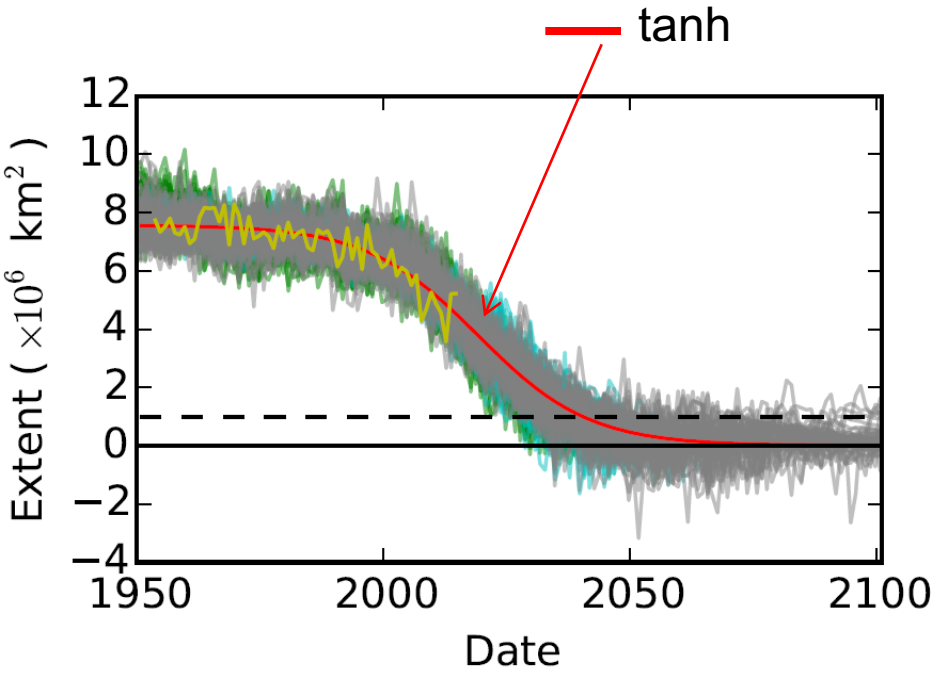
Bias correction (subtracting climatological error) can be useful in some case, but not all …



Flato, Swart and Fyfe, in prep.

Another possibility would be to try and extract a signal from model output, ‘calibrated’ by observations …





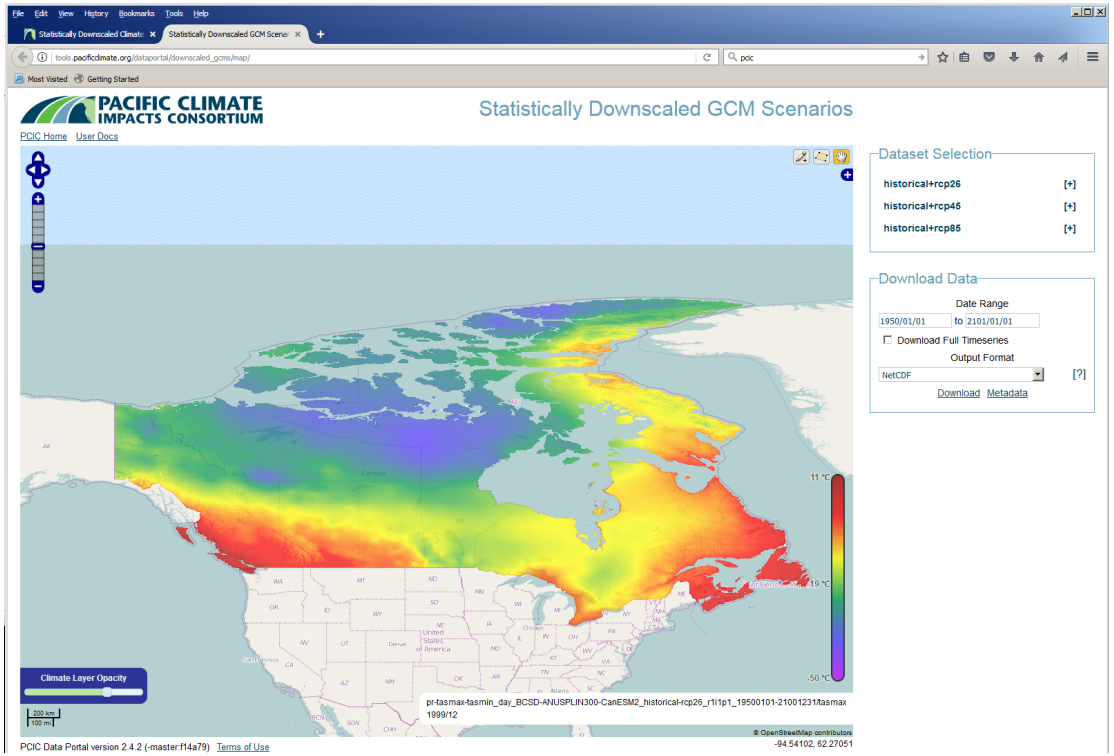
Flato, Swart and Fyfe, in prep.

Statistical post-processing techniques …

Useful for estimating changes in extreme events as shape of PDF is corrected.

e.g.:

- Bias-Correction Spatial Disaggregation (BCSD)
- Bias Correction/Constructed Analogues with Quantile mapping reordering (BCCAQ)

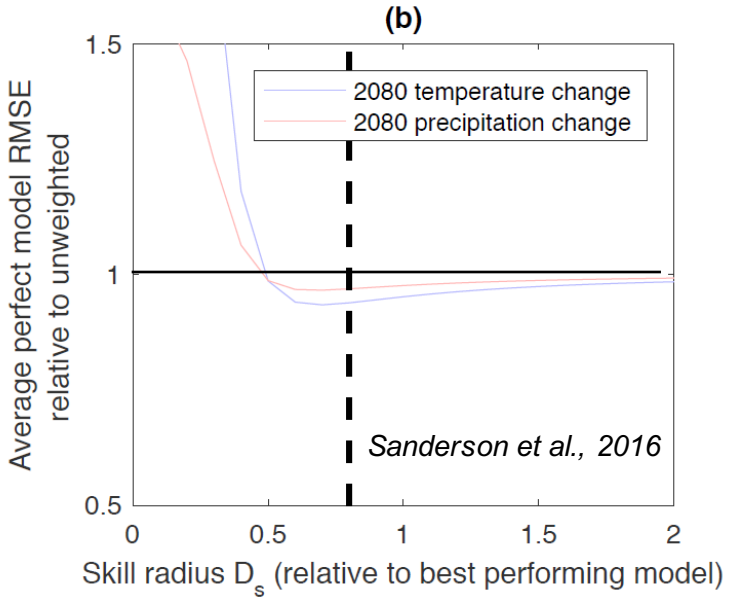


Concluding thoughts …

- Many traditional performance metrics (e.g. climatological bias), while revealing systematic errors, may not be particularly valuable in assessing model fitness-for-purpose (estimating TCR, future projection, b.c. for downscaling, …).
- More systematic evaluation will undoubtedly help, but we need to focus on things that matter (how do we identify those things?).
- Systematic evaluation without systematic (and comprehensive) model documentation will limit its value.
- The lack of progress on many longstanding biases suggests the need for more work on value-added post-processing (bias correction, 'signal extraction', other statistical approaches …).
- Model weighting/sub-setting may be helpful, but internal variability needs to be more carefully considered.

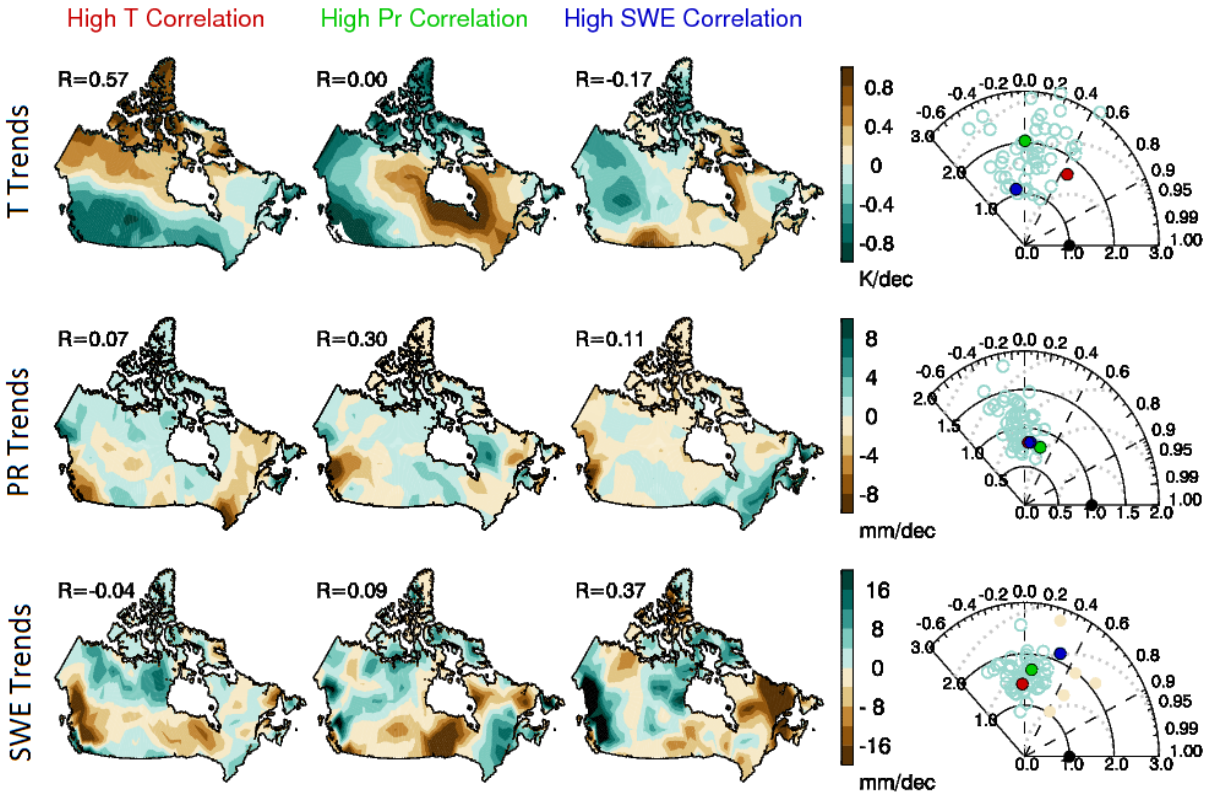
Concluding thoughts, cont'd …

- We should take more advantage of large ensembles and 'perfect model' studies to ask about the connection between historical performance and projection skill.
  - A recent example is that of Sanderson et al. (GMD, 2016)



Concluding thoughts, cont'd …

- But issues remain when we need consistent multi-variate information …

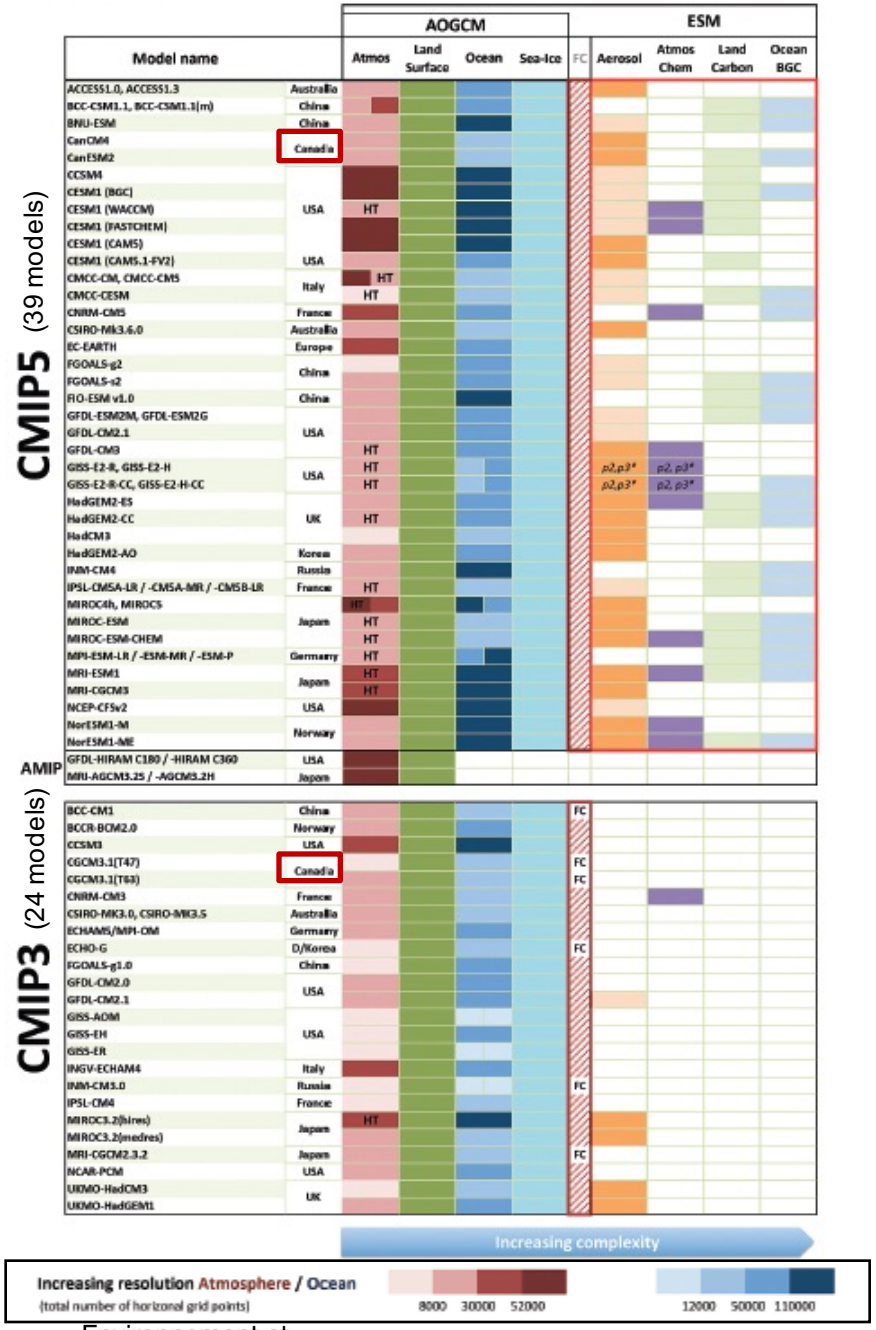


1981-2005 trends

L. Mudryk (pers. Comm.)
Based on Kushner et al., in press.

Thank you

More models; more comprehensive ...



Source: Table 9.1, IPCC AR5.

New capabilities, such as the widespread implementation of ‘satellite simulators’, were exploited.

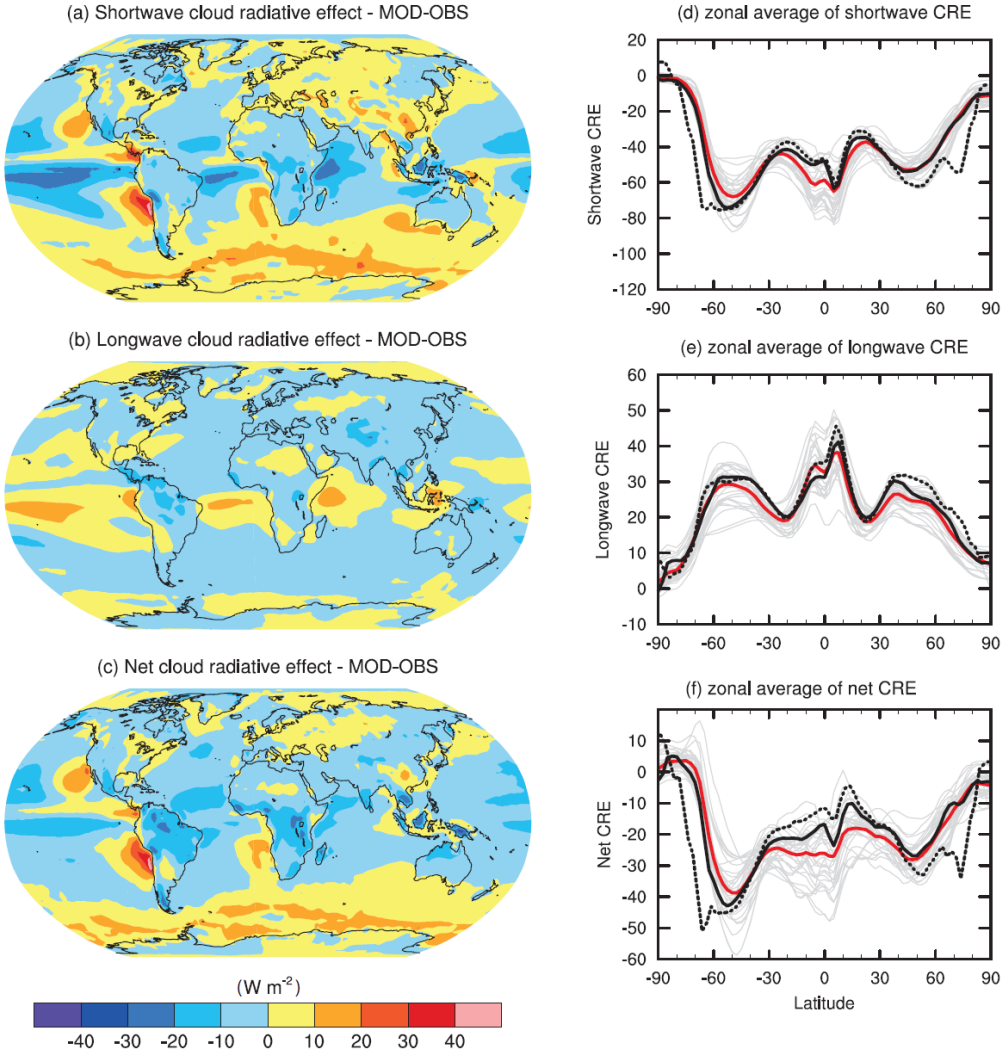


Figure 9.5 | Annual-mean cloud radiative effects of the CMIP5 models compared against the Clouds and the Earth's Radiant Energy System Energy Balanced and Filled 2.6 (CERES EBAF 2.6) data set (in W m^{-2}; top row: shortwave effect; middle row: longwave effect; bottom row: net effect). On the left are the global distributions of the multi-model-mean biases, and on the right are the zonal averages of the cloud radiative effects from observations (solid black: CERES EBAF 2.6; dashed black: CERES ES-4), individual models (thin grey lines), and the multi-model mean (thick red line). Model results are for the period 1985–2005, while the available CERES data are for 2001–2011. For a definition and maps of cloud radiative effect, see Section 7.2.1.2 and Figure 7.7.

We also made many comparisons of recent CMIP5 results to earlier CMIP3 results (used in the Fourth Assessment).

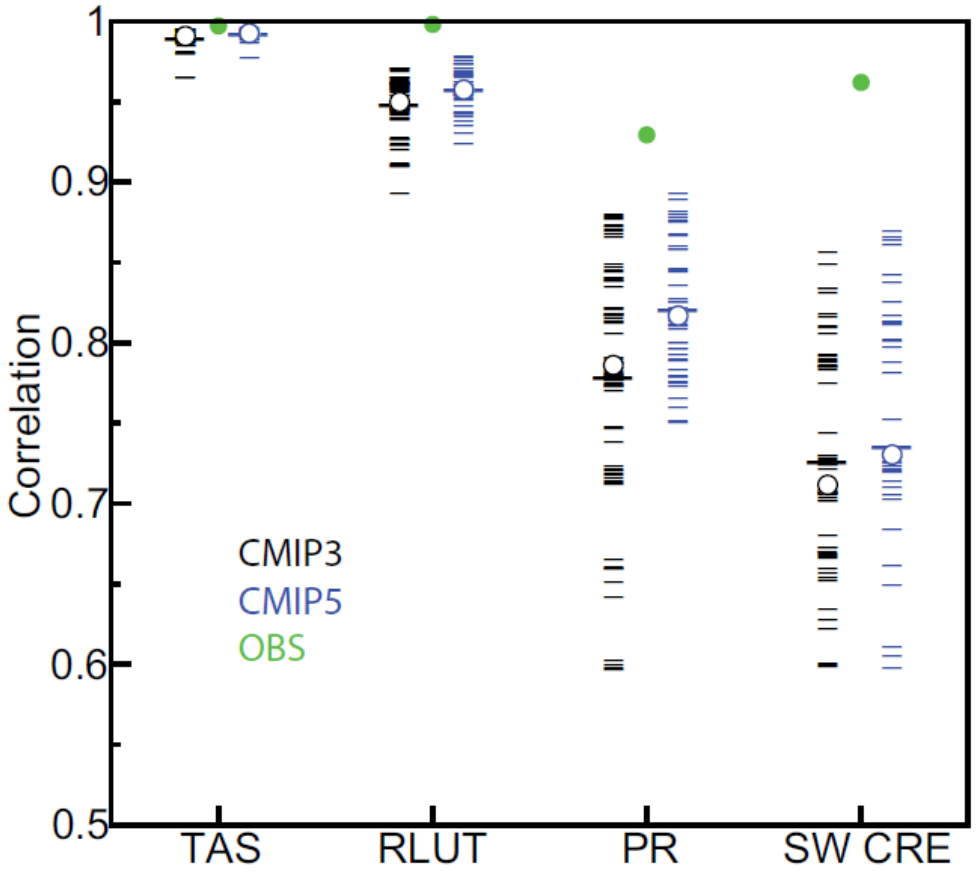


Figure 9.6 | Centred pattern correlations between models and observations for the annual mean climatology over the period 1980–1999. Results are shown for individual CMIP3 (black) and CMIP5 (blue) models as thin dashes, along with the corresponding ensemble average (thick dash) and median (open circle). The four variables shown are surface air temperature (TAS), top of the atmosphere (TOA) outgoing longwave radiation (RLUT), precipitation (PR) and TOA shortwave cloud radiative effect (SW CRE). The observations used for each variable are the default products and climatological periods identified in Table 9.3. The correlations between the default and alternate (Table 9.3) observations are also shown (solid green circles). To ensure a fair comparison across a range of model resolutions, the pattern correlations are computed at a resolution of 4° in longitude and 5° in latitude. Only one realization is used from each model from the CMIP3 20C3M and CMIP5 historical simulations.