Calibration and validation of seasonal forecasts

Laura Ferranti

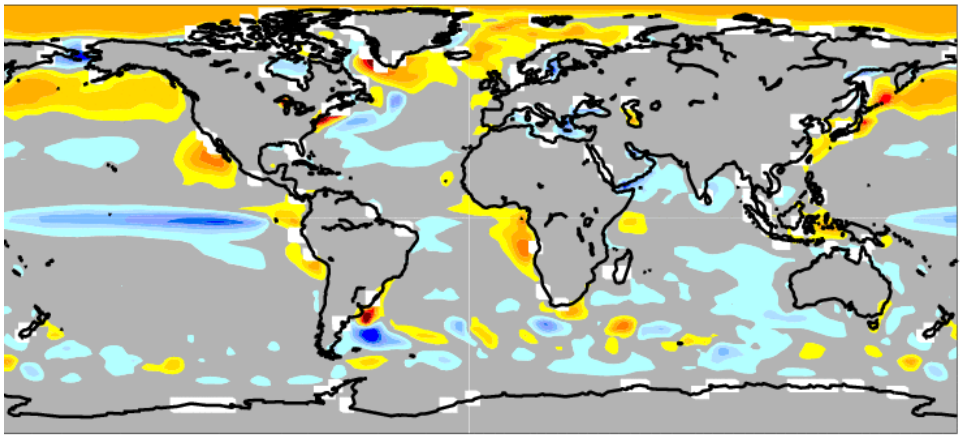
ECMWF

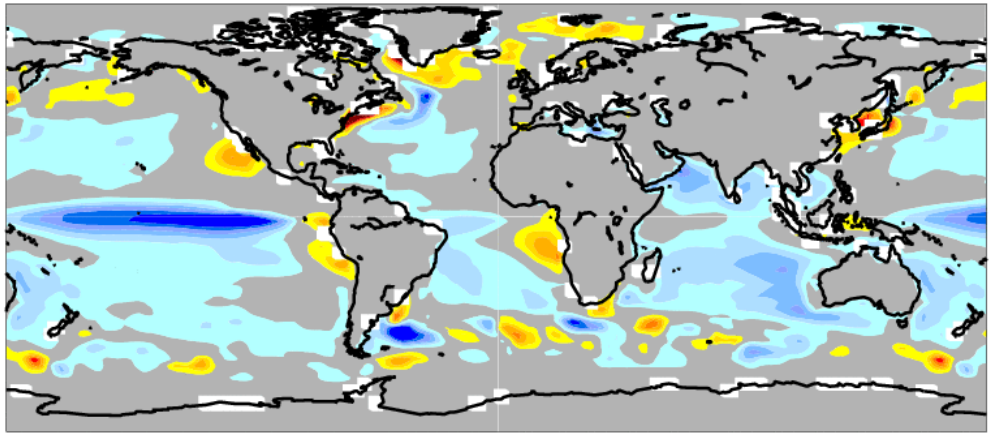
Outline:

- Calibration and systematic model errors
- How we assess the skill of the seasonal forecast
- Important outstanding issues that affect the skill assessment

SST Biases

forecasts initiated in May (1981-2010)





JJA
2-4 months

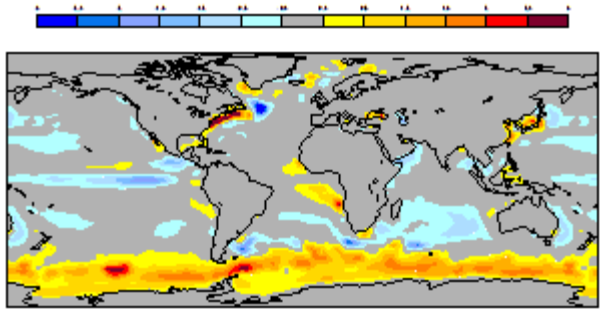
SON
5-7 months

ECMWF

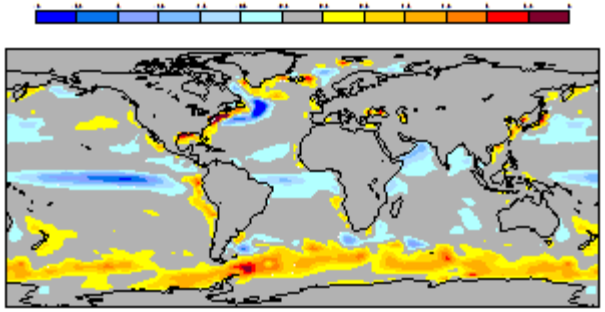
SST Biases for DJF

Biases from 4 independent coupled systems included in the EUROSIP multi-model (1996-2009)

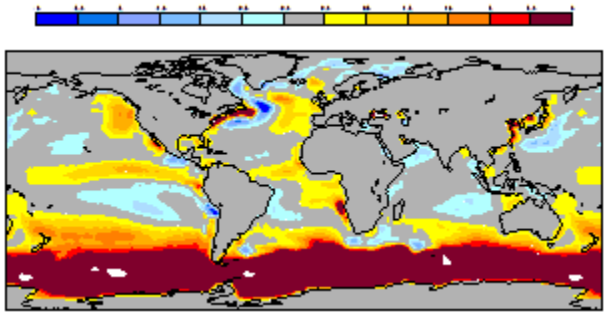
a



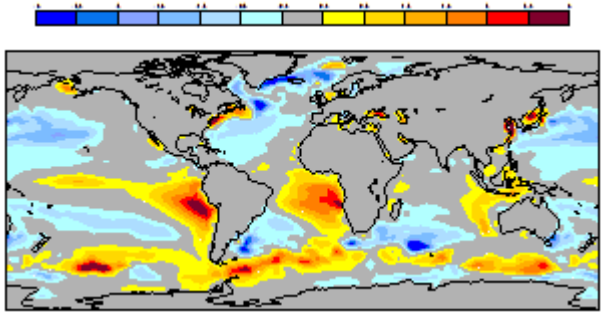
b



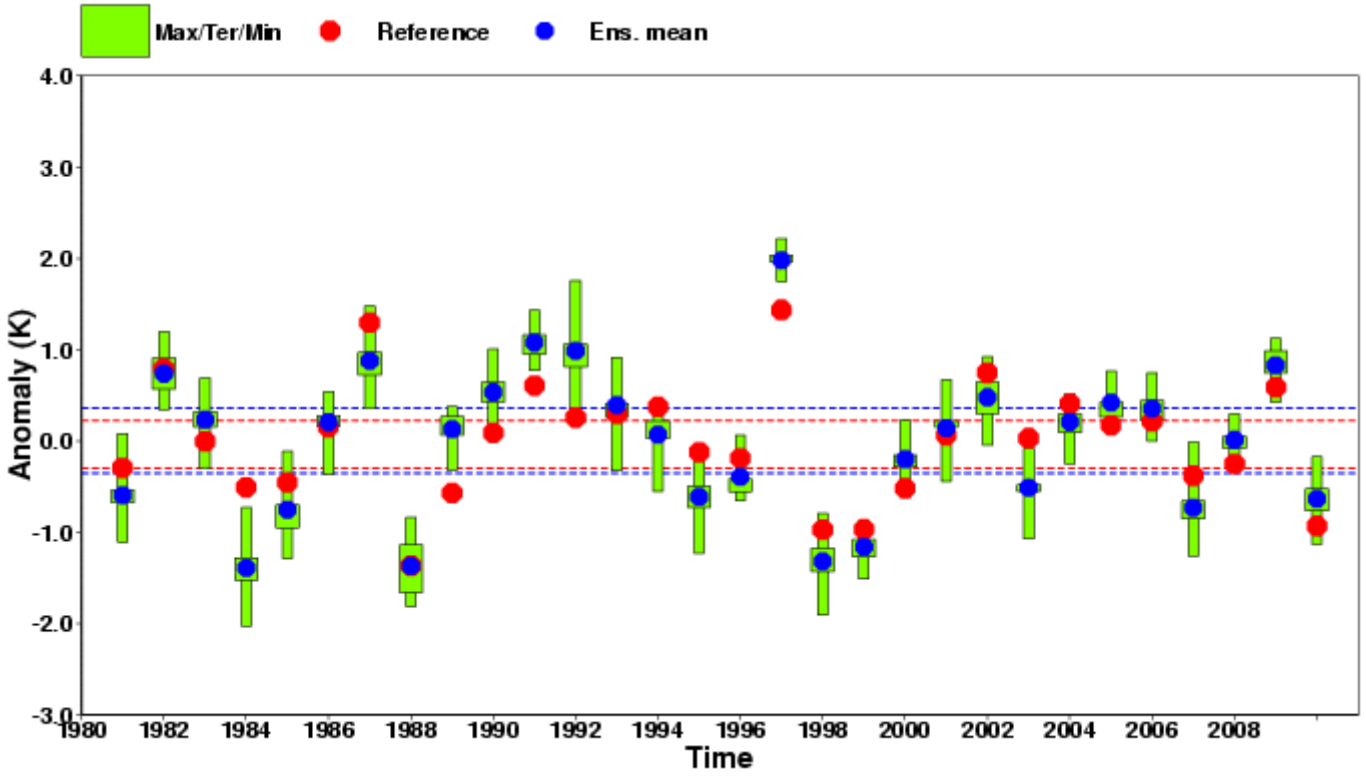
c



d



Nino 3.4 SST anomalies:



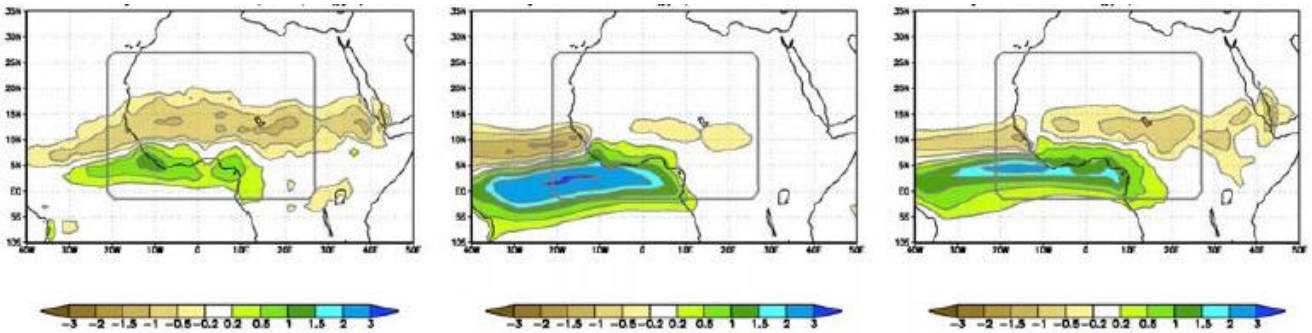
NINO 3.4	Correlation	Sd m/Sd obs
System 4	0.88	1.30

ECMWF

Assessing spatial errors : leading modes of rainfall variability

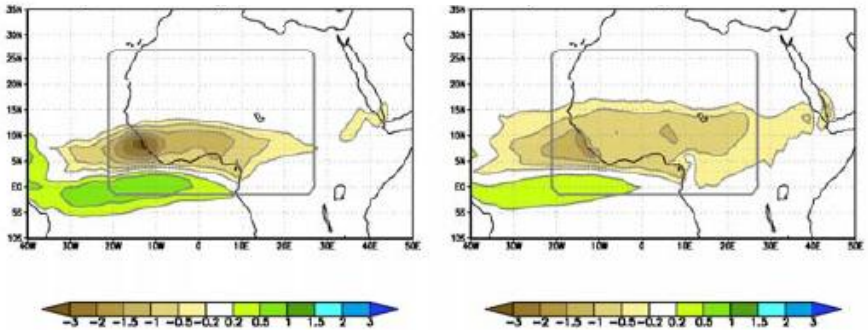
Observed System 3 r=0.33 System 4 r=0.71

EOF 1



r=0.0 r=0.44

EOF 2



Molteni et. al 2011

Figure 5.2.1 Left: Rainfall EOF-1 for West Africa from GPCP data. Centre: West Africa EOF-1 (top) and EOF-2 (bottom) from S3. Right: EOF-1 (top) and EOF-2 (bottom) from S4. The EOF domain is delimited by the grey box, shaded values are anomalies corresponding to 1 PC standard deviation.Correlation with GPCP EOF-1 is listed above each model EOF.

Calibration at ECMWF:

- Seasonal predictions for all the parameters are issued in term of anomalies.

- The amplitudes of the SST anomalies over the tropical Pacific (NINO indices) are re-scaled

ECMWF Calibration impact of variance correction:

Sys4
Sys4 cal.
Sys3

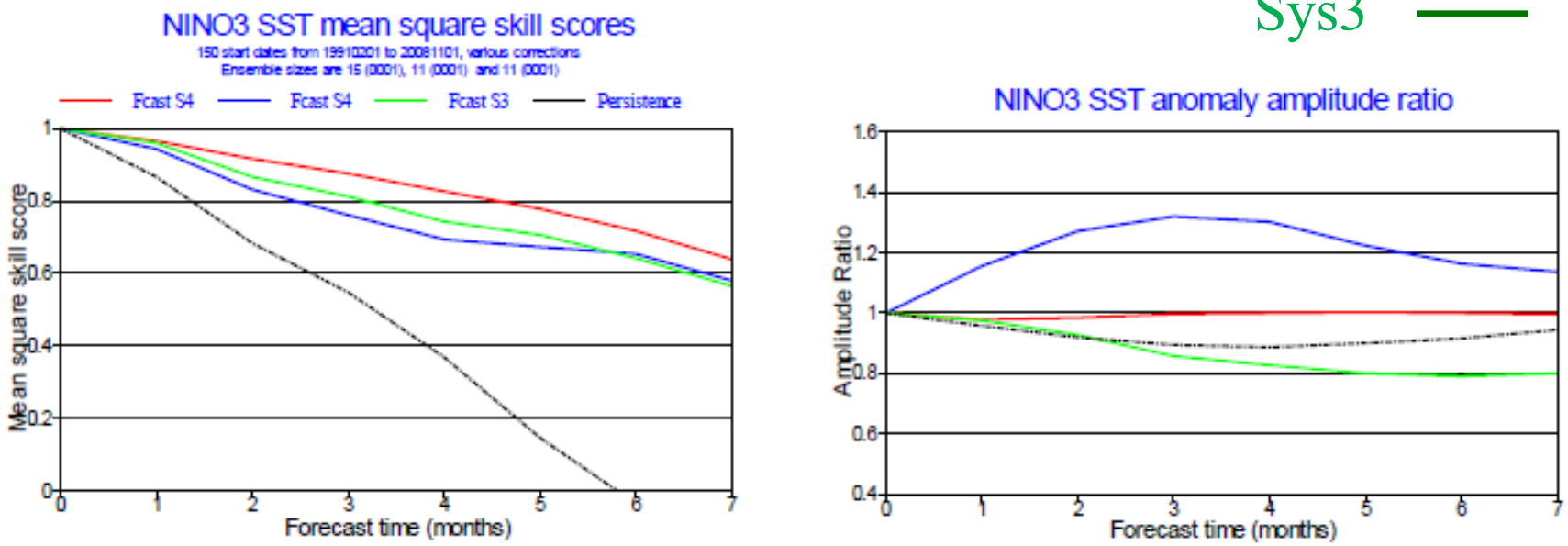


Figure3.2.2: NINO3 statistics in S4 with (red) and without (blue) variance correction and S3 (green). Left: mean-square skill score; right: anomaly amplitude w.r.t. observations.

Molteni et. al 2011

Calibration (Re-calibration):

- Calibration is a statistical adjustment of numerical forecasts to produce probabilistic forecasts that are more accurate (sharper and more reliable).

- Re-calibration adjusts the probability distribution produced by the model using information about its past performance.

- There is a large variety of Calibration methods. The choice of using a specific calibration procedure is conditioned by the users needs.

- Recalibration require long sets of stationary training data as well re-forecast data that covers a long period

Calibration :

- **Multi-model approach by combining output from several models, is an effective alternative to create calibrated probabilistic forecasts.**

- **The combination of several independent models widens the ensemble spread by sampling model errors.**

- **The multi-model forecast can better represent the full range of uncertainties. Its spread can represents better the unpredictable noise so that the multi-model forecast is more reliable .**

Re-forecasts:

- It is an integral part of the seasonal forecast system
- It is used to assess the systematic errors and therefore to calibrate the forecast
- It is used to assess the skill of the forecast
- It can be used to re-calibrate the forecast

Seasonal forecast skill assessment:

- A set of verification scores for deterministic and probabilistic forecast should be used.
- There is no single metric that can fully represent the quality of the probabilistic forecasts.
- The robustness of verification statistics is always a function of the sample size. WMO –SVSLRF suggests 20 years
- Typically verification is performed in cross-validation mode.
- The skill depends strongly on the season, so forecast evaluated separately for different starting months.

SST deterministic scores:

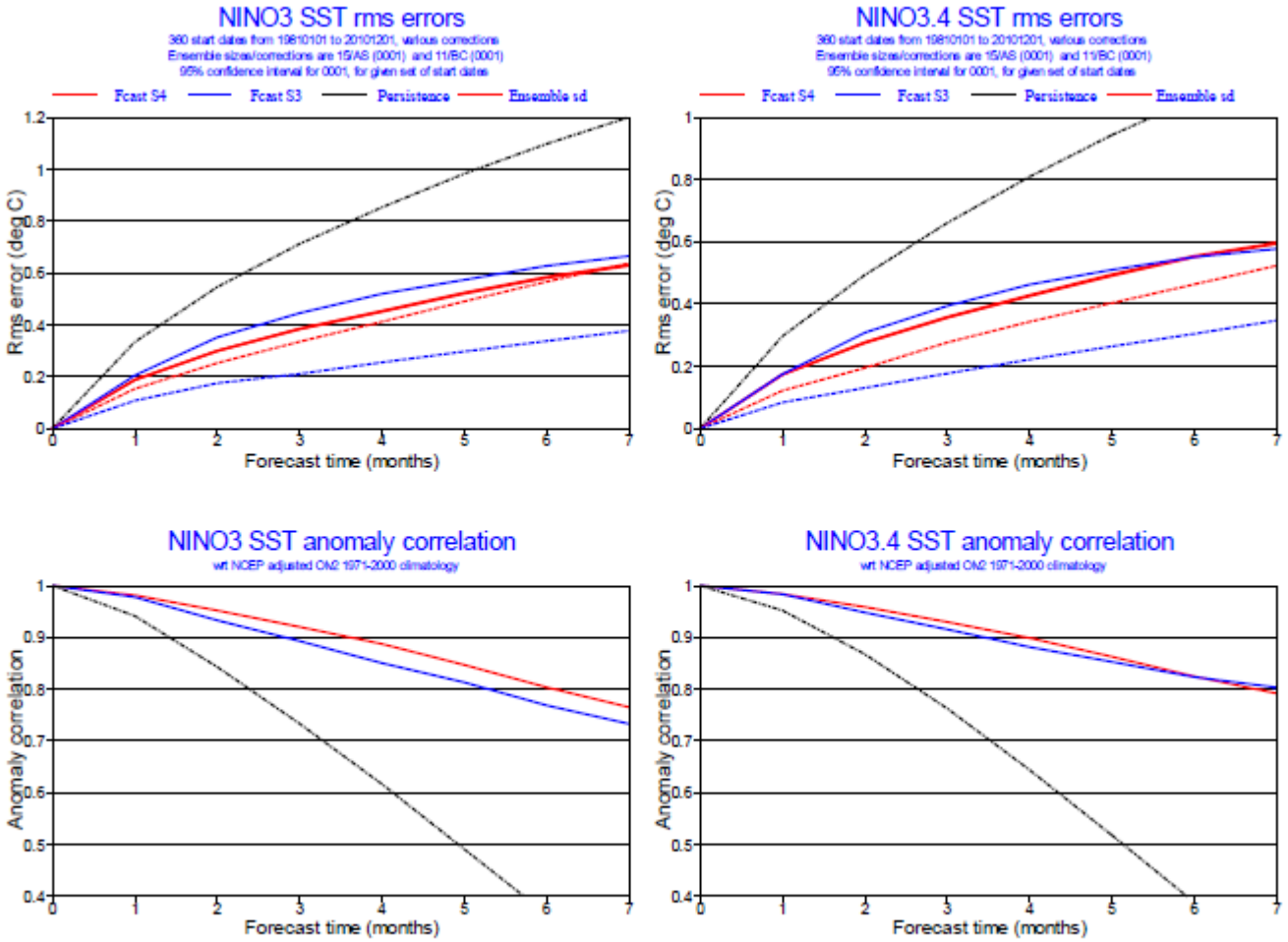


Figure4.1.1. S4 (red) and S3 (blue) NINO3 and NINO3.4 SST scores for the 30 year re-forecast period. S4 has decreased error (solid line) and increased ensemble spread (dashed line).

SST deterministic scores:

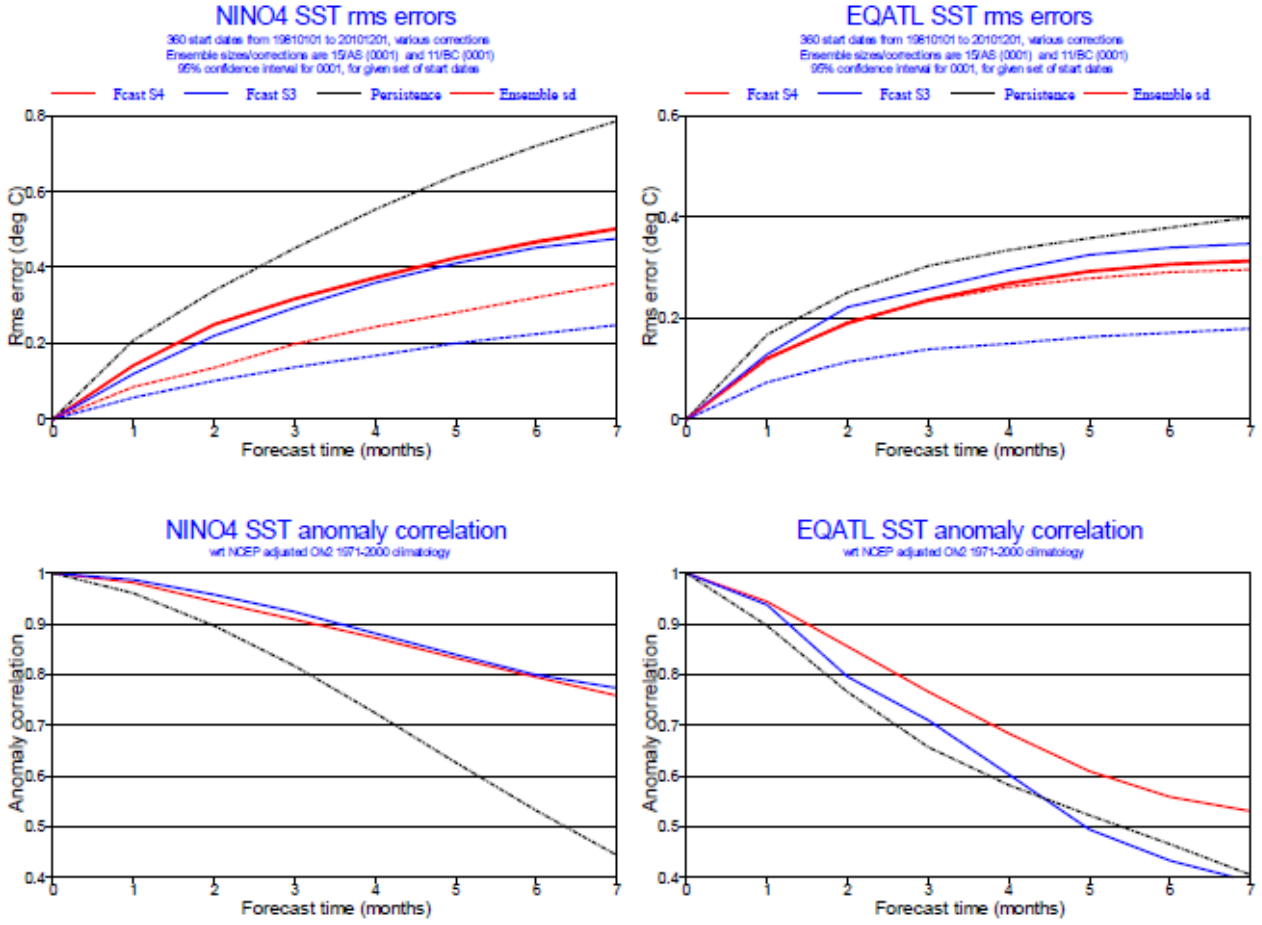
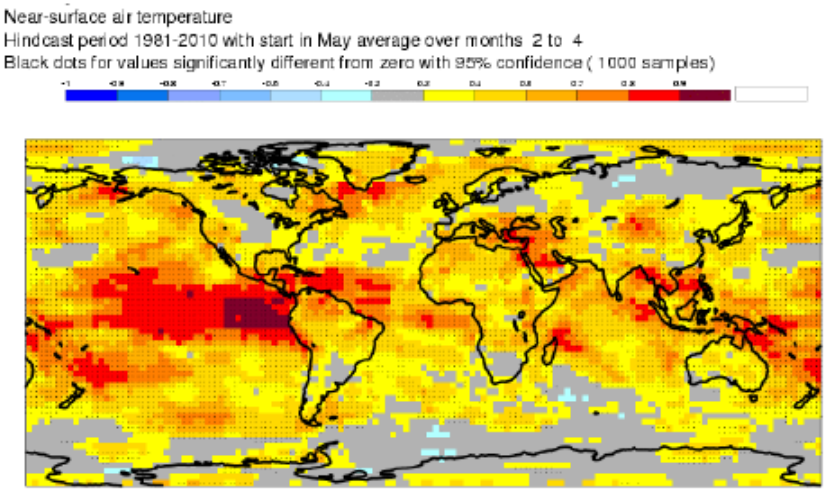


Figure4.1.2. As above, but for NINO4 and Equatorial Atlantic SST.

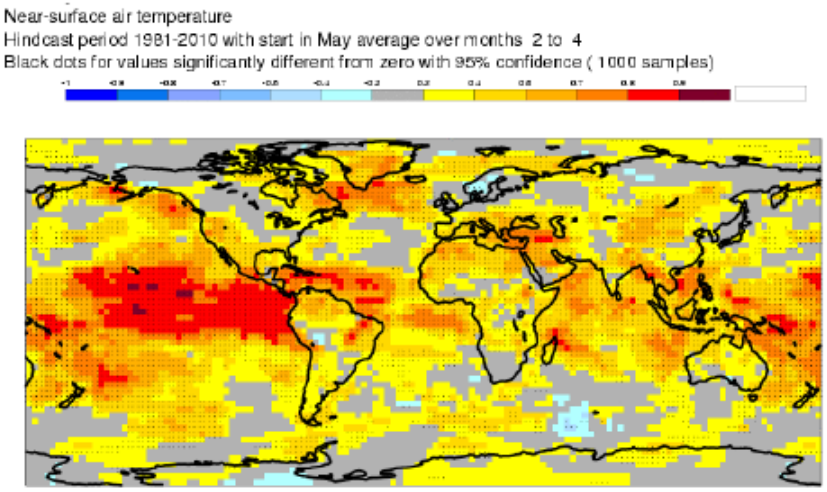
2m temp grid-point anomaly correlation:

Sys 4



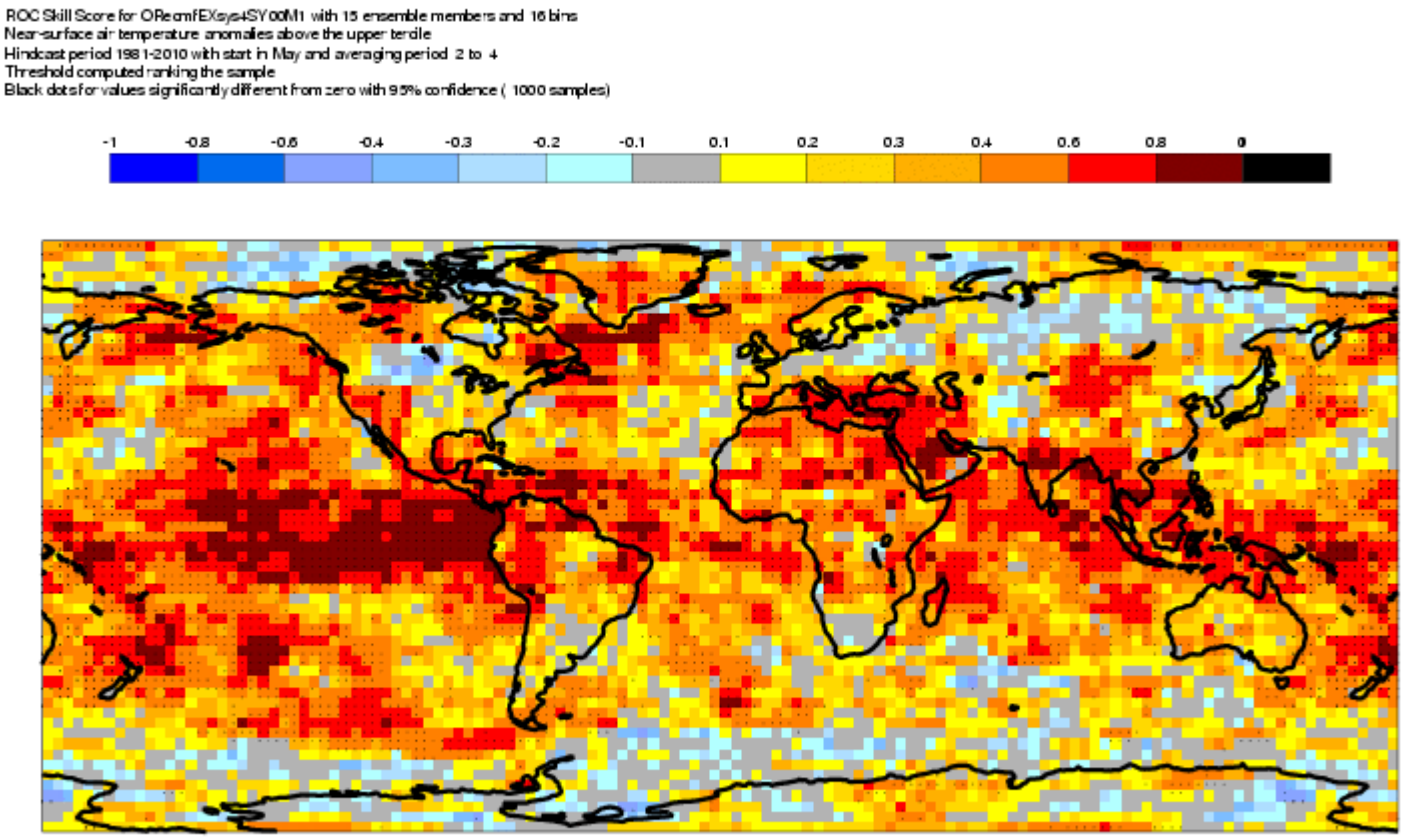
JJA month 2-4

Sys 3



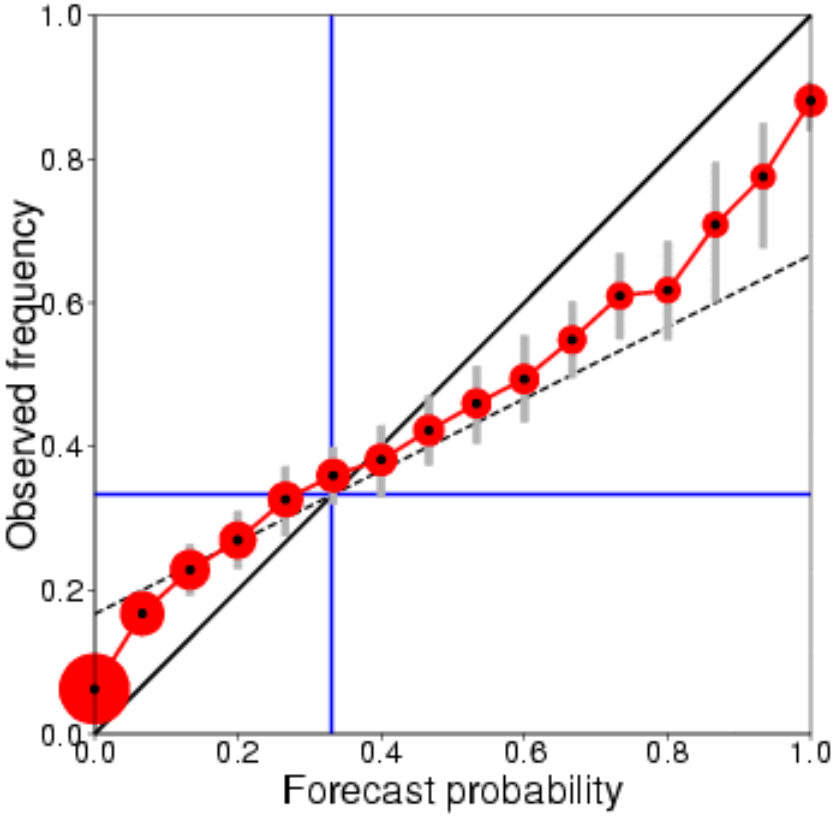
4.2.1: Ensemble-mean anomaly correlation for 2m_T in JJA: S4 (top), S3 (bottom).

Roc skill score



Reliability:

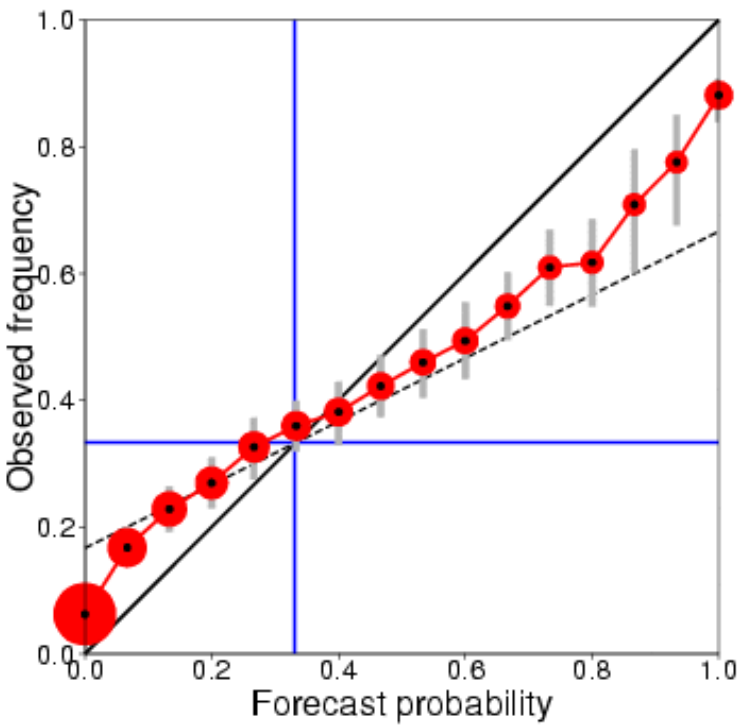
JJA 2m temp upper terc. Tropical band



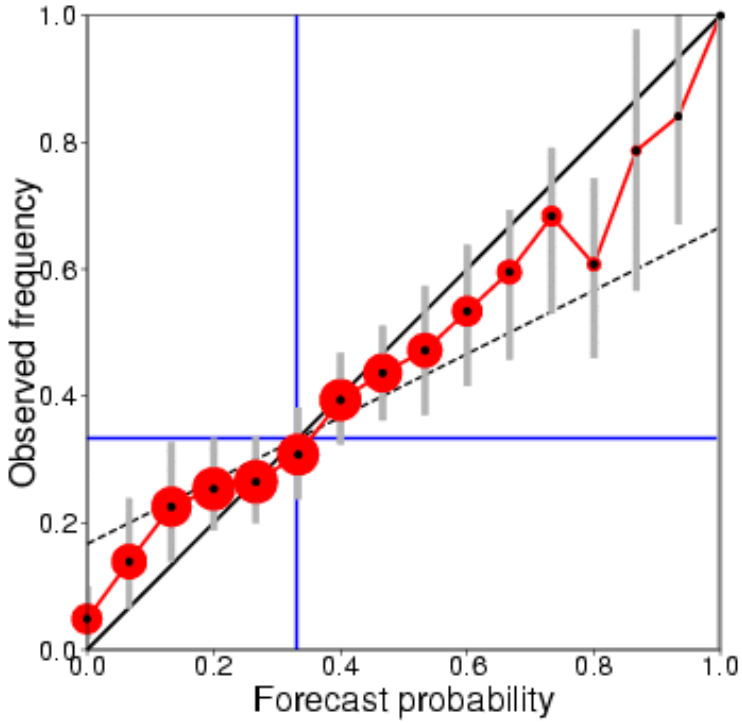
Reliability :

JJA 2m temp upper tercile

Tropical band



Europe



Seasonal forecast skill assessment:

- **The verification of ensemble forecasts requires a sufficient number of verification samples and involves the application of probabilistic skill metrics.**

- **Seasonal forecasts show high prediction skill in the tropics, particularly the ENSO region.**

- **Predictability is low in the extra-tropics.**

- **In central Europe, seasonal forecasts are at best only slightly better than climatology.**

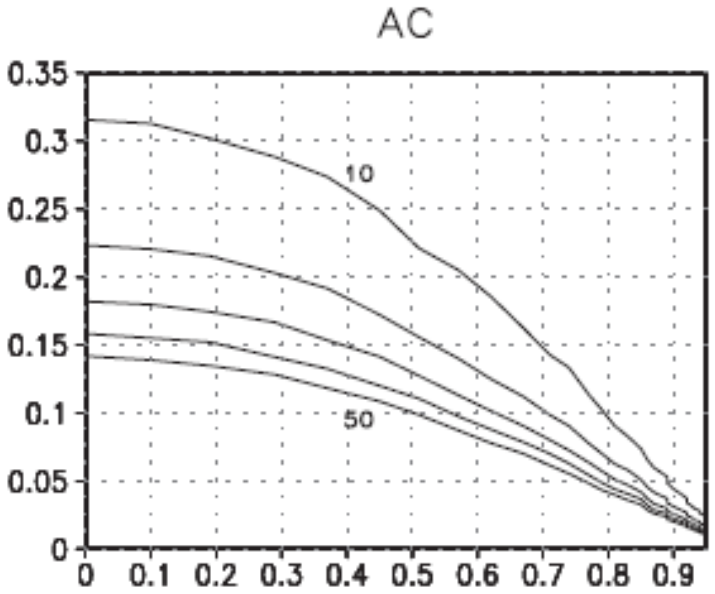
Seasonal forecast skill assessment:

- **The limitation associated with the sample size**
- **The effect of long term trend**
- **The effect of the ensemble size**

The limitation associated with of the sample size:

Variations in the spread of estimates of AC (y-axis) with the expected values of AC(x axis).

The differences in skill the AC estimates are due to the small verification time series. Spread is shown for verification size 10-50.



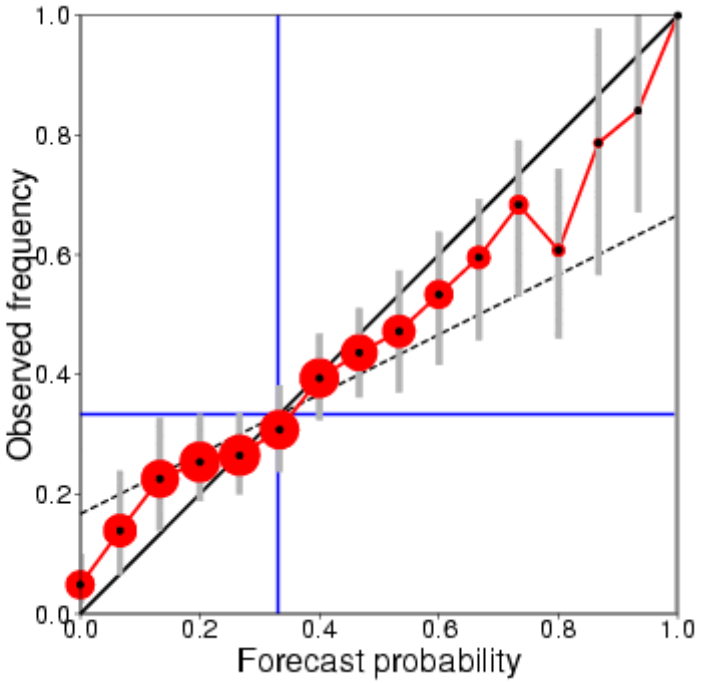
Kumar 2009

For an “accurate” estimate of deterministic skill over the tropics 20 years sample might be sufficient while over mid-latitudes a larger sample (>40 years) is needed.

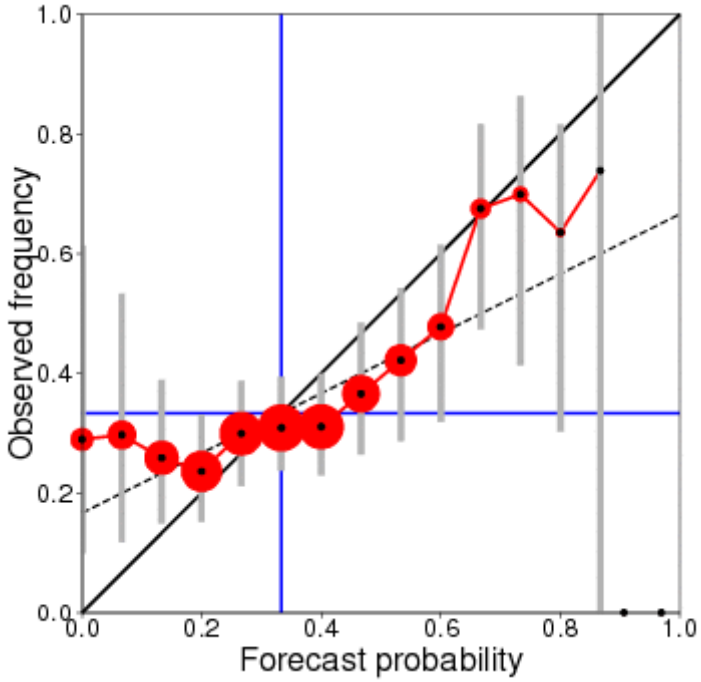
Sensitivity to the re-forecast period over Europe:

JJA - Reliability for 2m temp anomaly in the upper Terc.

1981-2010



1996-2010

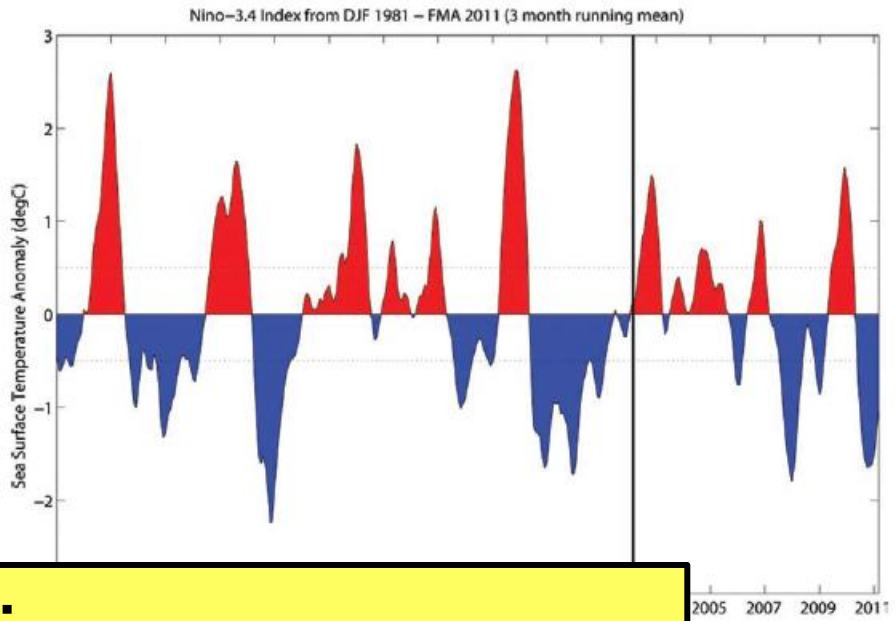


The limitation associated with of the sample size:

Barnston et al. 2012 analyzed the real-time ENSO prediction skills during the period 2002-2011.

The study showed that during the 9 years period (2001-2011) the ENSO events:

- had smaller amplitudes



aly with respect
egion for 1981–

Conclusions from Barnston :

Enso prediction skill is slightly higher using todays's models.

Decadal variability of ENSO predictability can hide the skill improvements

12

MWF

The limitation associated with of the sample size:

- The skill of the seasonal predictions is mainly associated with the ability to predict ENSO and its influence over remote regions (teleconnections)
- In the skill analysis we need to consider a sufficiently long period that sample the ENSO variability.
- The skill assessment of the seasonal predictions is based on the re-forecast performance.
- The size of the re-forecast (length and ensemble size) affect the skill estimates.
- Re-forecasts data should cover a period of at least 20 years.

Seasonal forecast skill assessment:

- The limitation associated with the sample size
- **The effect of long term trend**
- The effect of the ensemble size

The effect of long term trend in the sample

- The surface air temperature during the last 30 years exhibits a warming trend.

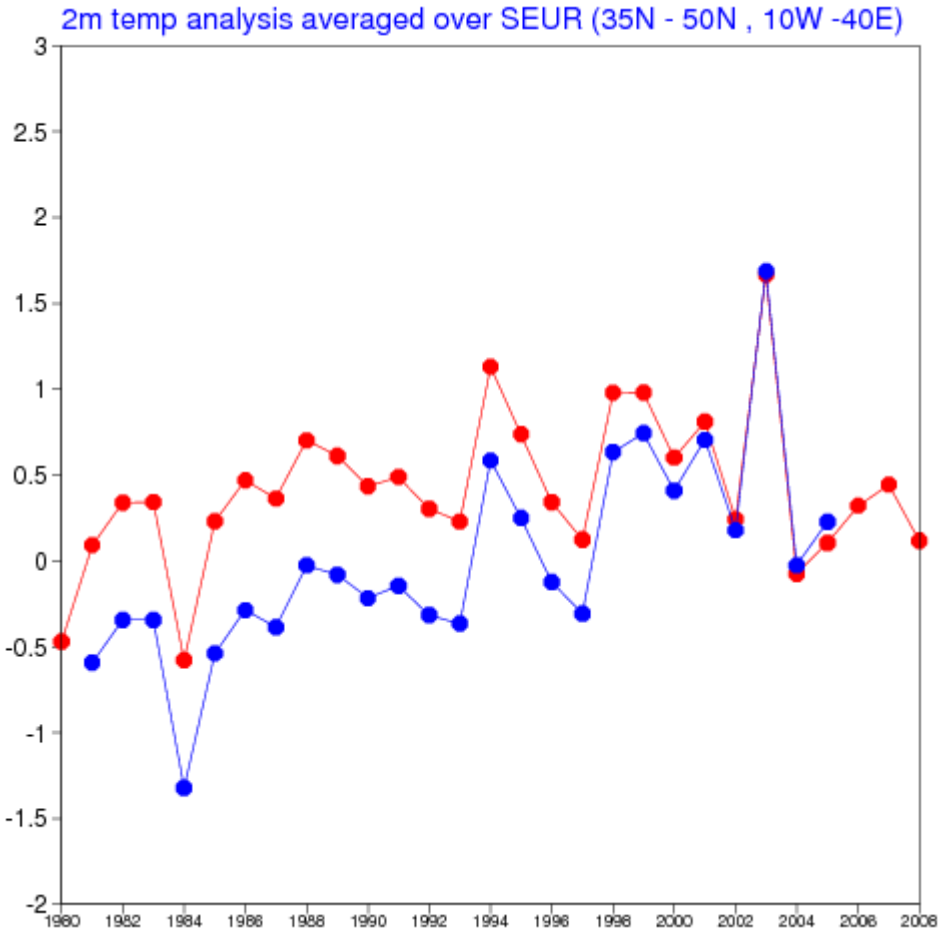
- This global warmth in the last decades is a continuation of the upward warming trend observed since the mid-20 century in response to the increase of GHGs (among others Hegerl at al. 2007, Hoerling et al. 2007).

- Several studies discussed the importance of an adequate representation of the GHGs in the coupled climate models used for seasonal predictions (Doblas-Reyes at al. 2006, Lininger et al. 2007, Cai et al. 2009)

The effect of long term trend in the verifying sample

- **In the skill assessment how we can distinguish the ability of reproducing the effect of climate change from the ability of predicting the year-to-year variations of anomalies?**

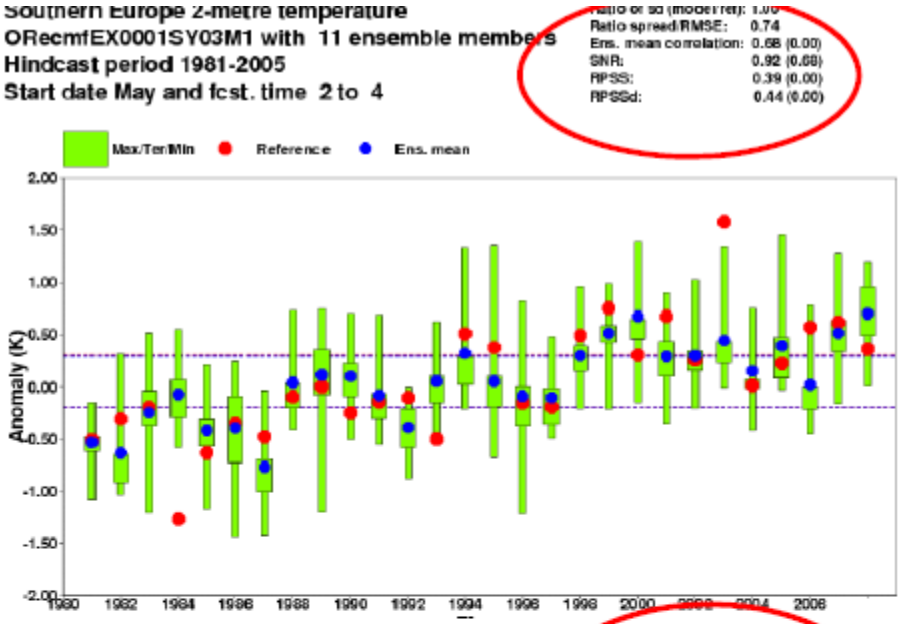
Verification with a moving climate to filter out the effects of long term trends:



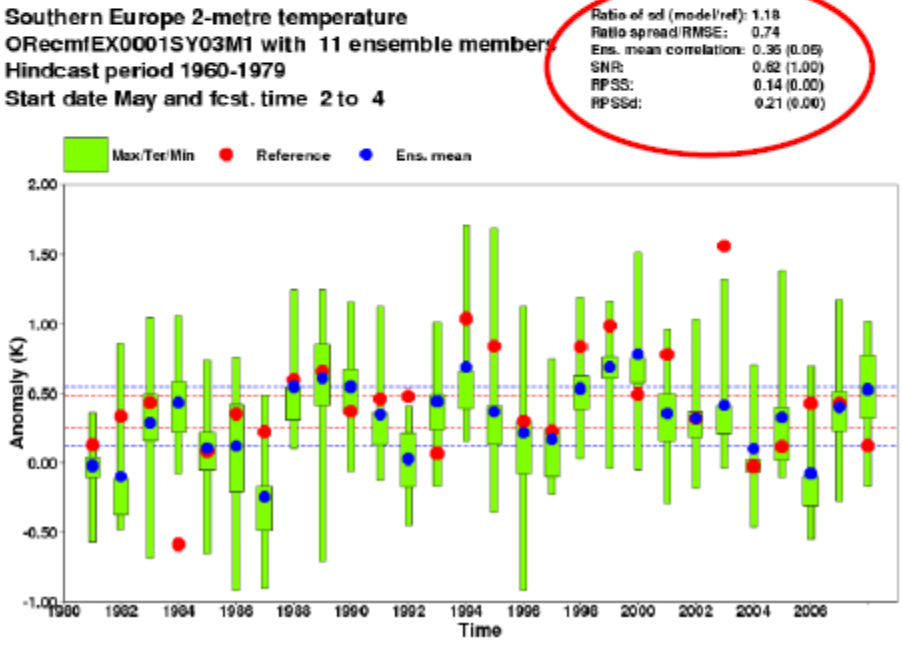
Anomalies with respect to a fixed climate (1981-2005)

Anomalies with respect to a moving climate (1960-1979, 1961-1980, ……..1988-2007)

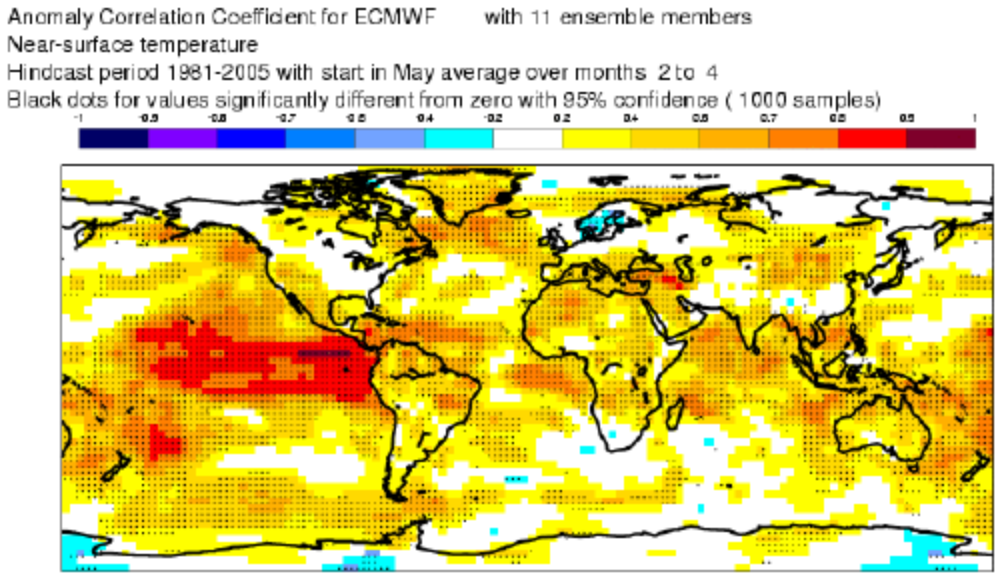
ECMWF



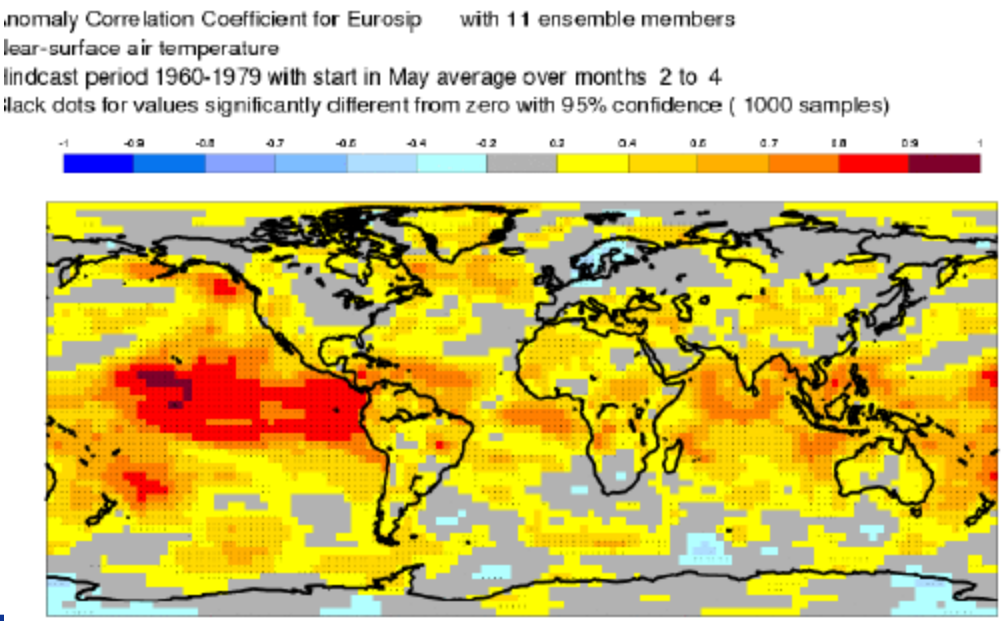
Fixed climate



Moving climate



Fixed climate



Moving climate

Seasonal forecast skill assessment:

- The limitation associated with the sample size
- The effect of long term trend
- **The effect of the ensemble size**

The relevance of the ensemble size:

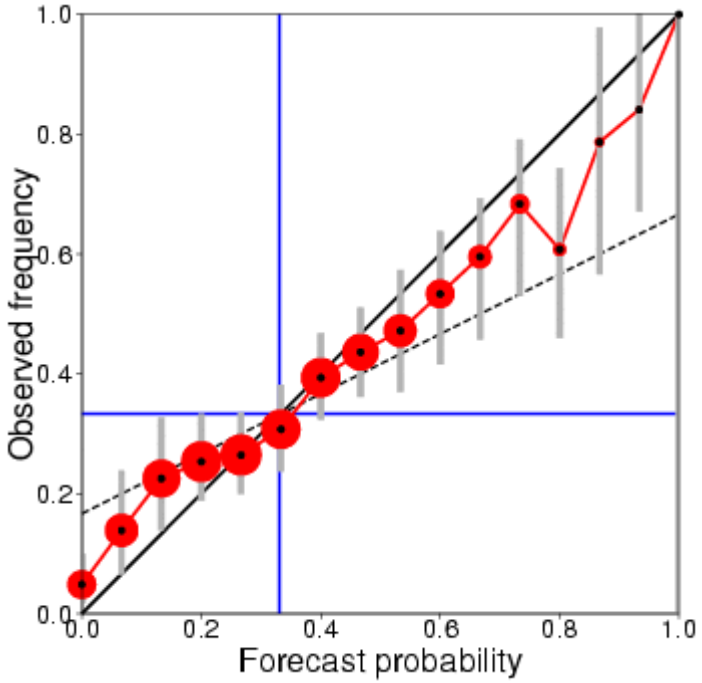
Several authors have studied the dependence on ensemble size of the probabilistic scores. (e.g. Richardson 2001; Kumar et al. 2001, Mason 2004, Müller et al. 2005, Ferro 2007)

Kumar et al. 2001 showed that the ensemble size of 10-20 members is sufficient to estimate the skill only for moderate ENSO cases.

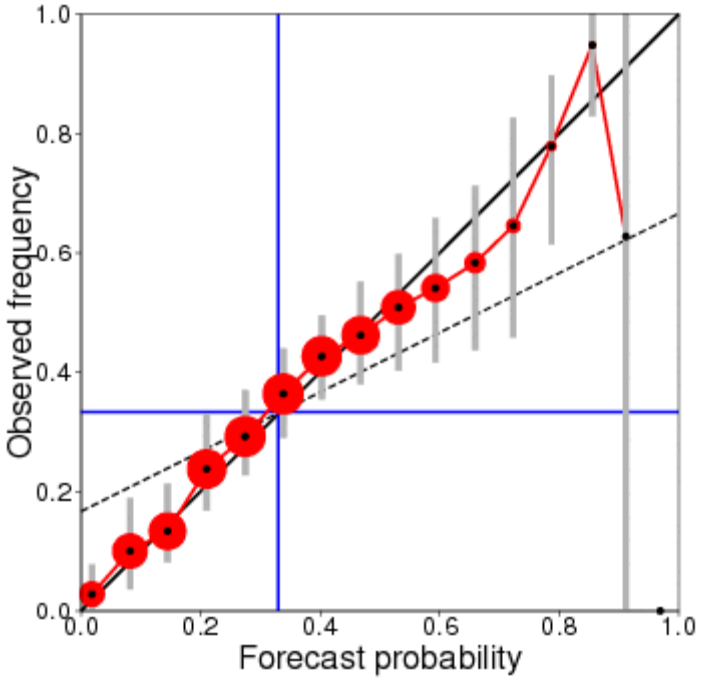
Müller et al. 2005 and Weigel et al. 2007 suggested the use of a de-biased Brier and ranked probability skill score to avoid the dependency on the ensemble size and to assess forecast with small ensemble size.

Sensitivity to ensemble size:

15 ensembles



31 ensembles



ECMWF

The relevance of the ensemble size:

- **Because of the atmospheric internal variability the seasonal predictability is limited.**

- **Ideally to estimate this upper limit in the average skill an ensemble with an infinite size will be needed.**

- **In reality the seasonal predictions are done with a limited ensemble size particularly for the re-forecast (**typically 10-25 members**)**

Verification or Validation?

Validation is a more general term, less quantitative than verification.

In the validation we can include any assessment of the model the climate statistics (e.g. NAO frequency …leading mode of variability etc.)

Verification assess the accuracy of a time-series of forecasts by comparing with a corresponding time-series of observations.

From Laurie Wilson:

to validate is to check that one is doing the right things,
to verify is to check that one is doing things right