Met Office
Hadley Centre

Assessing skill from retrospective forecasts

Doug Smith, Rosie Eade, Nick Dunstone, Leon Hermanson, Holger Pohlmann, Adam Scaife

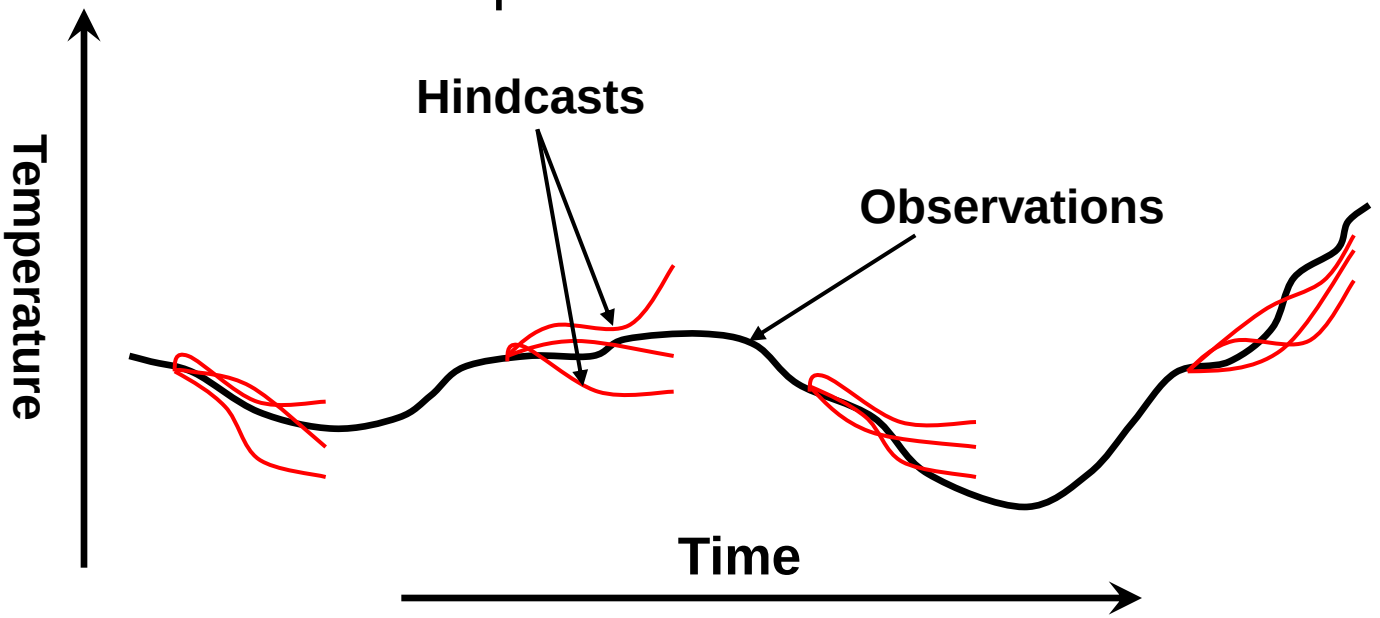
Contents

- Dealing with model bias
- Measuring skill
- Other issues
- Examples
  - Physical processes
  - Case studies

Hindcasts to assess skill

Ensembles to sample uncertainties:
- Uncertainties in the initial conditions
- Model errors

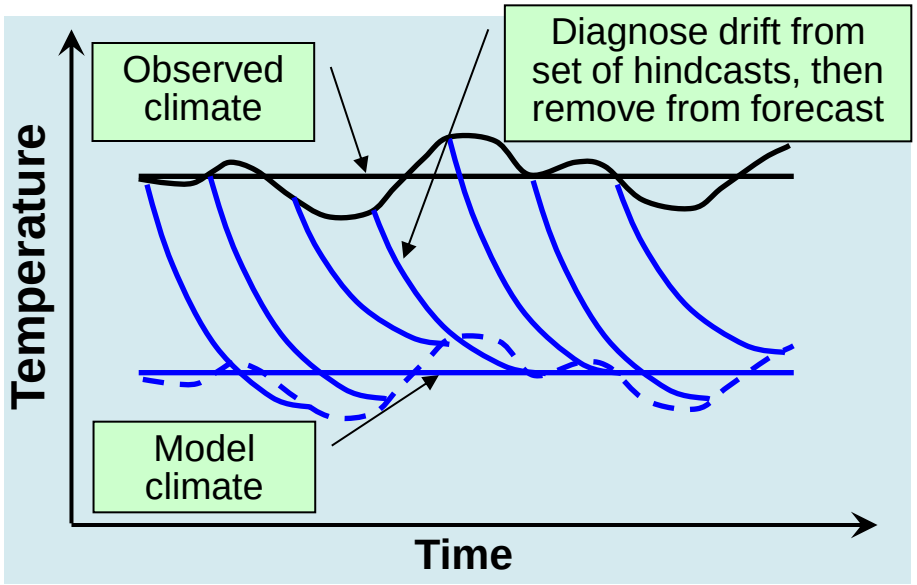
An optimistic view:



Perform historical tests ("retrospective forecasts" or "hindcasts" to assess likely skill and correct biases

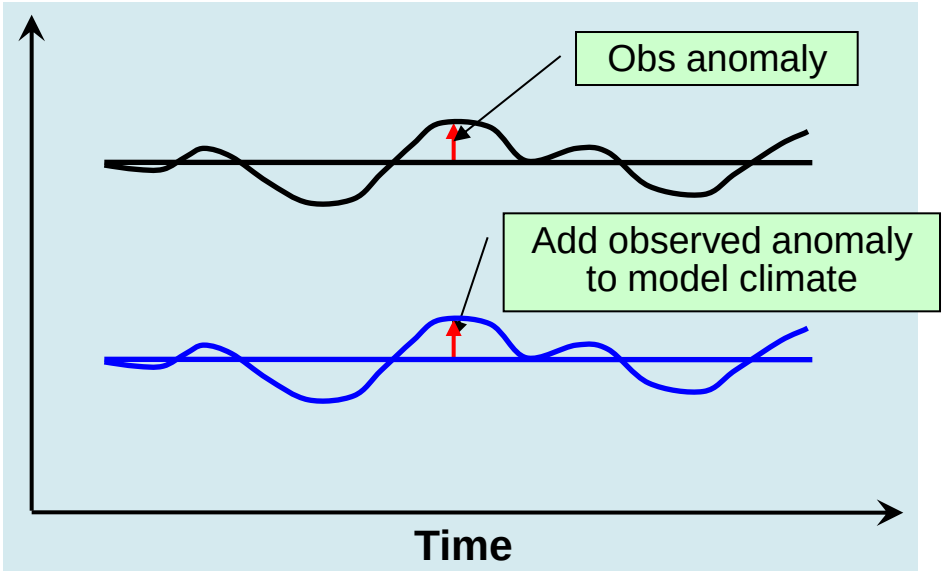
Models are imperfect: Dealing with model bias

Full field initialisation



- Routinely used in seasonal forecasting
- Ideally need large hindcast set, sampling multiple phases of variability
- Non-linearity?

Anomaly initialisation



- Needs model to be spun-up, together with simulation of recent period
- Observed anomalies could be in wrong location relative to model features
- Non-linearity?

Correcting the bias/drift

For hindcast j of N in total, and at lead time t:

Y_{jt} = raw forecast $\hat{Y}_{jt}$ = adjusted forecast O_{kt} = observation

Anomaly initialisation:

$$\hat{Y}_{jt} = Y_{jt} - \sum_{k=year1}^{year2} X_k / M + \sum_{k=year1}^{year2} O_k / M$$

X = independent model simulations (e.g. 20th century)

Full field initialisation:

$$\hat{Y}_{jt} = Y_{jt} - \sum_{k=1}^{N} (Y_{kt} - O_{kt}) / N$$

Raw forecast minus mean bias

…alternative without observations

$$\hat{Y}_{jt} = Y_{jt} - \sum_{k=1}^{N} Y_{kt} / N$$

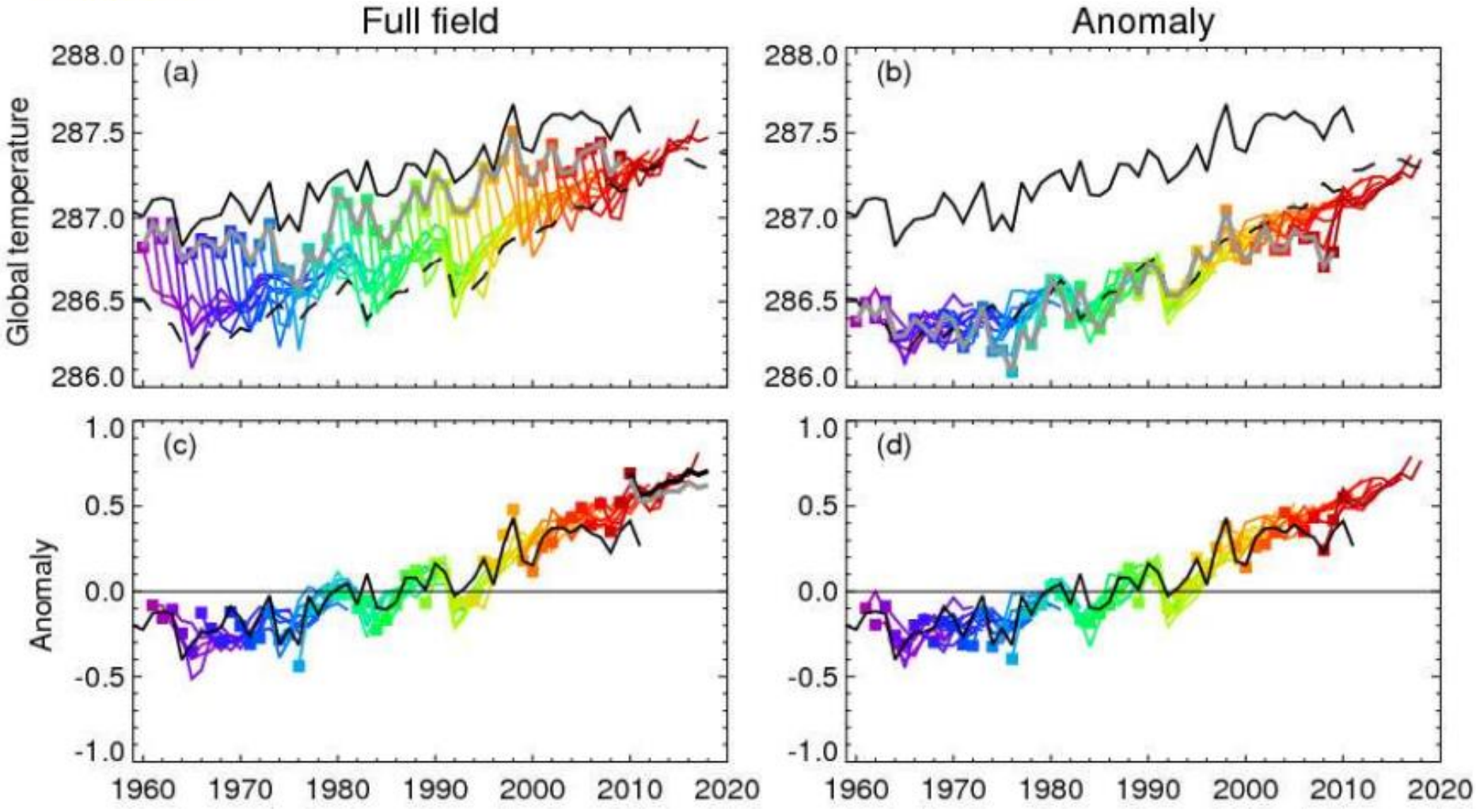
Raw forecast minus model climate for given lead time

…cross validated

$$\hat{Y}_{jt} = Y_{jt} - \sum_{\substack{k=1 \\ k \neq j}}^{N} (Y_{kt} - O_{kt}) / (N-1)$$

Bias computed over all **other** hindcasts

Correcting the bias/drift



(Smith et al. 2013)

Correcting the bias/drift

Full field initialisation:

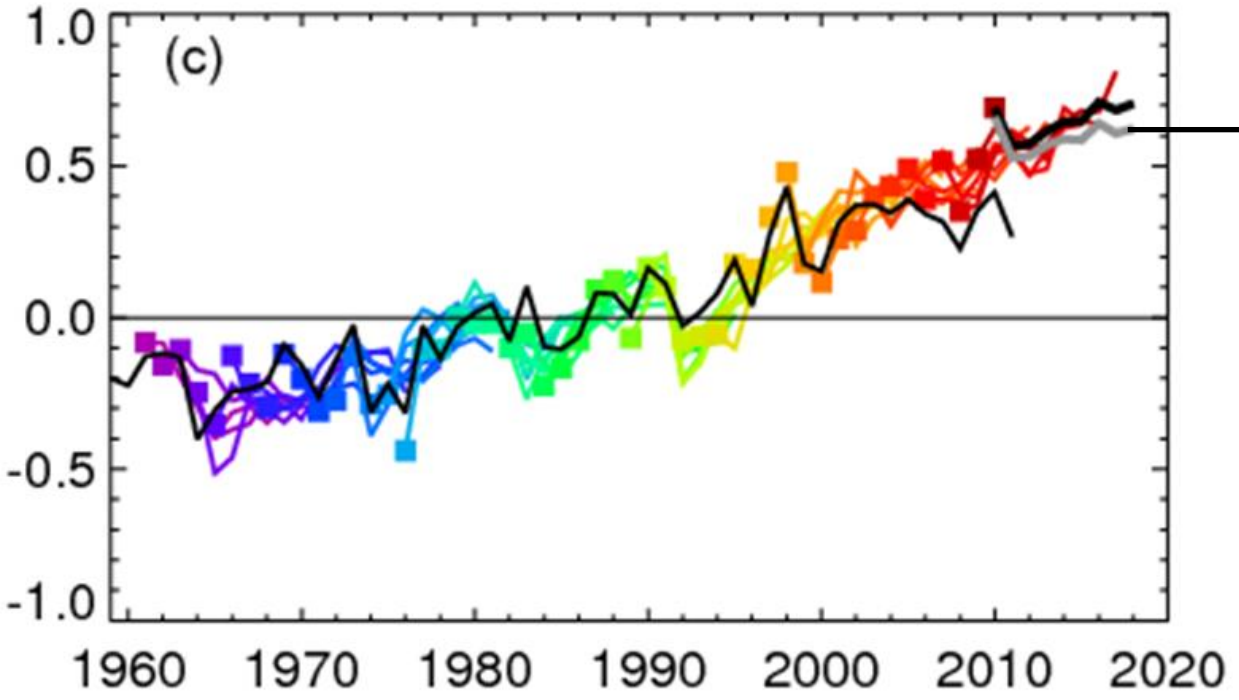
$$\hat{Y}_{jt} = Y_{jt} - \sum_{k=1}^{N} (Y_{kt} - O_{kt}) / N$$

Raw forecast minus mean bias

…alternative without observations

$$\hat{Y}_{jt} = Y_{jt} - \sum_{k=1}^{N} Y_{kt} / N$$

Raw forecast minus model climate for given lead time



Model climate for different lead times samples different periods:
e.g.
Year 1 = 1961 to 2001
Year 9 = 1969 to 2009

- **Year 9 climate warmer than year 1 climate**
- **Potentially removes the trend in forecasts**
- **Difficult to interpret time series**

(Smith et al. 2013)

Contents

- Dealing with model bias
- Measuring skill
- Other issues
- Examples
  - Physical processes
  - Case studies

Skill measures

N hindcast start dates

$\hat{Y}_{kt}$, O_{kt} Forecast and observation for hindcast k lead time t

Mean squared error

$$mse = \frac{\sum_{k=1}^{N} \left(\hat{Y}_{kt} - O_{kt}\right)^2}{N}$$

Root mean squared error

$$rmse = \frac{\sqrt{mse}}{\sigma_o}$$

(normalised)

≥0 no upper limit

Mean squared skill score

$$msss = 1 - \frac{mse}{mse_{ref}}$$

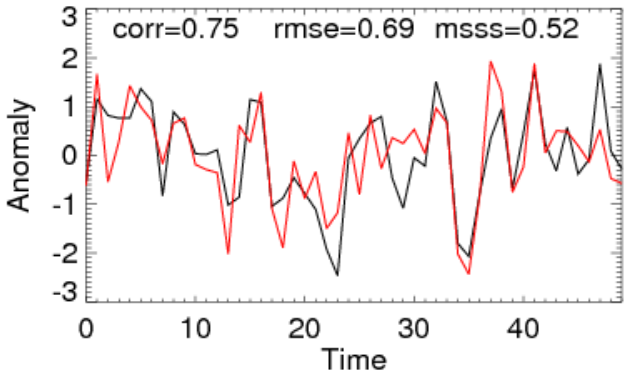
-infinity to +1

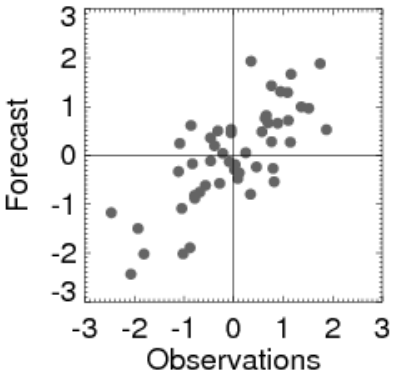
Anomaly correlation

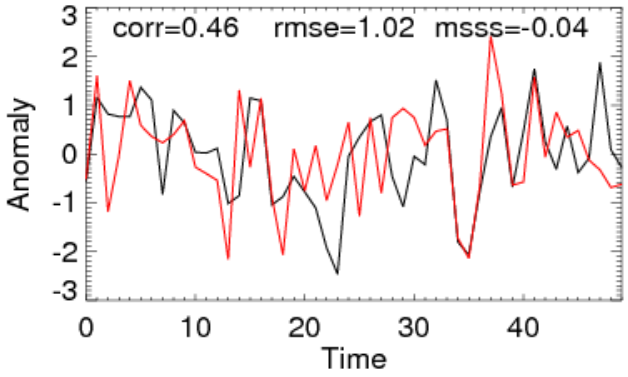
$$corr = \frac{\mathrm{cov}}{\sigma_o \sigma_y}$$

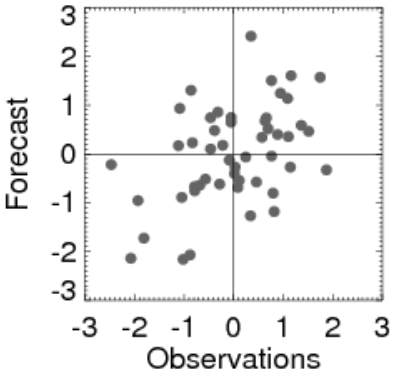
-1 to +1

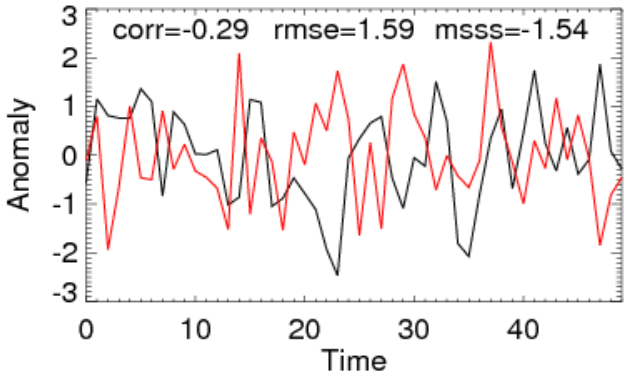
potential skill

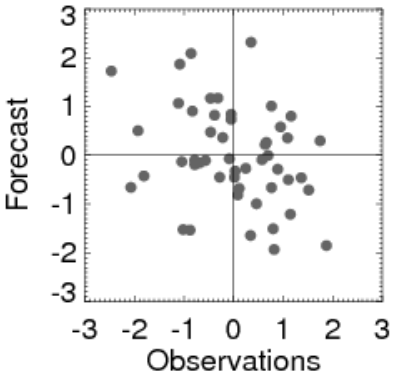




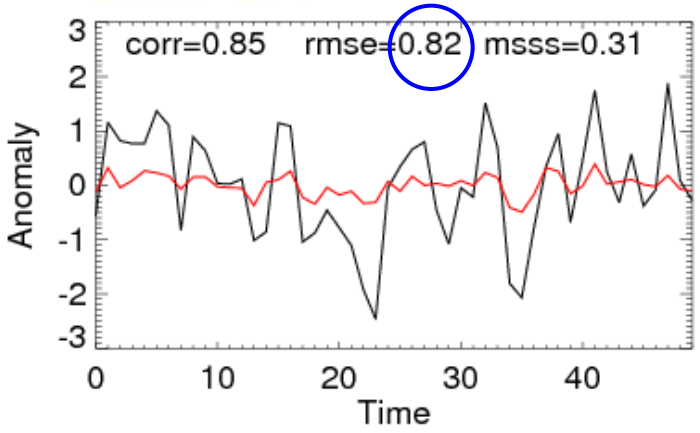


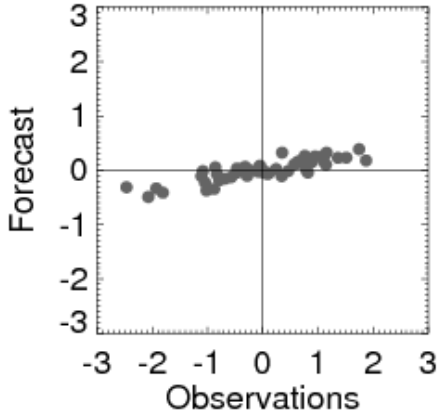


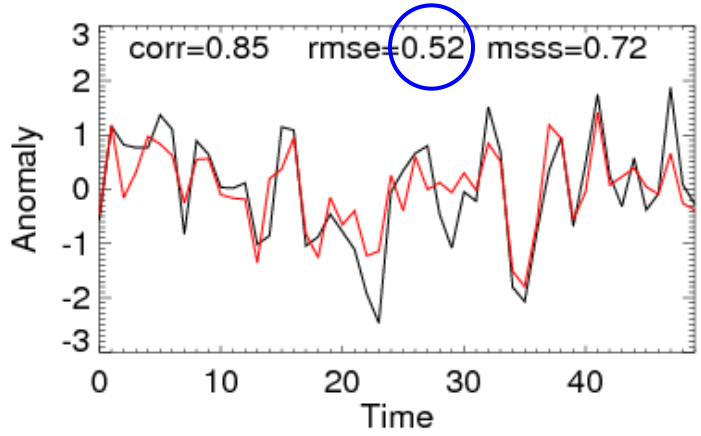


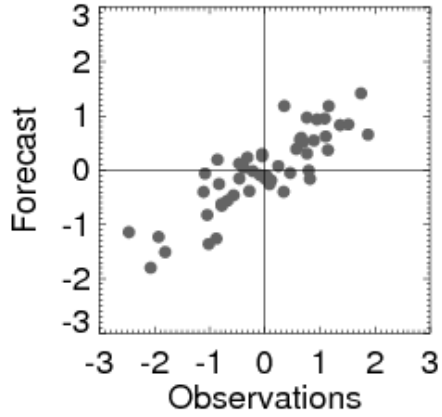


Skill measures: potential skill



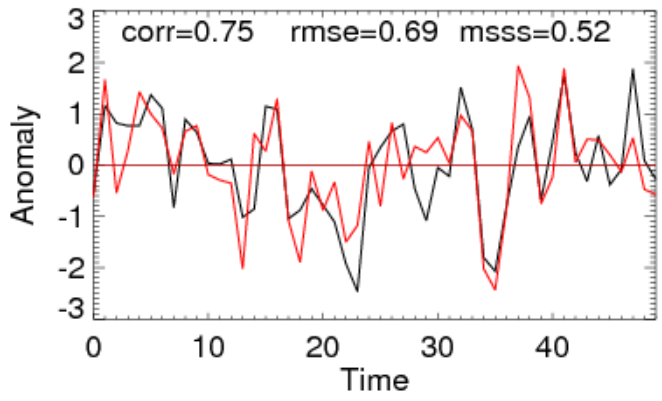


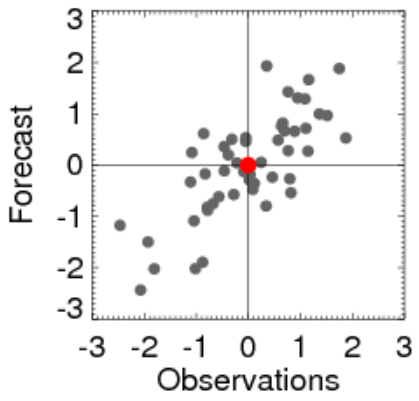


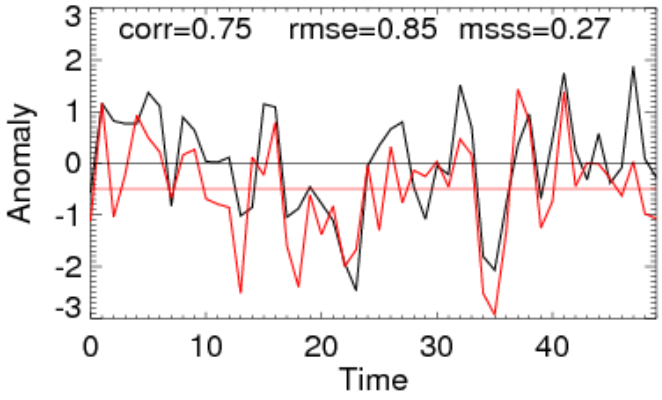


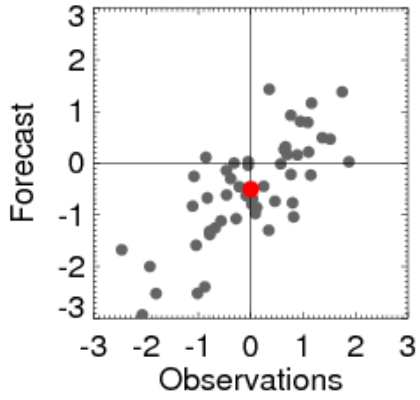
- Correlation measures **potential** skill
- Need to post process the forecast to achieve this i.e. minimize rmse
- Adjust the variance to be equal to the predictable component (corr*corr)

Skill measures: imperfect bias removal









- rmse, msss reduced
- correlation insensitive to bias

$$corr = \frac{\mathrm{cov}}{\sigma_o \sigma_y} = \frac{\sum_{k=1}^{N} (\hat{Y}_{kt} - \bar{Y})(O_{kt} - \bar{O})}{\sqrt{\sum_{k=1}^{N} (\hat{Y}_{kt} - \bar{Y})^2} \sqrt{\sum_{k=1}^{N} (O_{kt} - \bar{O})^2}}$$

- use uncentred correlation:

$$corr_{un} = \frac{\sum_{k=1}^{N} (\hat{Y}_{kt} - \bar{O})(O_{kt} - \bar{O})}{\sqrt{\sum_{k=1}^{N} (\hat{Y}_{kt} - \bar{O})^2} \sqrt{\sum_{k=1}^{N} (O_{kt} - \bar{O})^2}}$$

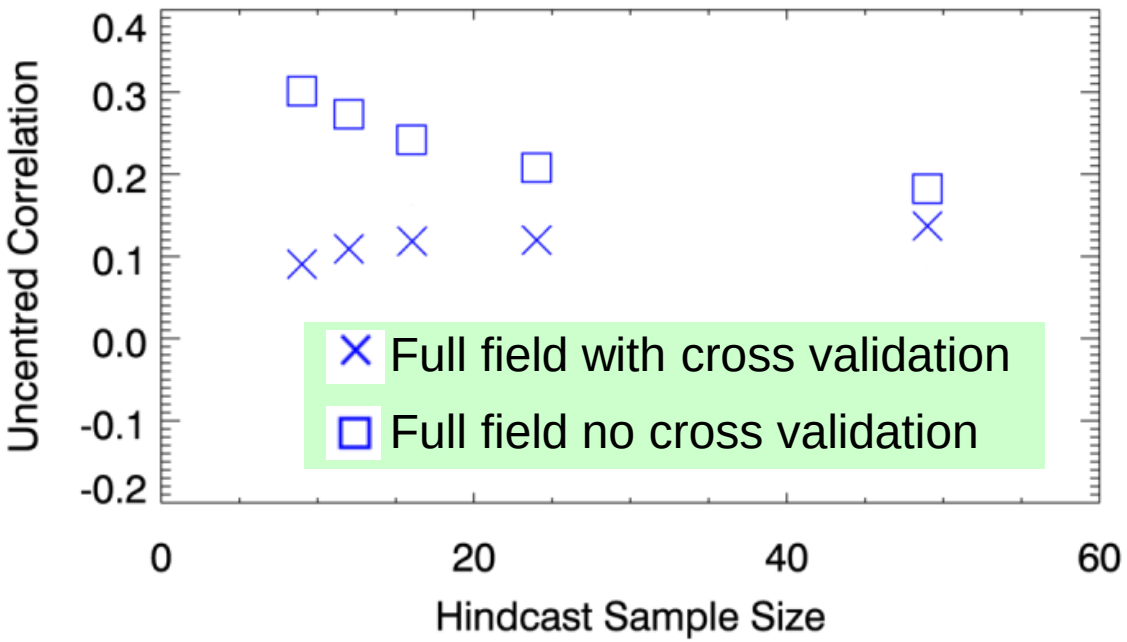
- $corr_{un} = 0.67$

Cross validation

…cross validated

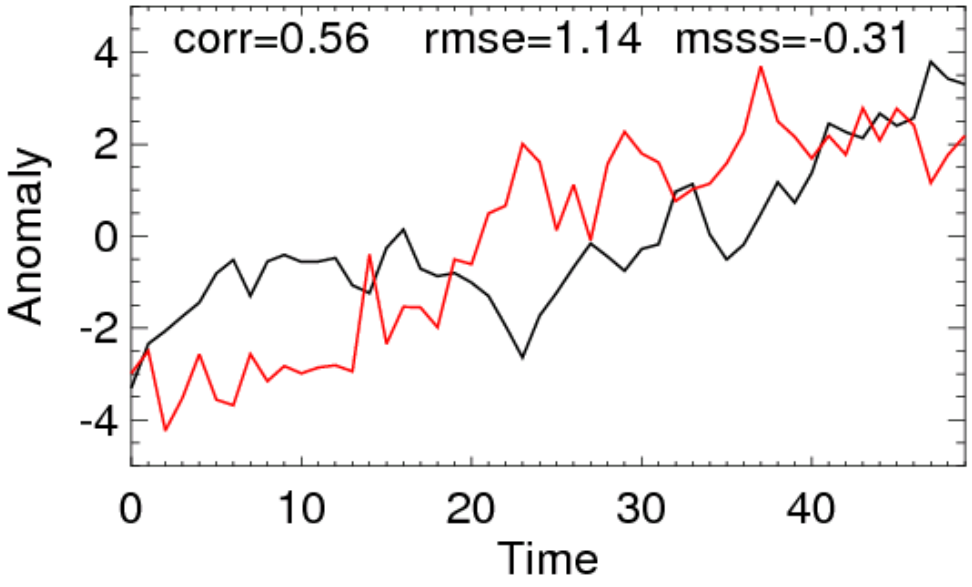
$$\hat{Y}_{jt} = Y_{jt} - \sum_{\substack{k=1 \\ k \neq j}}^{N} (Y_{kt} - O_{kt})/(N-1)$$

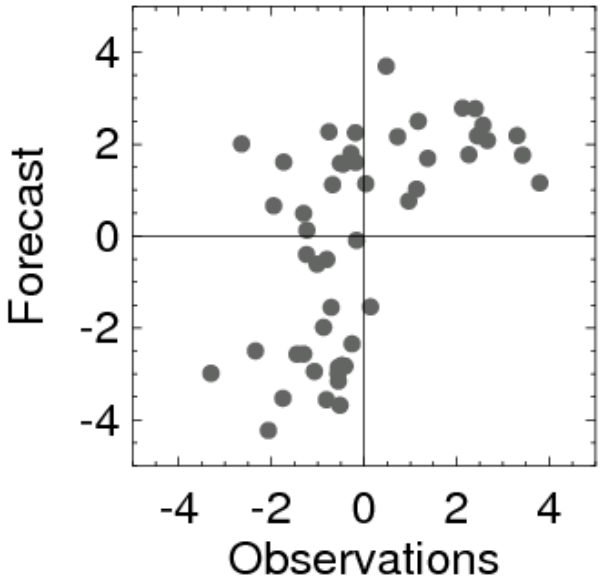
Bias computed over all **other** hindcasts



- Bias correction is imperfect, especially with small hindcast samples
- Skill over estimated without cross validation
- But under estimated with cross validation!

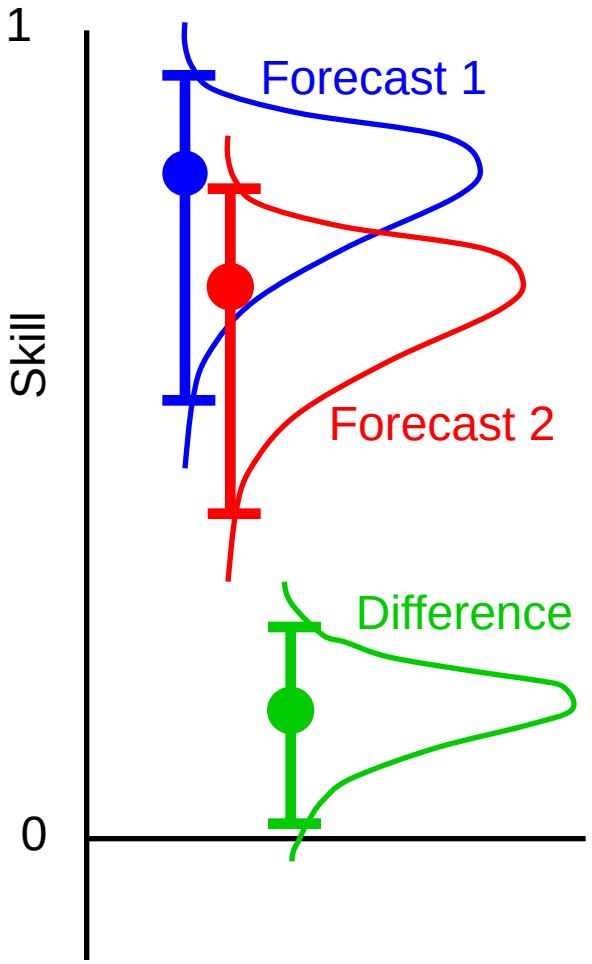
Beware of trends!





- Positive correlation caused by trend
- rmse > 1 and msss < 0
- Detrended correlation = - 0.7
- Need to examine the time series!

Significance



- Many approaches
- Complicated by autocorrelation
- None are perfect ⇒ just a guide

e.g. Block bootstrap

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ⇒ skill_0

7 8 9 13 14 15 3 4 5 10 11 12 2 3 4 ⇒ skill_1

6 7 8 2 3 4 10 11 12 12 13 14 8 9 10 ⇒ skill_2

8 9 11 8 9 10 1 2 3 2 3 4 12 13 14 ⇒ skill_3

12 13 14 4 5 6 2 3 4 1 2 3 13 14 15 ⇒ skill_4

… many times (e.g. 5000)

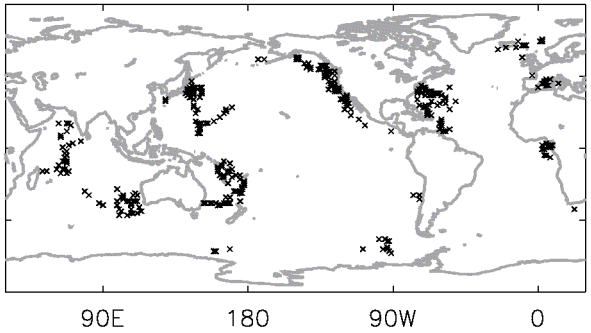
- build pdf (skewed)
- build pdf of differences in skill

Contents

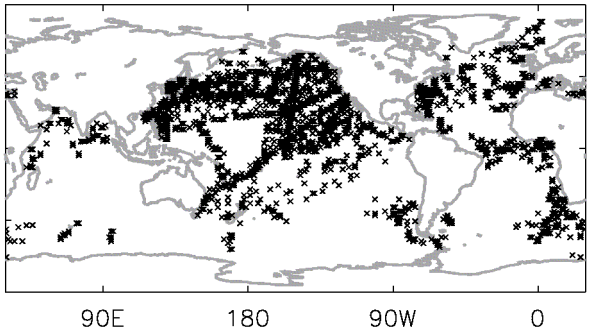
- Dealing with model bias
- Measuring skill
- Other issues
- Examples
  - Physical processes
  - Case studies

Sub-surface ocean observations

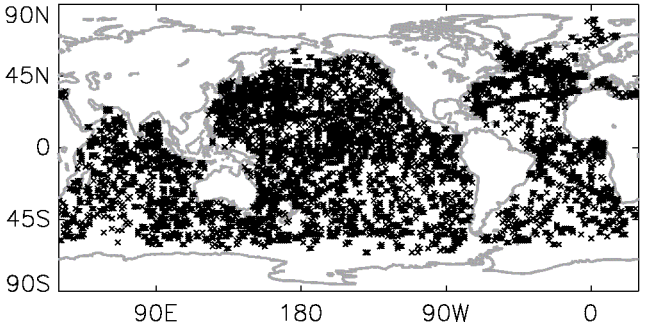
1960



1980

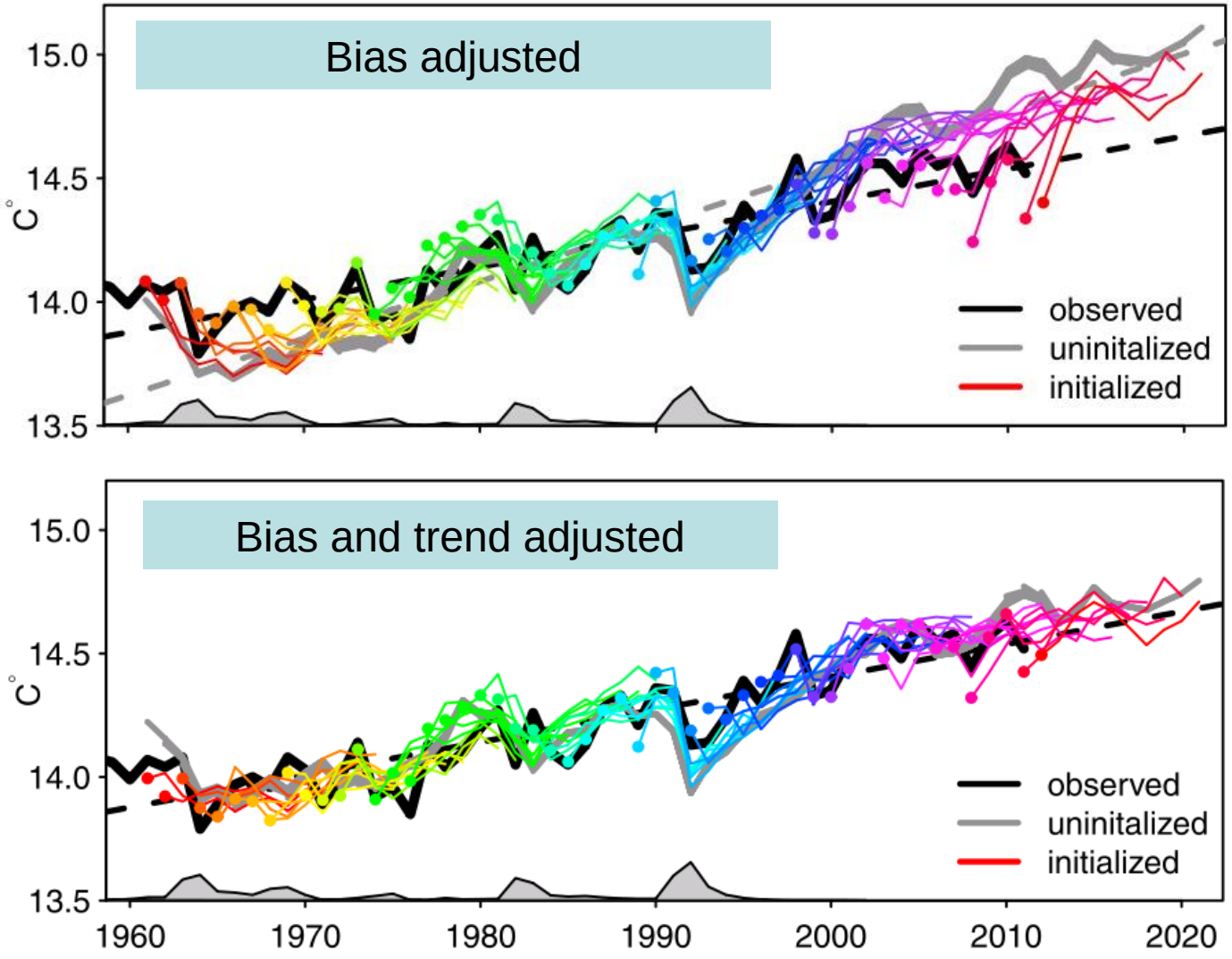


2007



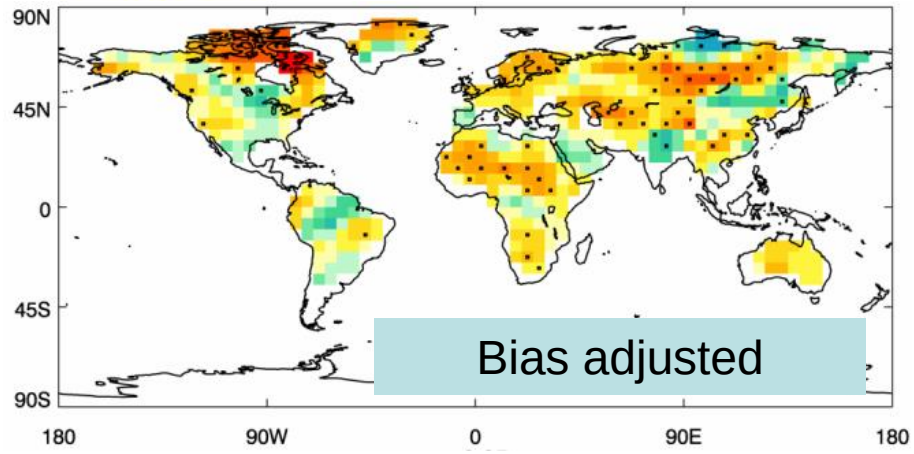
- Need historical tests to assess likely skill of forecasts
- Far fewer sub-surface ocean observations in the past
- Could forecasts be more accurate than hindcasts?

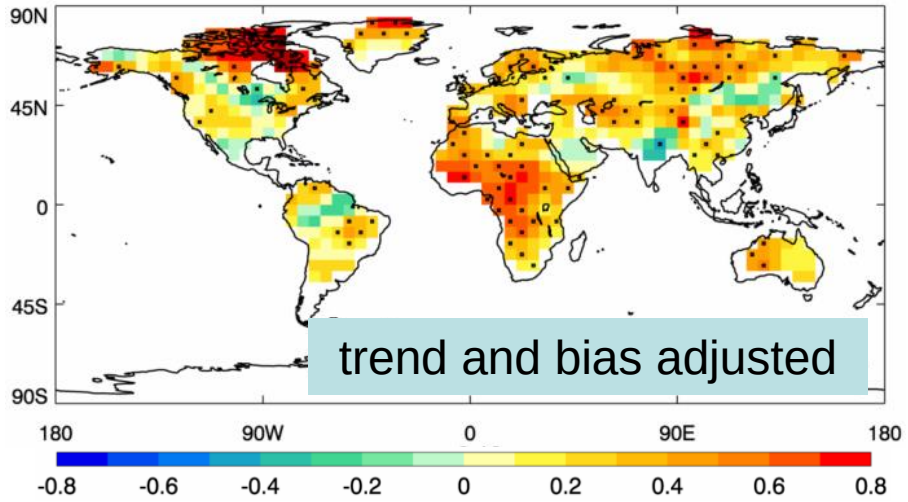
Additional trend adjustment

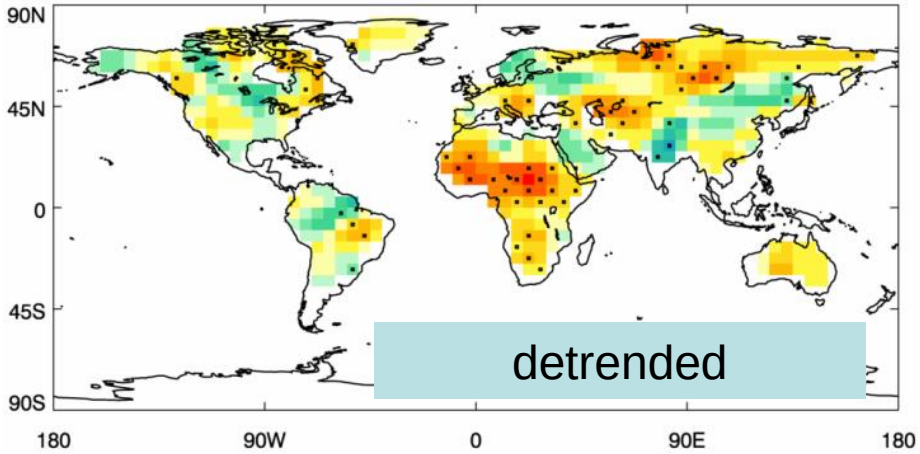


(Kharin et al. 2012)

Additional trend adjustment: years 1-5 precipitation correlation

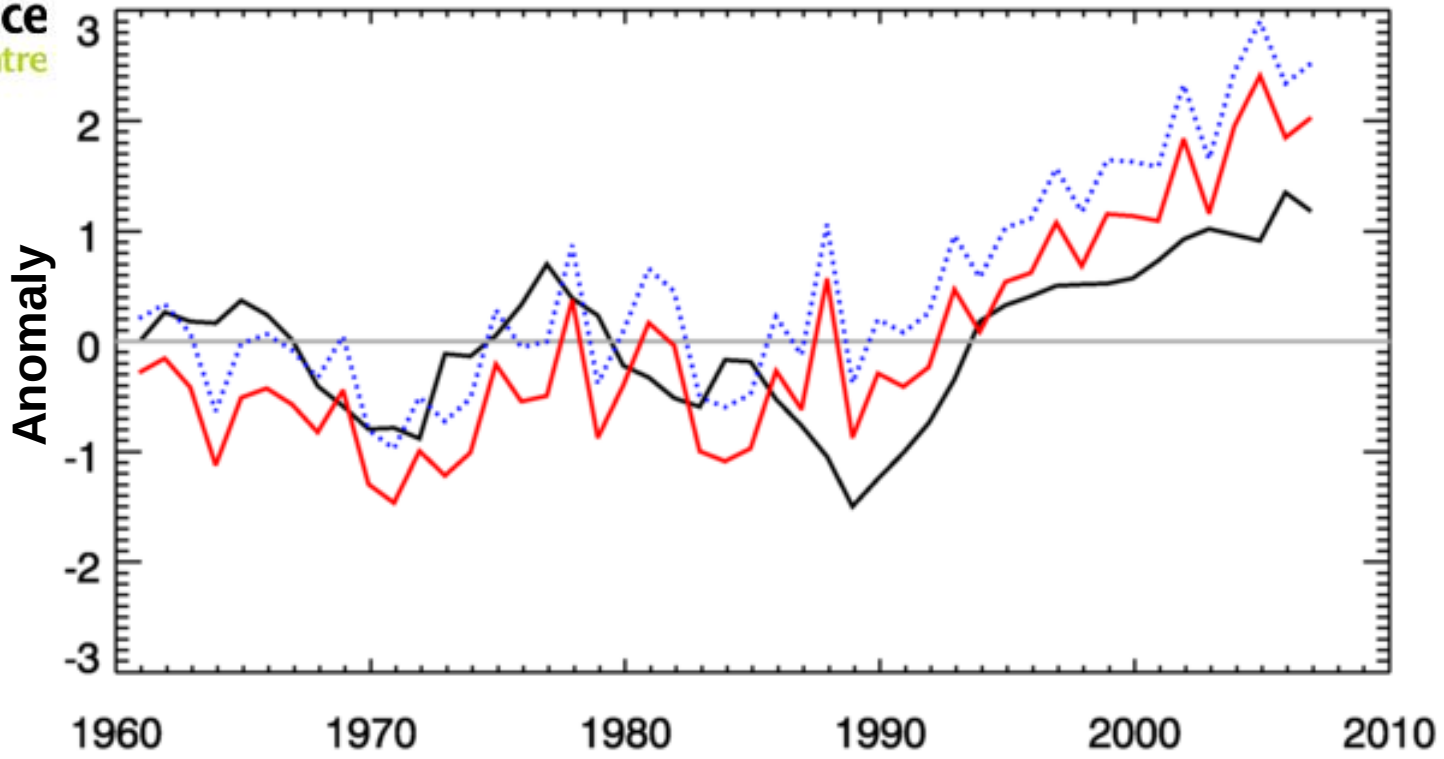






- Need to assess detrended skill!

Low frequency variability and bias correction



Observations

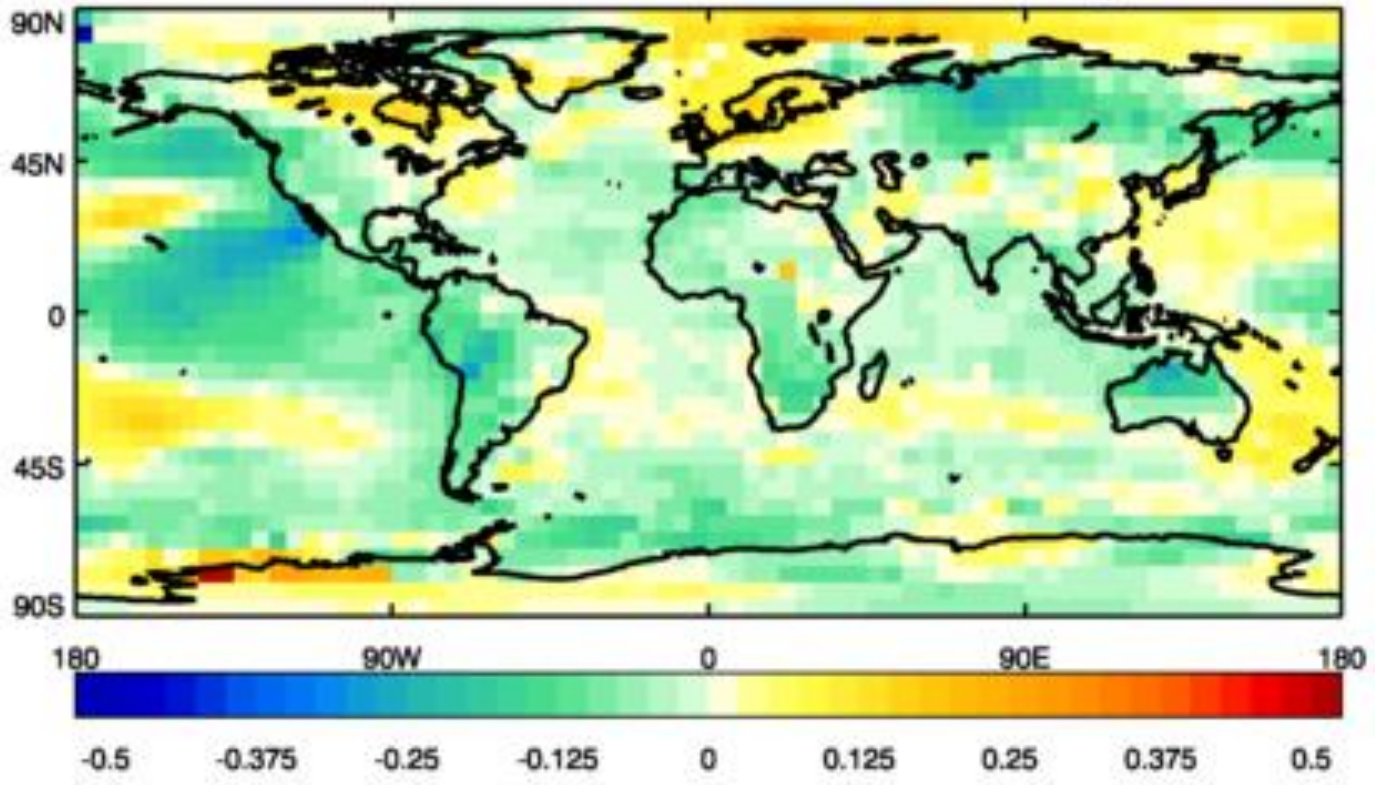
Bias correction using all years (1960-2005)

Bias correction sampling –ve phase (1960-85)

Need to bias correction to sample both phases to minimize errors

Volcanoes and bias correction

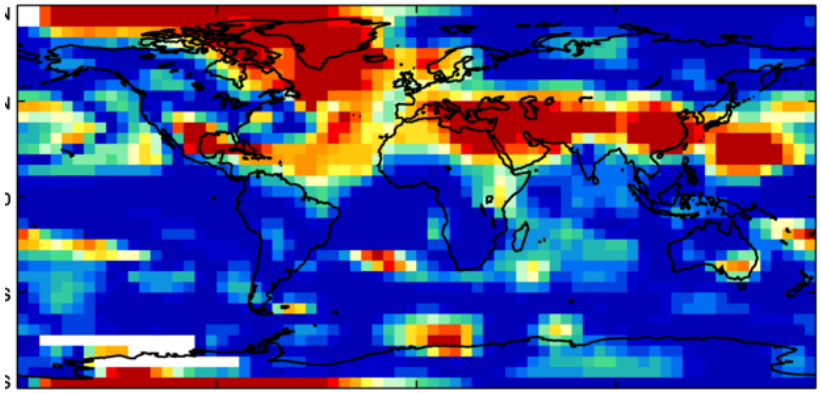
Bias correction difference if 3 years following volcanoes are excluded



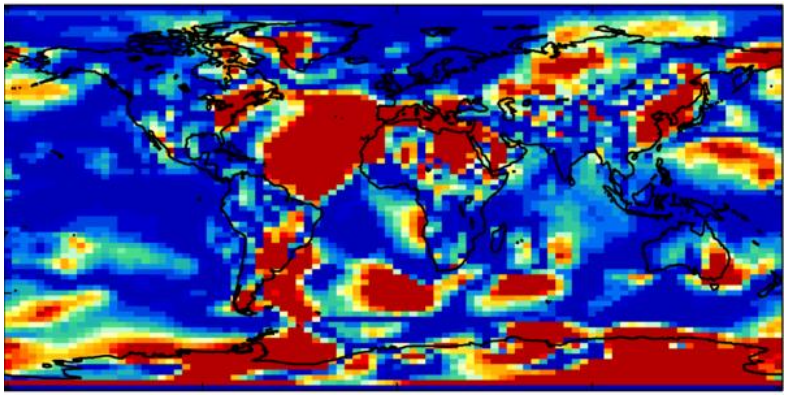
- Imperfect model response to volcanoes
- Hindcasts are generally too cool following volcanoes
- Bias correction too warm ⇒ forecasts too warm

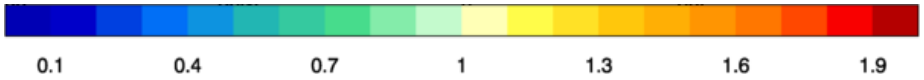
Signal to noise and correlation: Years 2-5

Temperature



Pressure





- Is predictable component of obs and model the same?
- predictable component of obs P(obs) = r^2
- predictable component of model

  P(model) = var(ensemble mean)/var(ensemble member)
- Plot ratio P(obs)/P(model)
- Each member not necessarily a potential realisation of reality
  - Need large ensemble, and to adjust variance

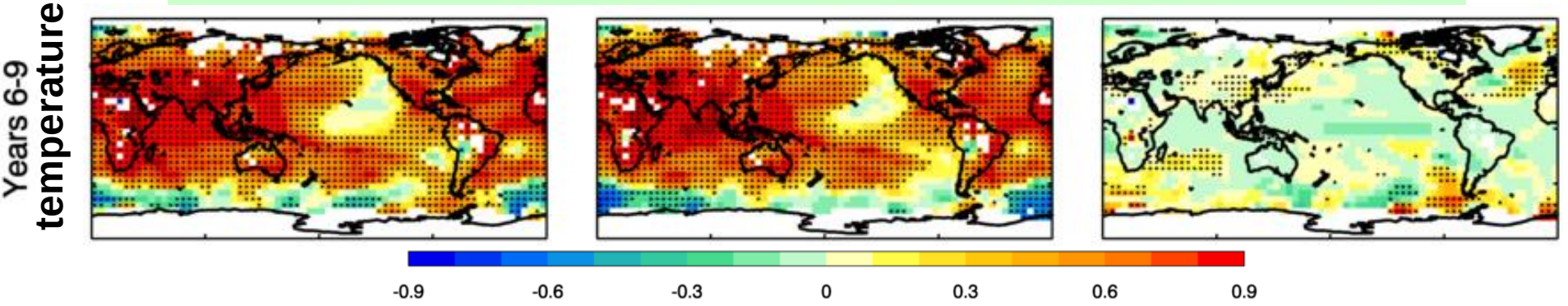
(Eade et al. In prep)

Contents

- Dealing with model bias
- Measuring skill
- Other issues
- Examples
  - Physical processes
  - Case studies

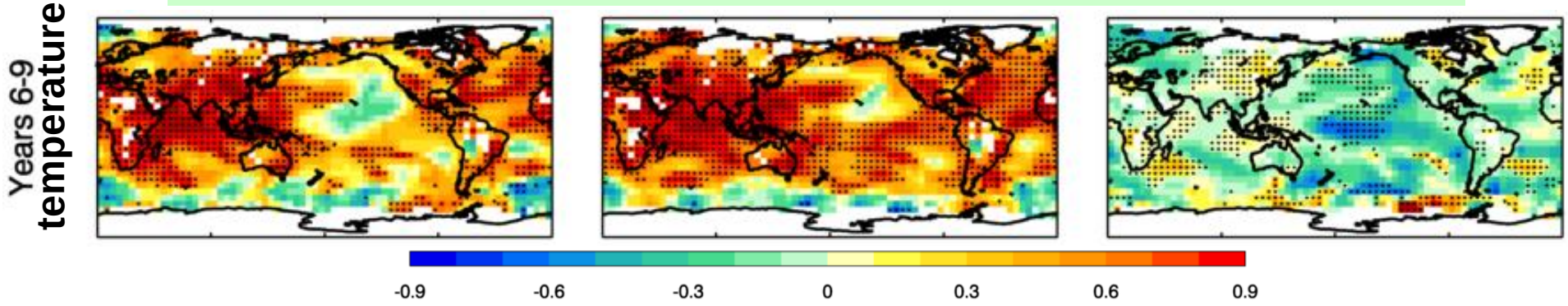
Full fields versus anomaly initialisation: years 6-9 temp

50 start dates (Nov 1st every year from 1960 to 2009)



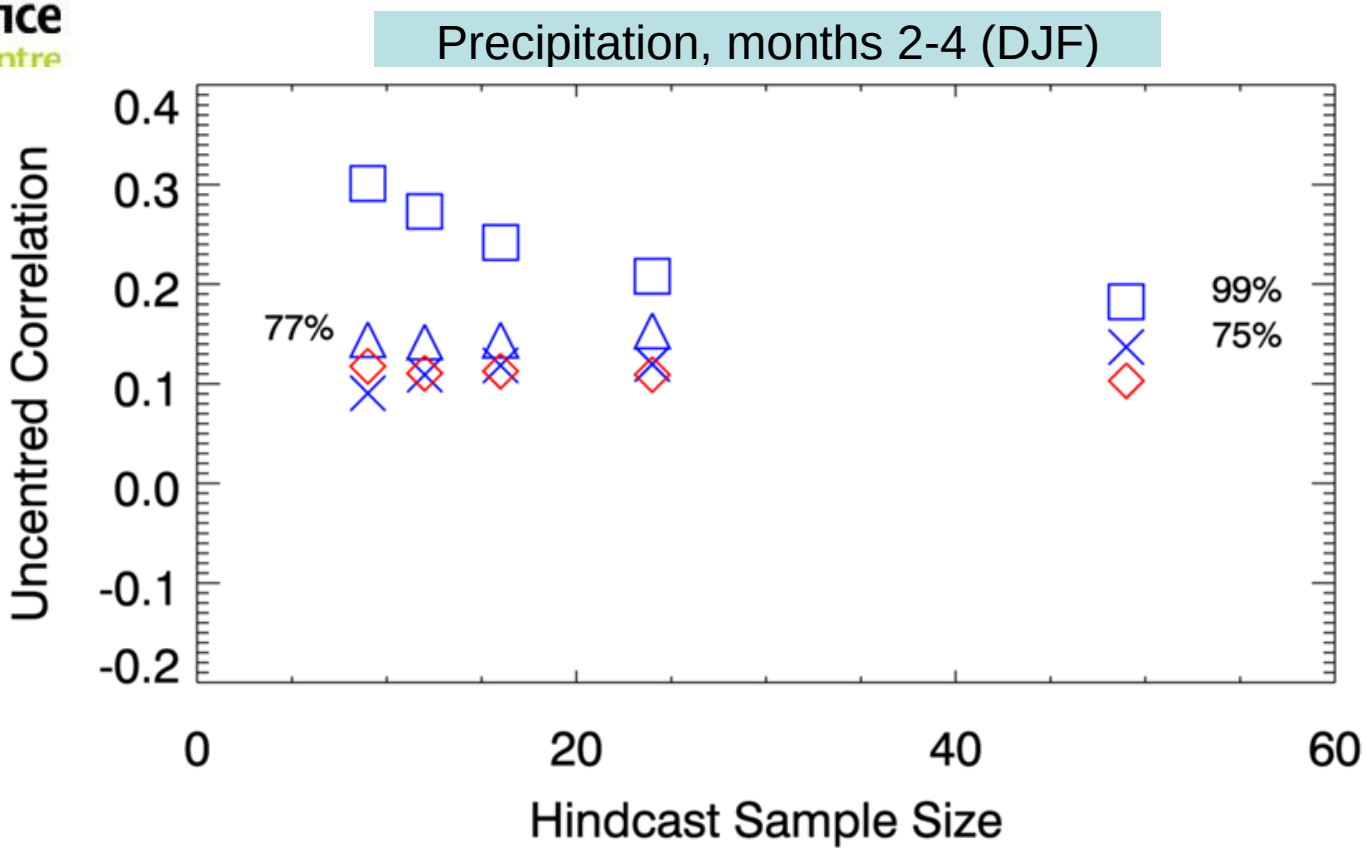
Full field | Anomaly | Difference

10 start dates (Nov 1st every 5 years from 1960 to 2005)



(Smith et al. 2013)

Assessing skill in retrospective forecasts
Global average of regional correlations



◇ Anomaly initialisation

× Full field with cross validation

□ Full field no cross validation

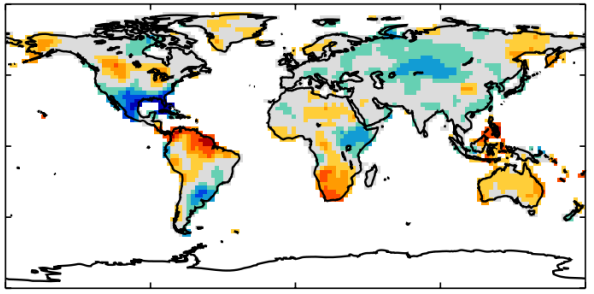
Cross validation: compute bias correction from all hindcast **except** the one being corrected

(Smith et al. 2013)

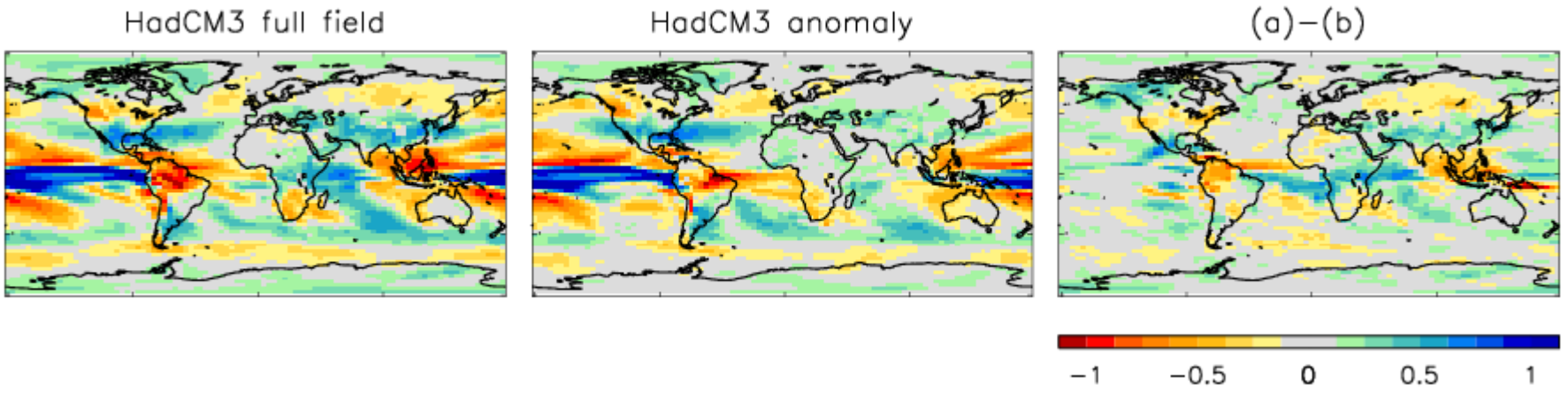
DJF precip (months 2-4)

Met Office Hadley Centre

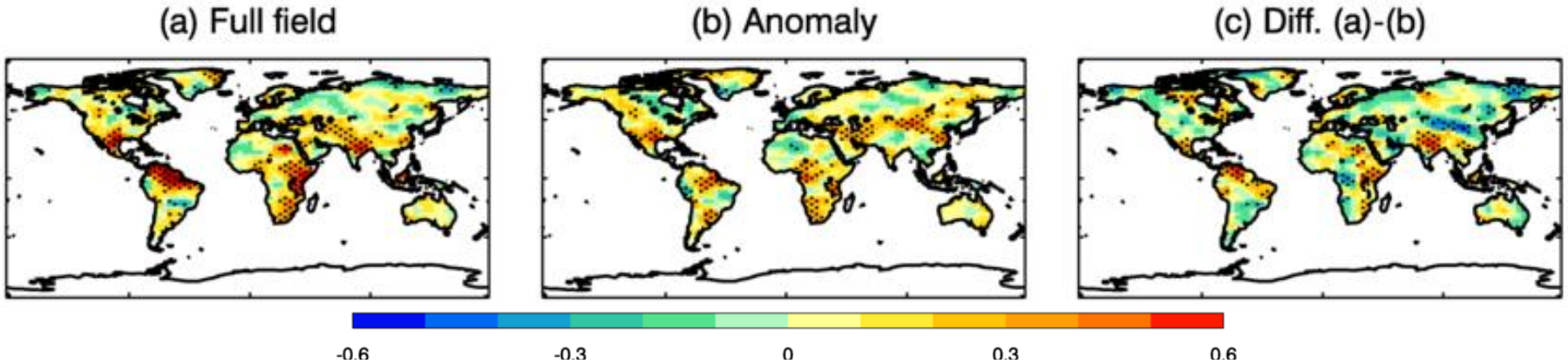
Observed Nino composite



Model composites



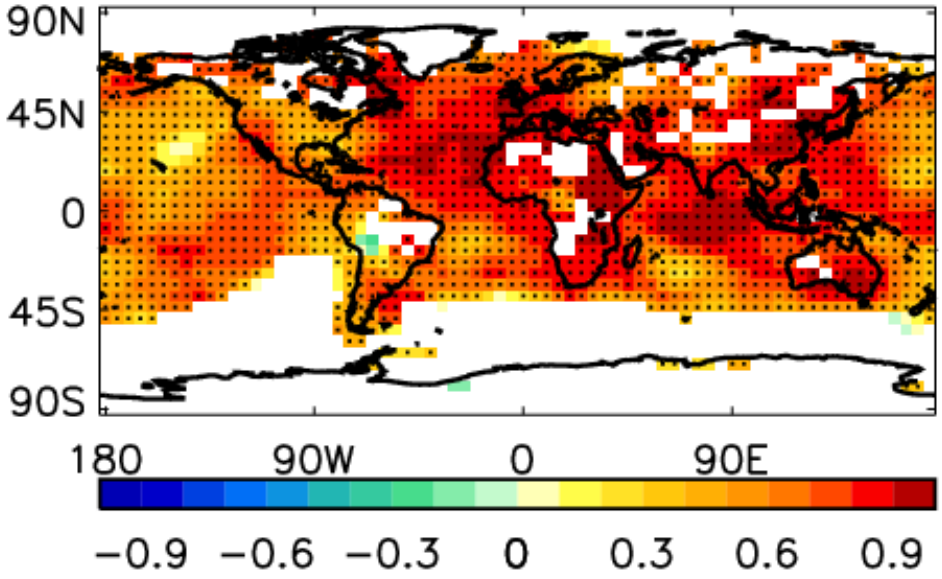
Hindcast skill



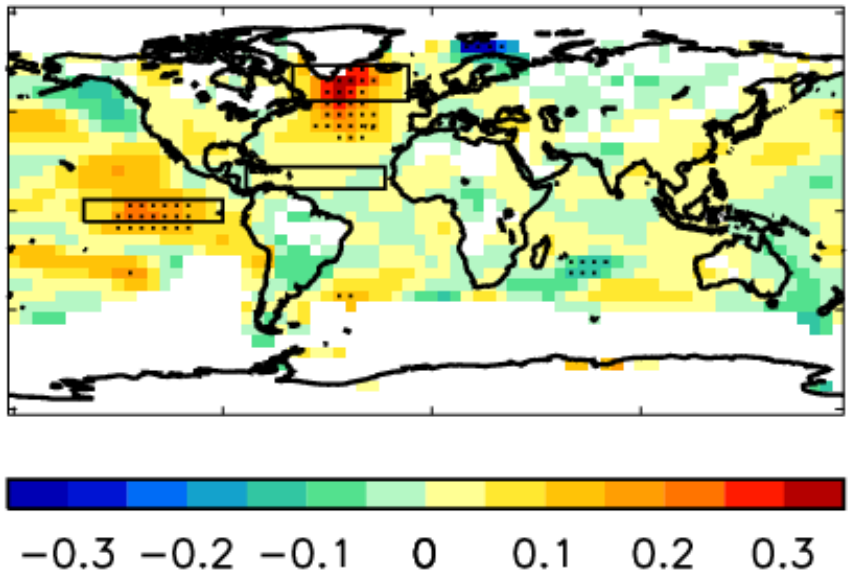
(Smith et al. submitted)

Surface temperature predictions (five year means)

Skill of initialised predictions



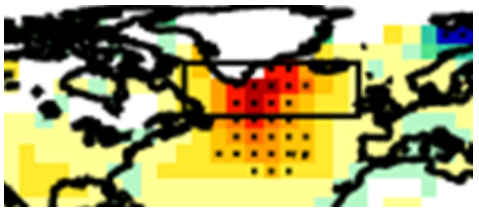
Initialised - Uninitialised

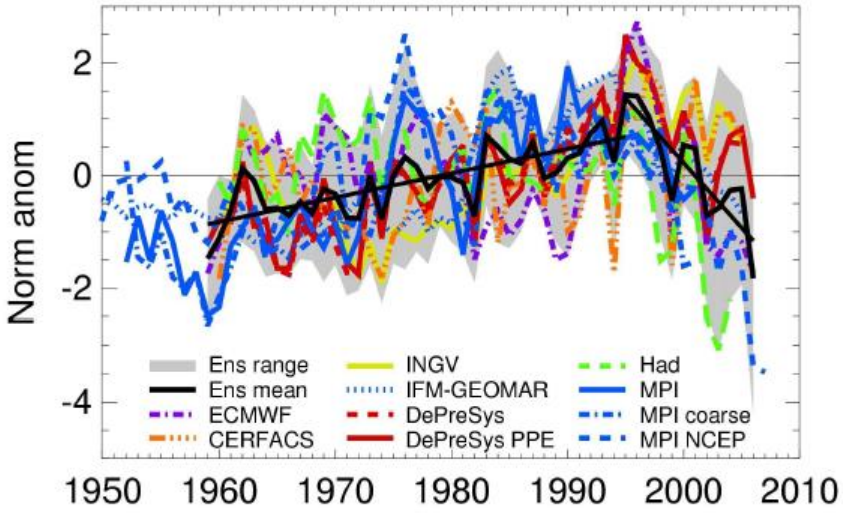


- Skilful almost everywhere (positive correlations)
- Mostly due to external forcing
- Initialisation gives improved skill mainly in North Atlantic and tropical Pacific

(Smith et al. 2010)

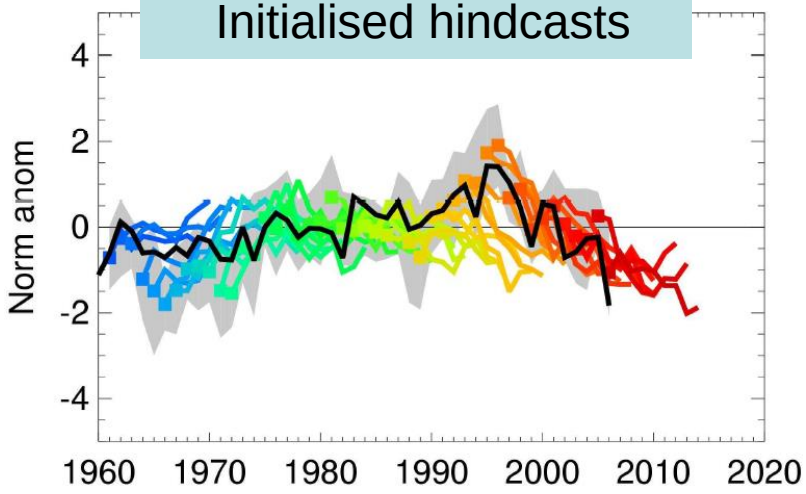
Physical basis for improved skill



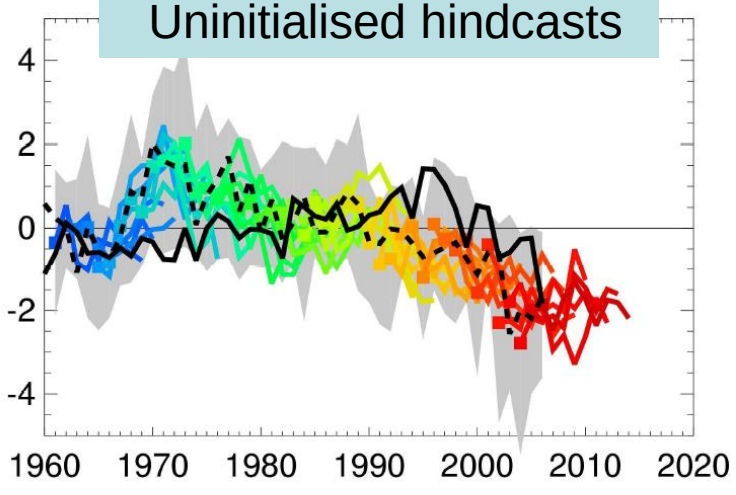


- No historical observations – must rely on models
- Consistent signal: increase from 1960 to 1995, decrease thereafter
- Agrees with related observations
- Some skill in initialised predictions, but not in uninitialised predictions

Initialised hindcasts



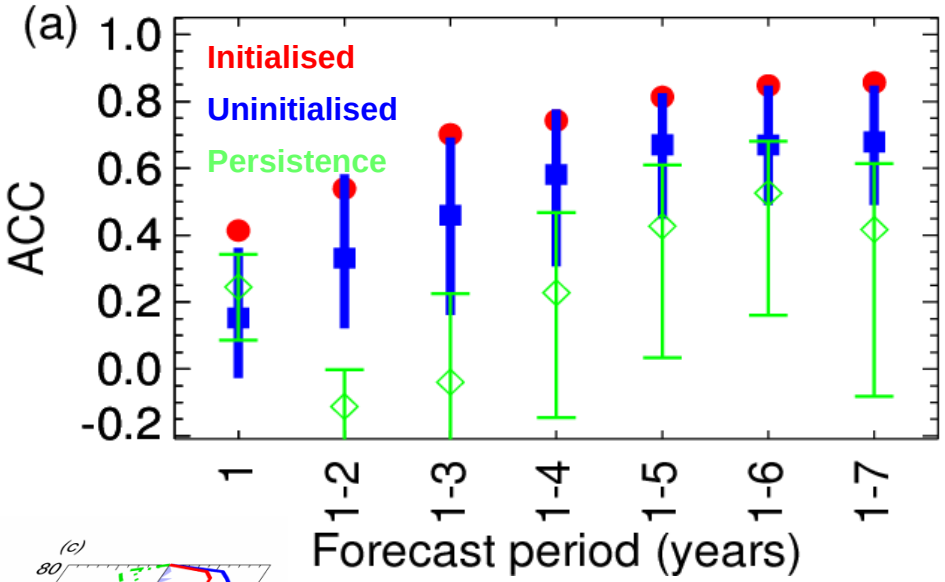
Uninitialised hindcasts

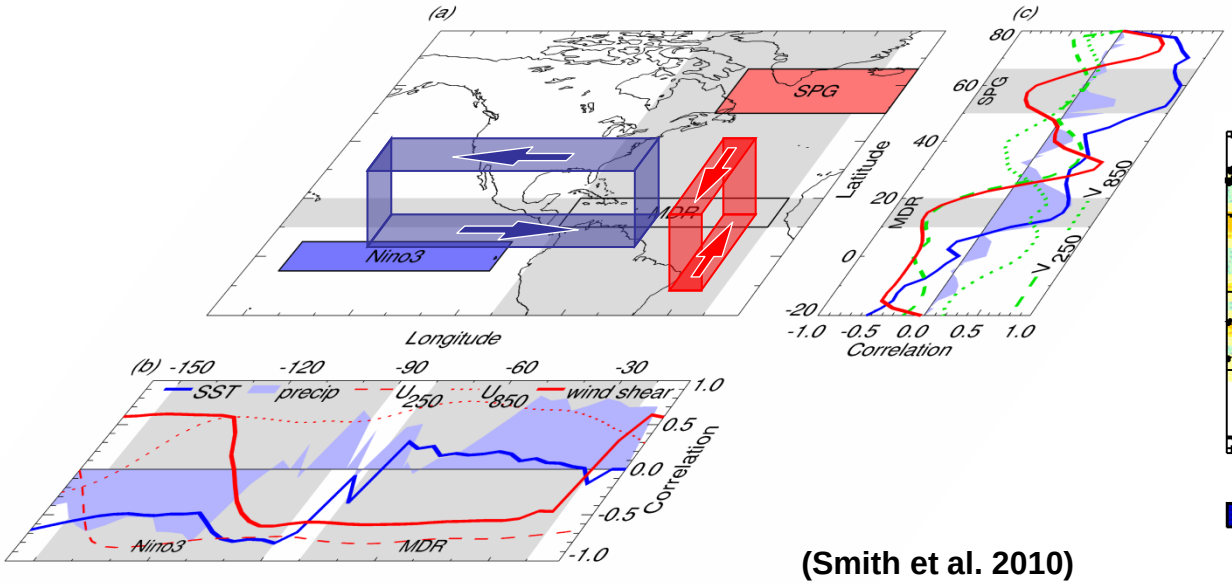


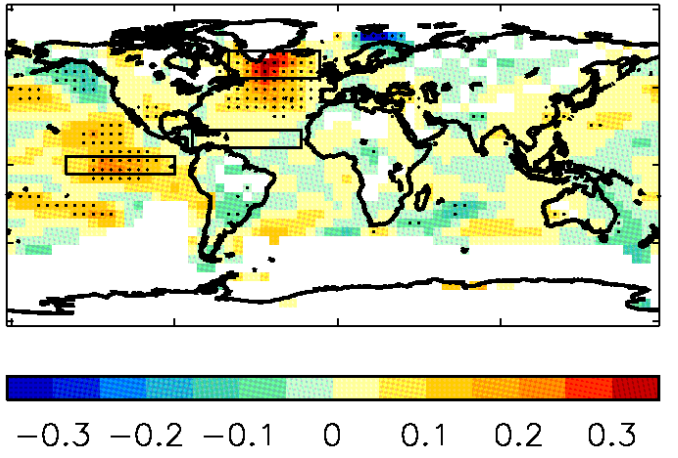
(Pohlmann et al. 2013)

Physical mechanisms: Atlantic tropical storm predictions

- Skill from both initialisation and external forcings
- Improved through initialisation
- Consistent with remote influences from improved SST predictions

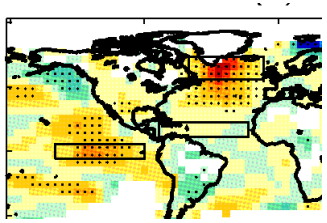


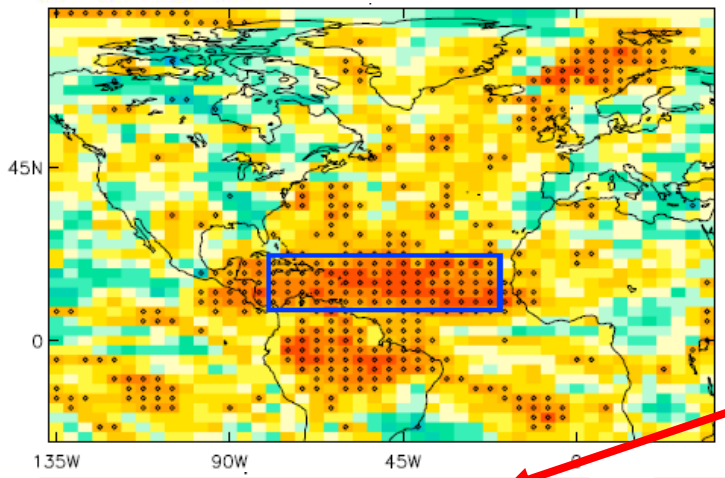


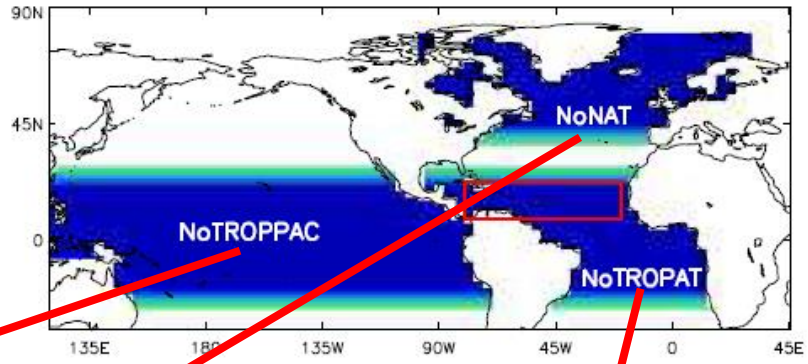


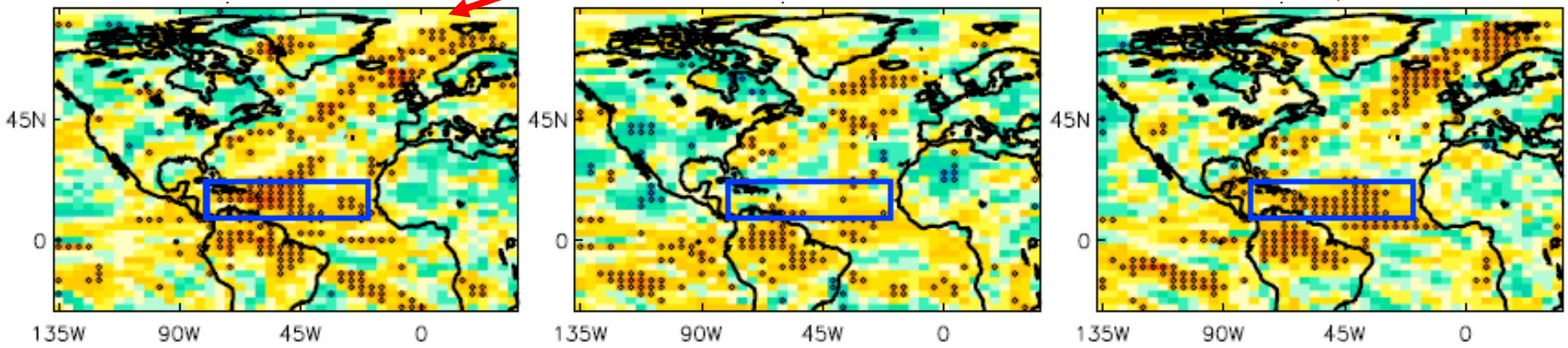
(Smith et al. 2010)

Remote influences on tropical Atlantic atmosphere









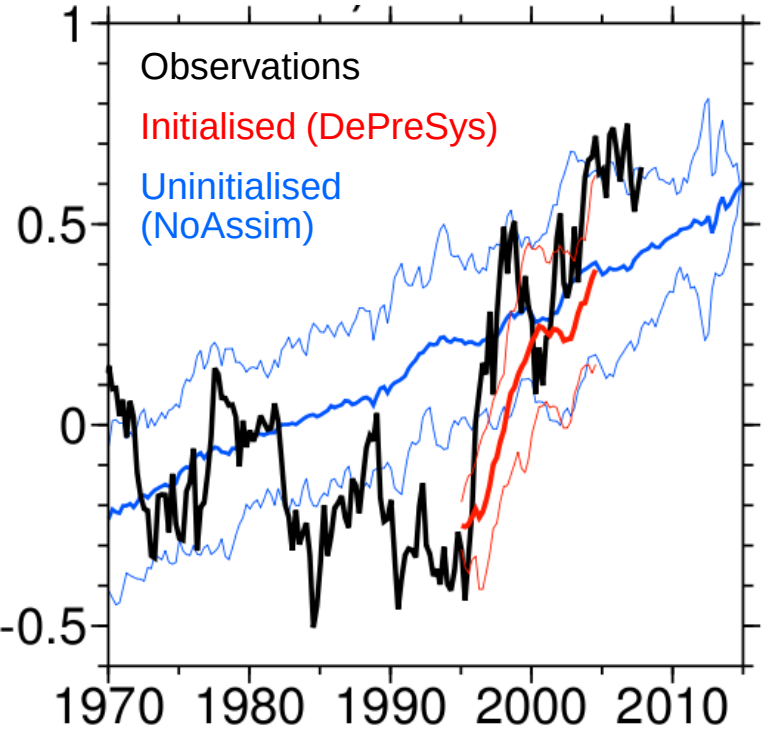
- Idealised predictions of rainfall 2 – 5 years ahead
- Tropical Atlantic atmosphere is relatively predictable
- Skill originates from sub-polar North Atlantic

Dunstone et al, 2011

North Atlantic sub-polar gyre (SPG)

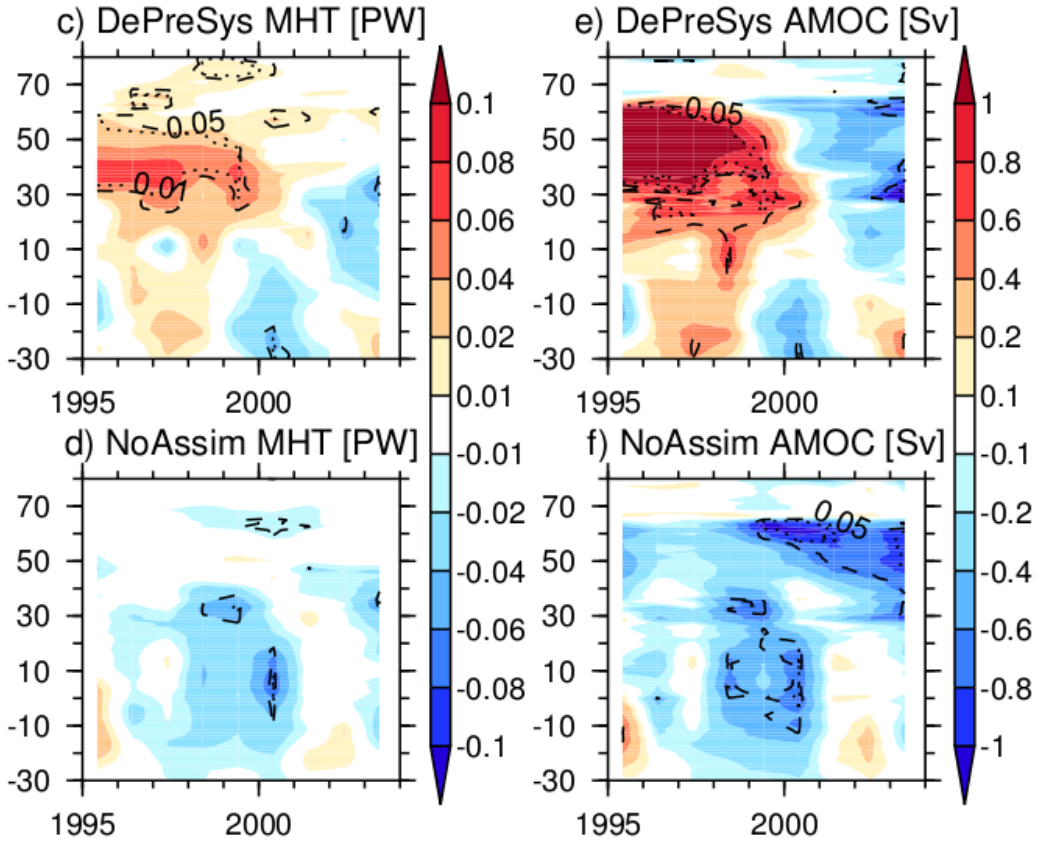
Met Office Hadley Centre

SPG 500m temp



Meridional heat transport

Overturning circulation



- Improved skill for 1995 rapid warming results from initialisation of increased Atlantic overturning circulation and meridional heat transport

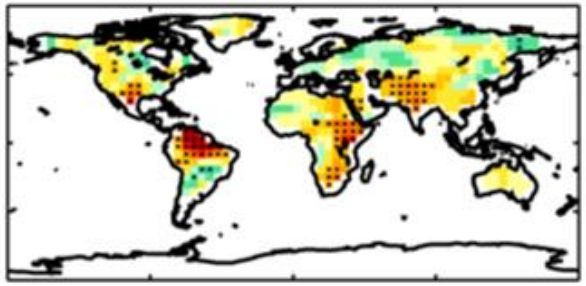
(Robson et al. 2012, also Yeager et al. 2012)

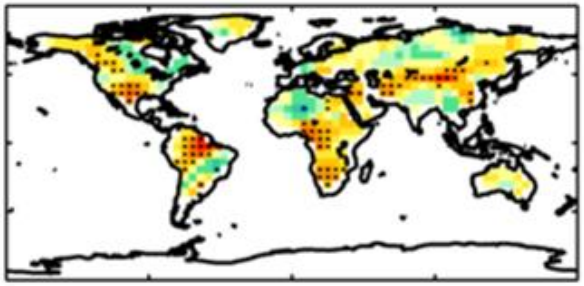
Summary

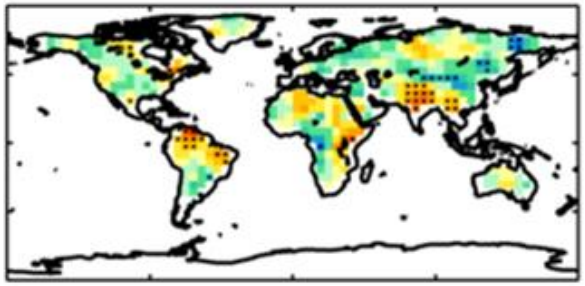
- Assessing skill is not easy!
- Models are not perfect
  - …but they may still provide useful guidance
  - Each ensemble member not necessarily a realisation of reality
- Need to assess potential skill
  - but beware of trends!
  - look at time series and more than one skill measure
- And understand **physical mechanisms** to gain confidence in forecasts:
  - **skill measures alone are not enough**

Full field versus anomaly initialisation: DJF precipitation correlation

50 start dates (Nov 1st every year from 1960 to 2009)

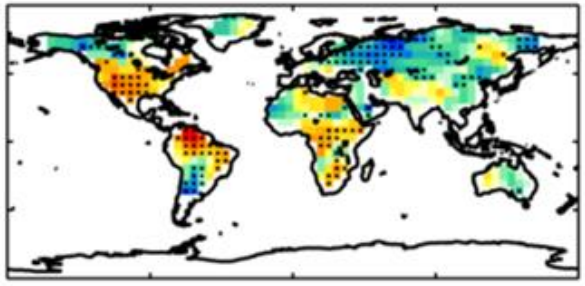


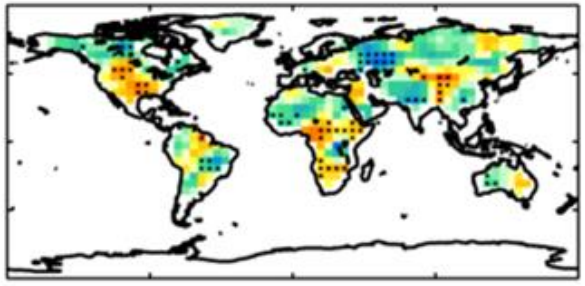


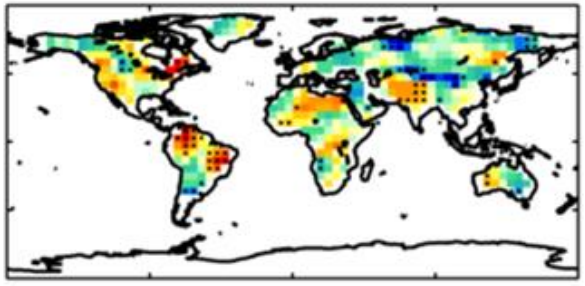


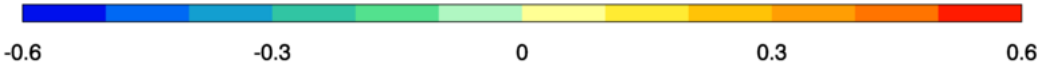
Full field | Anomaly | Difference

10 start dates (Nov 1st every 5 years from 1960 to 2005)









(Smith et al. 2013)