The Seasonal Forecast System at ECMWF

Tim Stockdale

European Centre for Medium-range Weather Forecasts

t.stockdale@ecmwf.int

Contents

- Seasonal forecasting with coupled GCMs
  - Why use GCMs?
  - How we make a forecast
  - Basic calibration – how, why, and what are the problems?
  - Model error

- Operational forecasts: ECMWF System 4 (and preview of SEAS5)
  - System design
  - How good are the El Nino forecasts?
  - How good are the atmospheric forecasts?

- Forecast interpretation
  - Meaning of probabilistic forecasts
  - Outlook

Sources of seasonal predictability

– KNOWN TO BE IMPORTANT:
  - El Nino variability - biggest single signal
  - Other tropical ocean SST - important, but multifarious
  - Climate change - impact is substantial in many regions
  - Local land surface conditions - e.g. soil moisture in spring

– OTHER FACTORS:
  - Volcanic eruptions - definitely important for large events
  - Mid-latitude ocean temperatures - complicated
  - Remote soil moisture/ snow cover - evidence stronger in some cases
  - Sea ice anomalies - local effects, but remote?
  - Dynamic memory of atmosphere - most likely on 1-2 months
  - Stratospheric influences - polar vortex, solar cycle, QBO, ozone, …

– Unknown or Unexpected - ???

Methods of seasonal forecasting

- Empirical forecasting
  - Use past observational record and statistical methods
  - Works with reality instead of error-prone numerical models ☺
  - Limited number of past cases means that it works best when observed variability is dominated by a single source of predictability ☹
  - A non-stationary climate is problematic ☹
- Two-tier forecast systems
  - First predict SST anomalies (ENSO or global; dynamical or statistical)
  - Use ensemble of atmosphere GCMs to predict global response
  - Some people still use regression of a predicted El Nino index on a local variable of interest
- Single-tier GCM forecasts
  - Include comprehensive range of sources of predictability ☺
  - Predict joint evolution of SST and atmosphere flow ☺
  - Includes indeterminacy of future SST, important for prob. Forecasts ☺
  - Model errors are an issue! ☹

Step 1: Build a coupled model

- IFS (atmosphere)
  - $T_L255L91$ Cy36r4, 0.7° grid for physics, full stratosphere
  - Modifications to stratospheric physics and lakes
  - Singular vectors from EPS system to perturb atmosphere initial conditions
  - Ocean currents coupled to atmosphere boundary layer calculations

- NEMO (ocean)
  - Global ocean model, 1x1 mid-latitude resolution, 0.3 near equator
  - Sophisticated 3D-VAR ocean analysis system, including analysis of salinity, multivariate bias corrections and use of altimetry.

- Coupling
  - Fully coupled, no flux adjustments, except no physical model of sea-ice

Step 2: Make some forecasts

- Initialize coupled system (cf. Magdalena's lecture earlier today)
  - Aim is to start system close to reality. Accurate SST is particularly important, plus ocean sub-surface.
  - Don't worry too much about "imbalances", want to minimize non-linear errors

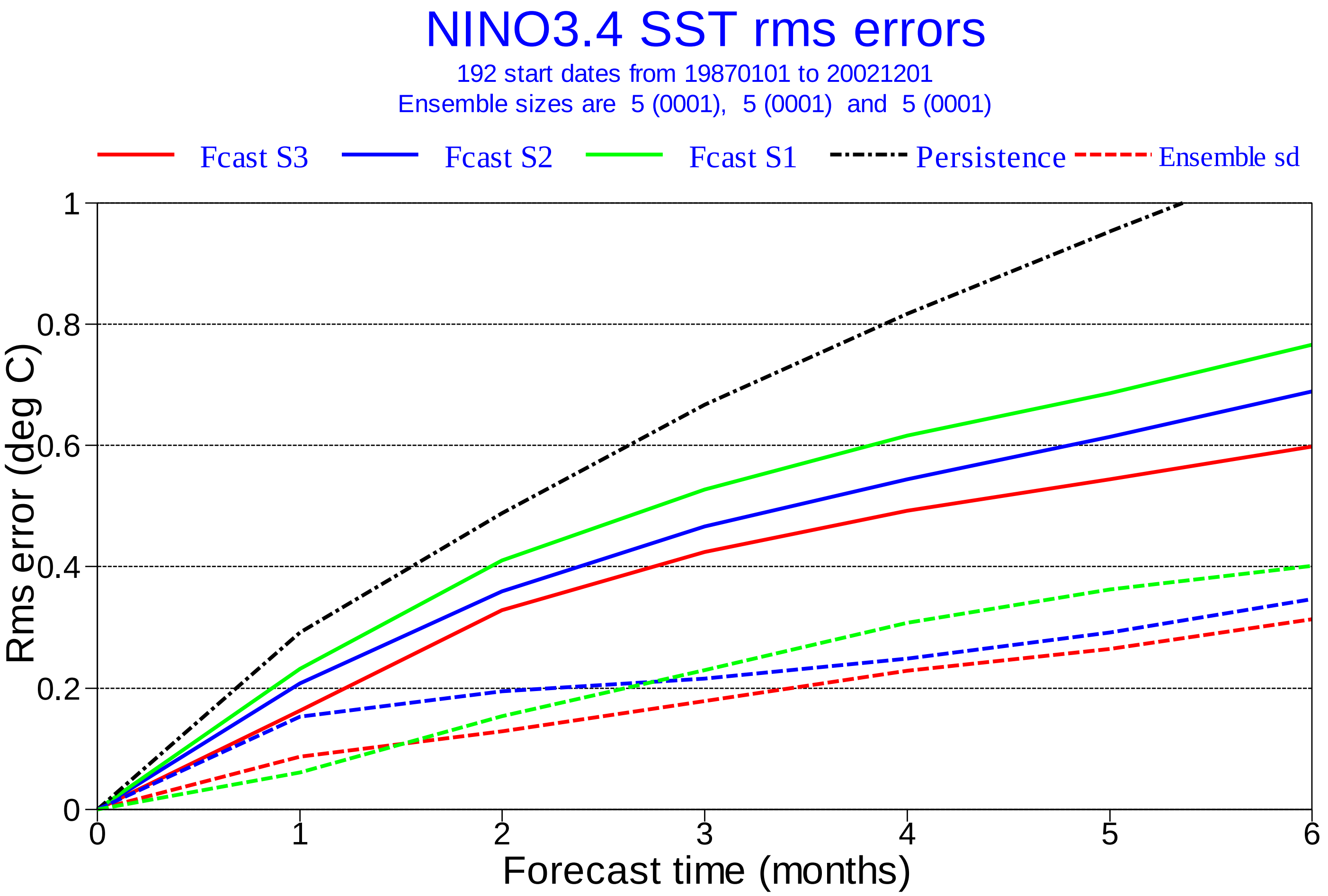
- Run an ensemble forecast
  - Explicitly generate an ensemble on the 1st of each month, with perturbations to represent the uncertainty *in the initial conditions*; run forecasts for 7 months
  - Stochastic physics to represent *indeterminacy* of large scale (possibly more)

- Worry about model biases later ….

Creating the ensemble

- Wind perturbations
  - Perfect wind would give a good ocean analysis, but uncertainties are significant. We represent these by adding perturbations to the wind used in the ocean analysis system.
  - BUT only have 5 member ensemble, and only limited representation of other sources of uncertainty in ocean analysis (e.g. observation error)
- SST perturbations
  - SST uncertainty is not negligible
  - SST perturbations added to each ensemble member at start of forecast.
  - (Some question marks on consistency of SST)
- Atmospheric unpredictability
  - Atmospheric ‘noise’ soon becomes the dominant source of spread in an ensemble forecast. This sets a fundamental limit to forecast quality.
  - To ensure that noise grows rapidly enough in the first few days, we activate ‘stochastic physics’ and use EPS singular vectors.
  - In System 4, stochastic physics increases spread at all timescales.

RMSE and spread in different systems



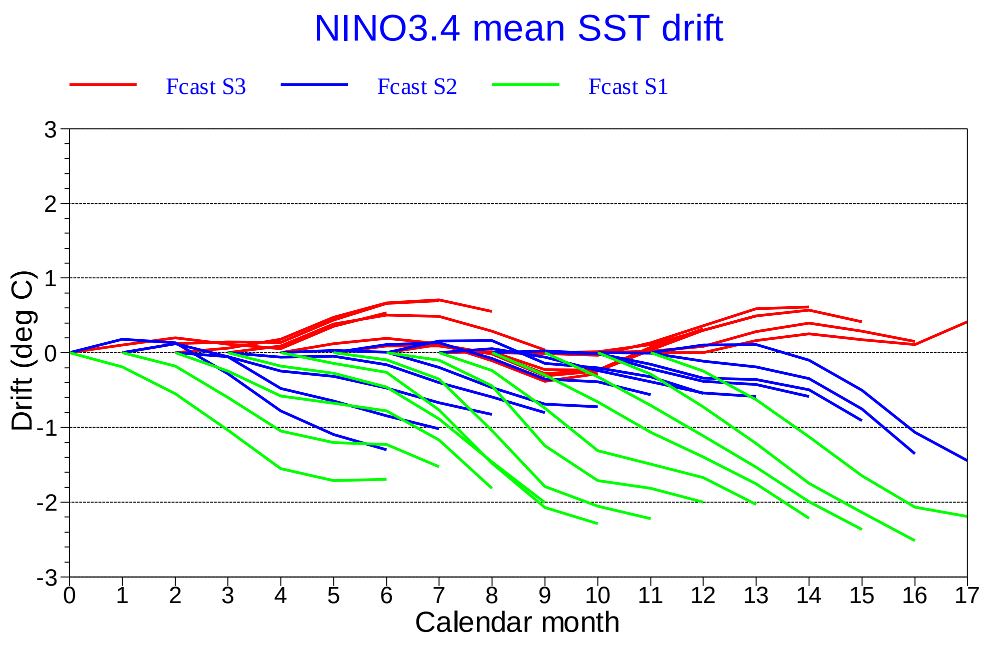
r.m.s. error of forecasts has been systematically reduced (solid lines) ….

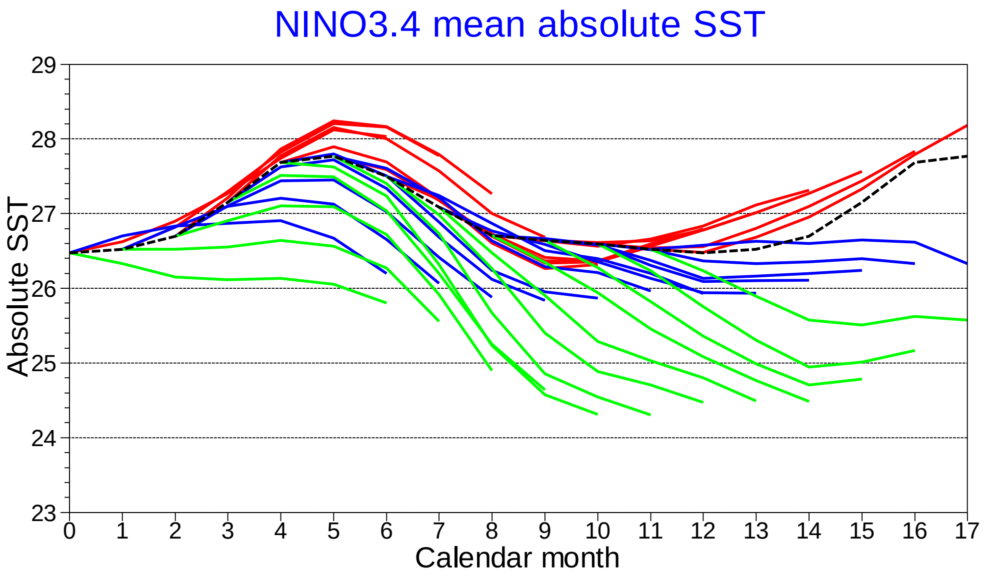
.. but ensemble spread (dashed lines) is still substantially less than actual forecast error.

Substantial amounts of forecast error are not from the initial conditions.

Step 3: Remove systematic errors

- Model drift is typically comparable to signal
  - Both SST and atmosphere fields
- Forecasts are made *relative* to past model integrations
  - Model climate estimated from 30 years of forecasts (1981-2010), all of which use a 15 member ensemble. Thus the climate has 450 members.
  - Model climate has both a mean and a distribution, allowing us to estimate e.g. tercile boundaries.
  - Model climate is a function of start date and forecast lead time.
  - EXCEPTION: Nino SST indices are bias corrected to absolute values, and anomalies are displayed w.r.t. a 1971-2000 climate.
- Implicit assumption of linearity
  - We implicitly assume that a shift in the model forecast relative to the model climate corresponds to the expected shift in a true forecast relative to the true climate, despite differences between model and true climate.
  - Most of the time, assumption seems to work pretty well. But not always.





SST bias is a function of lead time and season.

Some systems have less bias, but it is still large enough to require correcting for.

Nino plumes: variance scaling

- Model Nino SST anomalies in S4 have **too large amplitude**
- Problem is especially acute in boreal spring and early summer (model bias of “permanent La Nina” does not allow spring relaxation physics to apply; this was something S3 did very well)

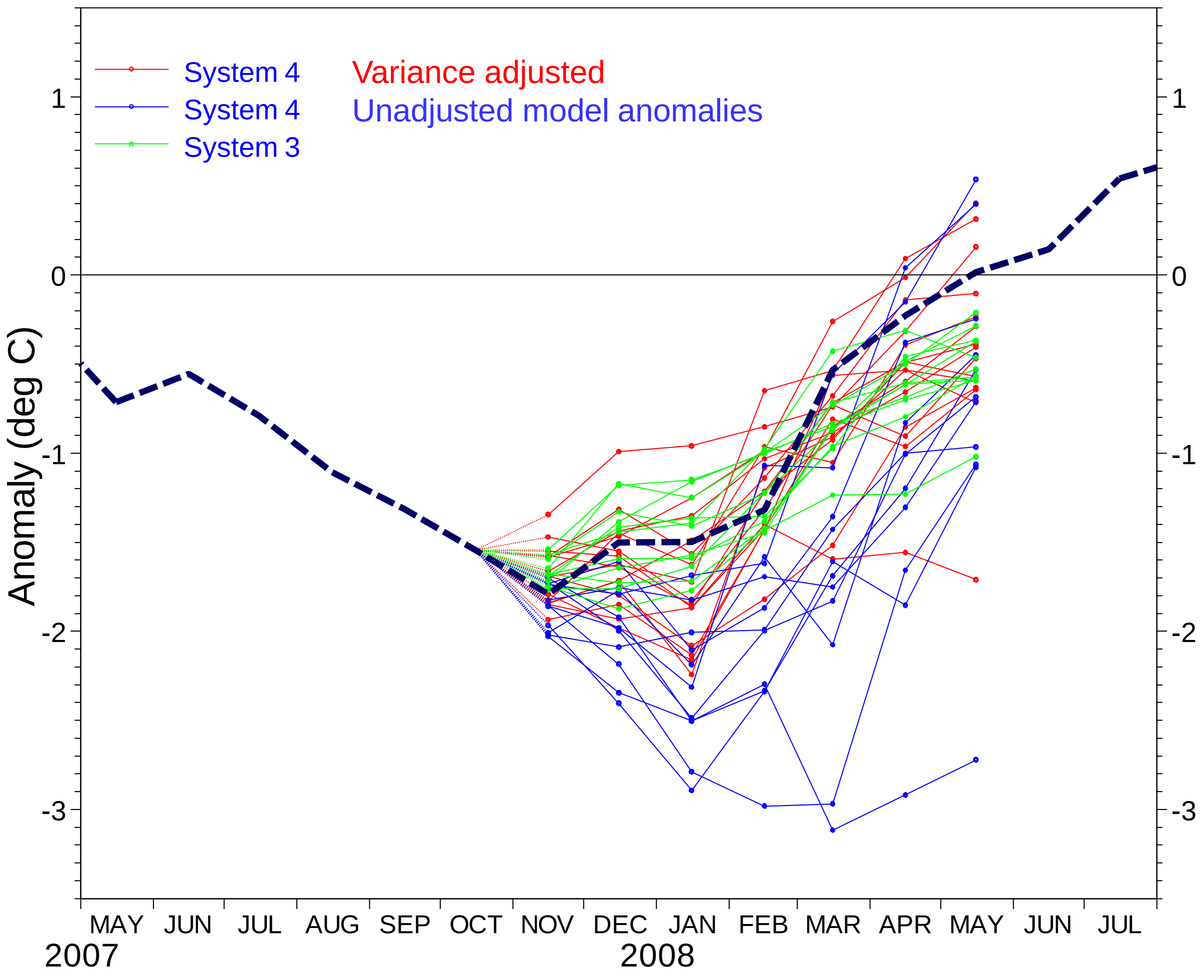
- We plot the “Nino plumes” corrected for both mean **and** variance, instead of just the mean.
- This is done by scaling the model anomalies so that the model variance matches the observed variance in the calibration period
- We use the same approach (cross-validated) when calculating scores

- This affects the *plotting*, not the model data itself
- The spatial maps are not affected: the tercile and quintile probability maps are already implicitly standardized w.r.t. model variance

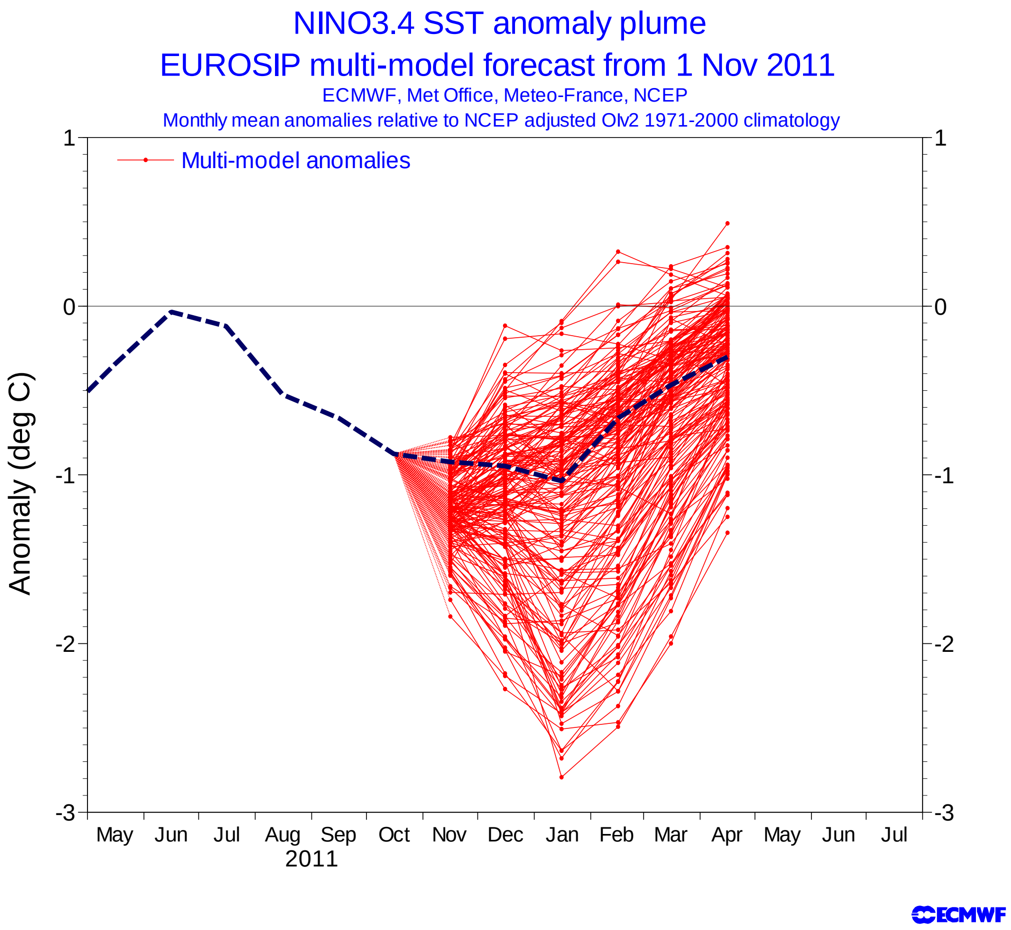
- **General technique**: is also used in our multi-model system

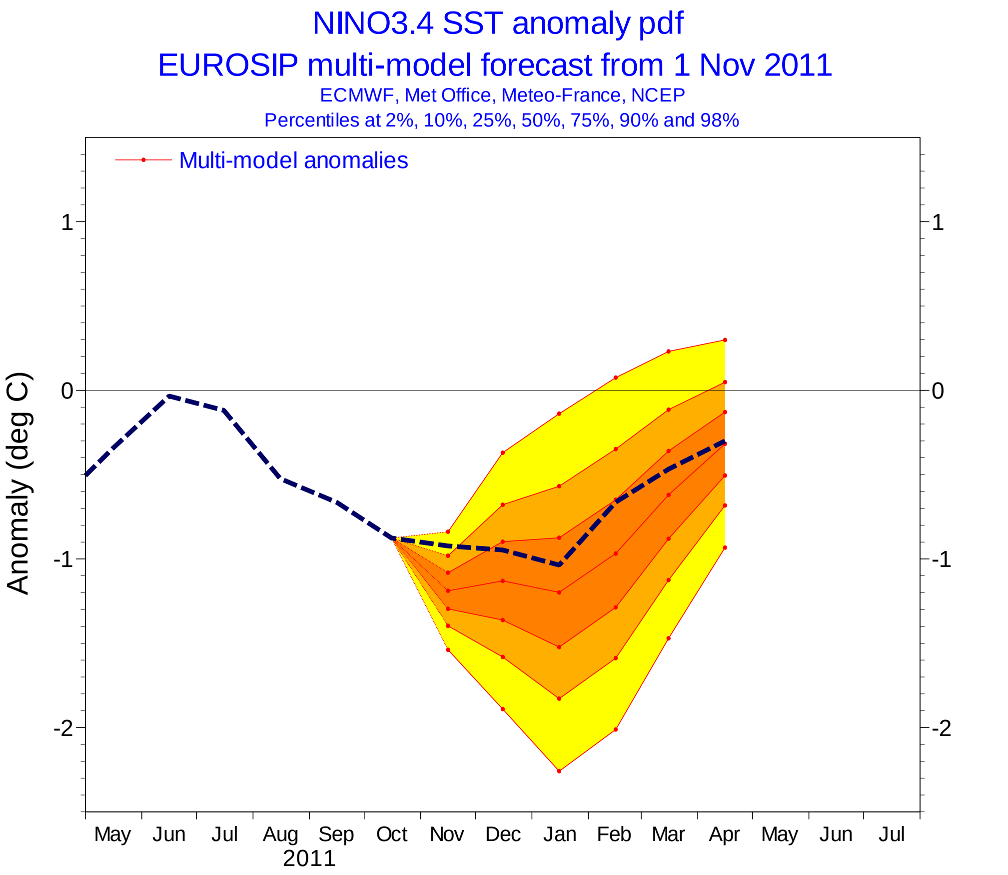
NINO3 SST anomaly plume
ECMWF forecasts from 1 Nov 2007

Monthly mean anomalies relative to NCEP adjusted OIv2 1971-2000 climatology



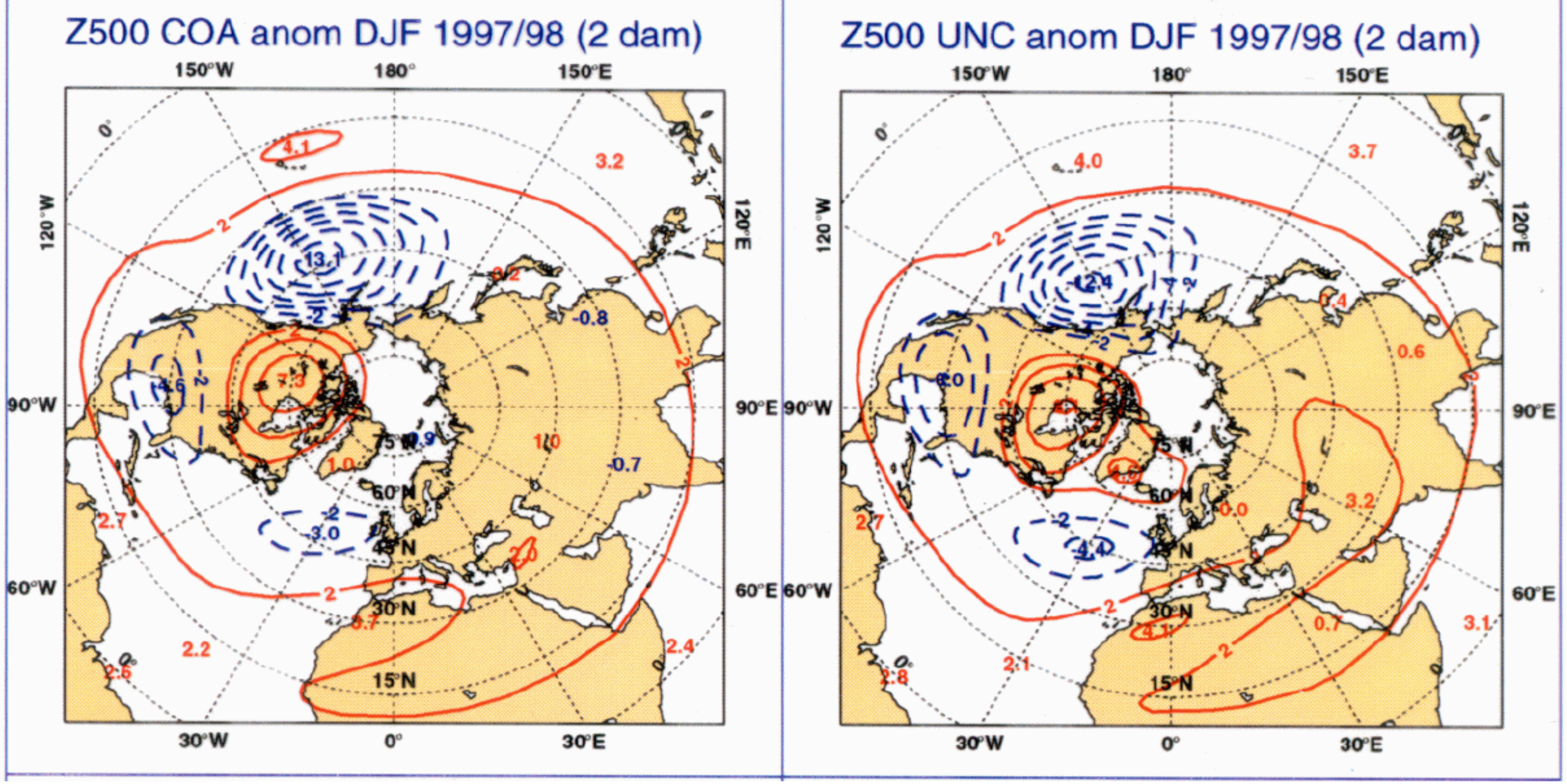
Nino 3.4 plume and pdf – calibrated multi-model forecast



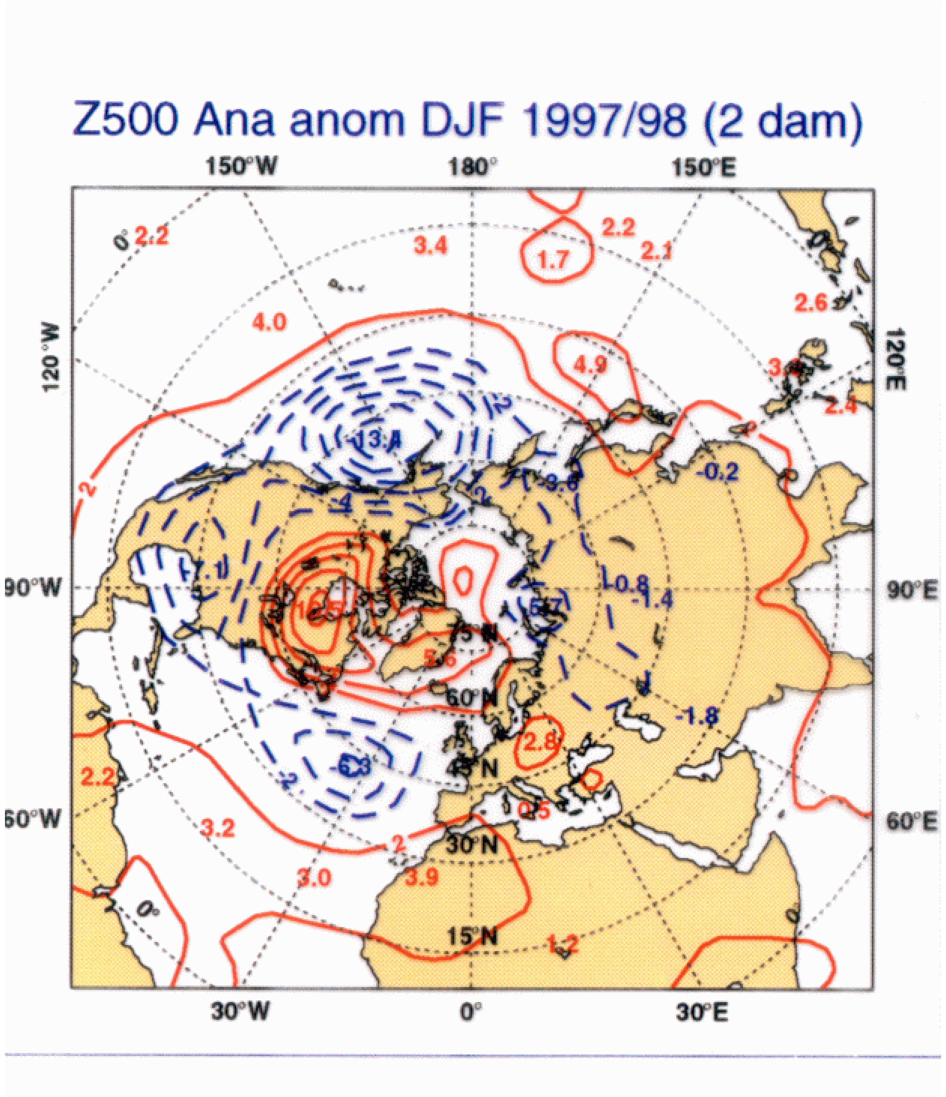


Each model bias and variance corrected, then multi-model ensemble formed, then calibrated to give pdf ("t" distribution)

Despite SST bias and other errors, anomalies in the coupled system can be remarkably similar to those obtained using observed (unbiased) SSTs …..



… and can also verify well against observations



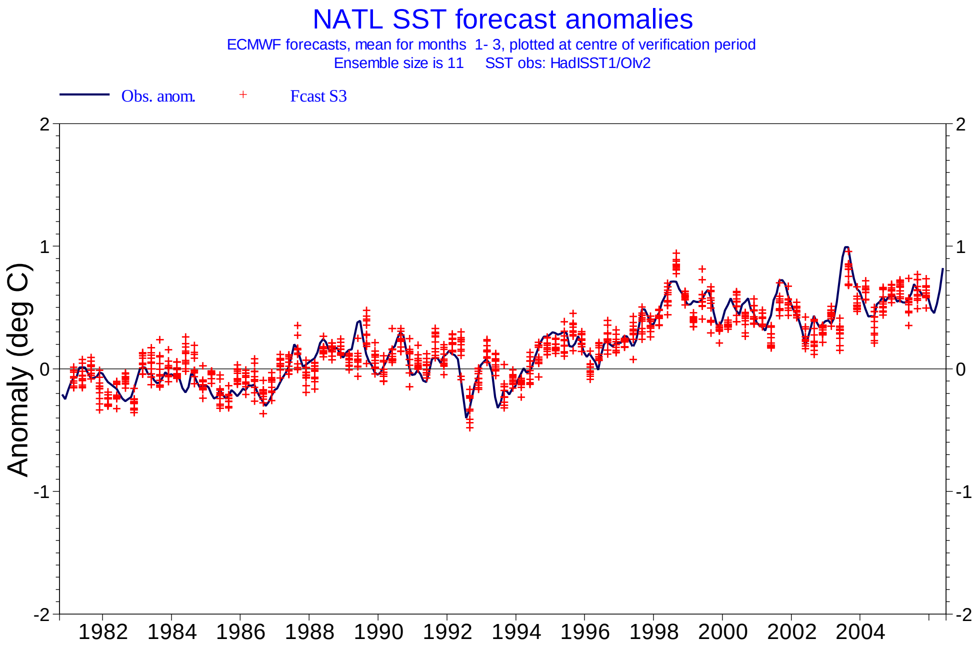
Model errors are still serious …

- Models have errors other than mean bias
  - E.g. System 2 had weak wind and SST variability
  - Past models underestimated MJO activity (S4 better)
  - Suspected too-weak teleconnections to mid-latitudes

- Mean state errors interact with model variability
  - Nino 4 region is very sensitive (cold tongue/warm pool boundary)
  - Atlantic variability suppressed if mean state is badly wrong

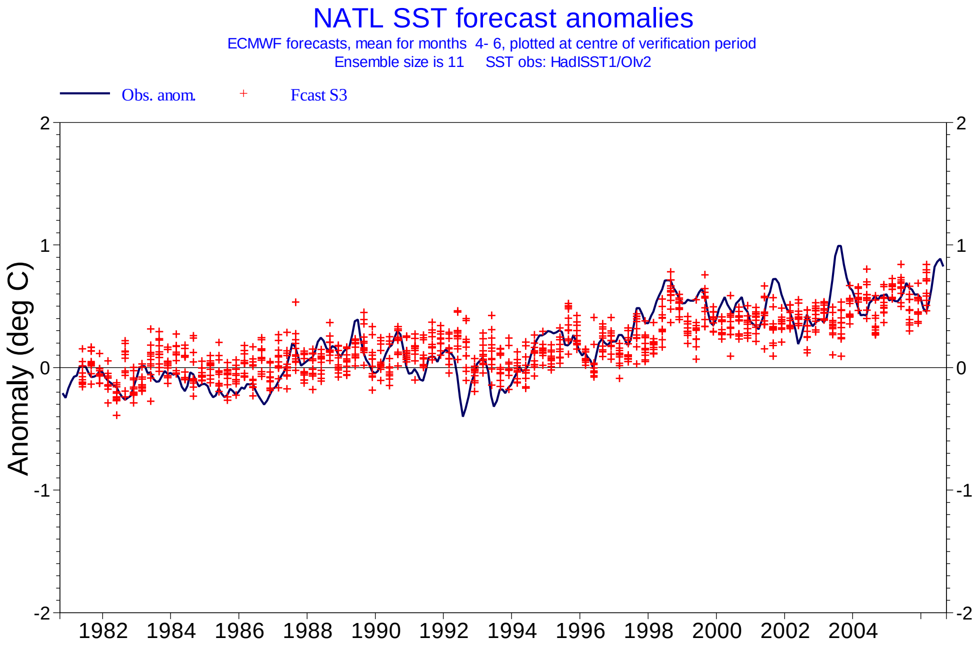
- Forecast errors are often larger than they should be
  - With respect to internal variability estimates and (occasionally) other prediction systems
  - Reliability of probabilistic forecasts is often not particularly high (S4 better)

Trends



Capturing trends is important. Time-varying CO2 and other factors are important in this.

There is a strong link between seasonal prediction and decadal/ multi-decadal climate prediction.



Operational seasonal forecasts

- Real time forecasts since 1997
  - “System 1” initially made public as “experimental” in Dec 1997
  - System 2 started running in August 2001, released in early 2002
  - System 3 started running in Sept 2006, operational in March 2007
  - System 4 started running in July 2011, operational in November 2011
  - SEAS5 re-forecasts have just started
- Burst mode ensemble forecast
  - Initial conditions are valid for 0Z on the 1st of a month
  - Forecasts are usually complete by late on the 2nd.
  - Forecast and product release date is 12Z on the 8th.
- Range of operational products
  - Moderately extensive set of graphical products on web
  - Raw data in MARS
  - Formal dissemination of real time forecast data

System 4 configuration

- Real time forecasts:
  - **51 member ensemble forecast to 7 months**
  - SST and atmos. perturbations added to each member

  - **15 member ensemble forecast to 13 months**
  - Designed to give an ‘outlook’ for ENSO
  - Only once per quarter (Feb, May, Aug and Nov starts)

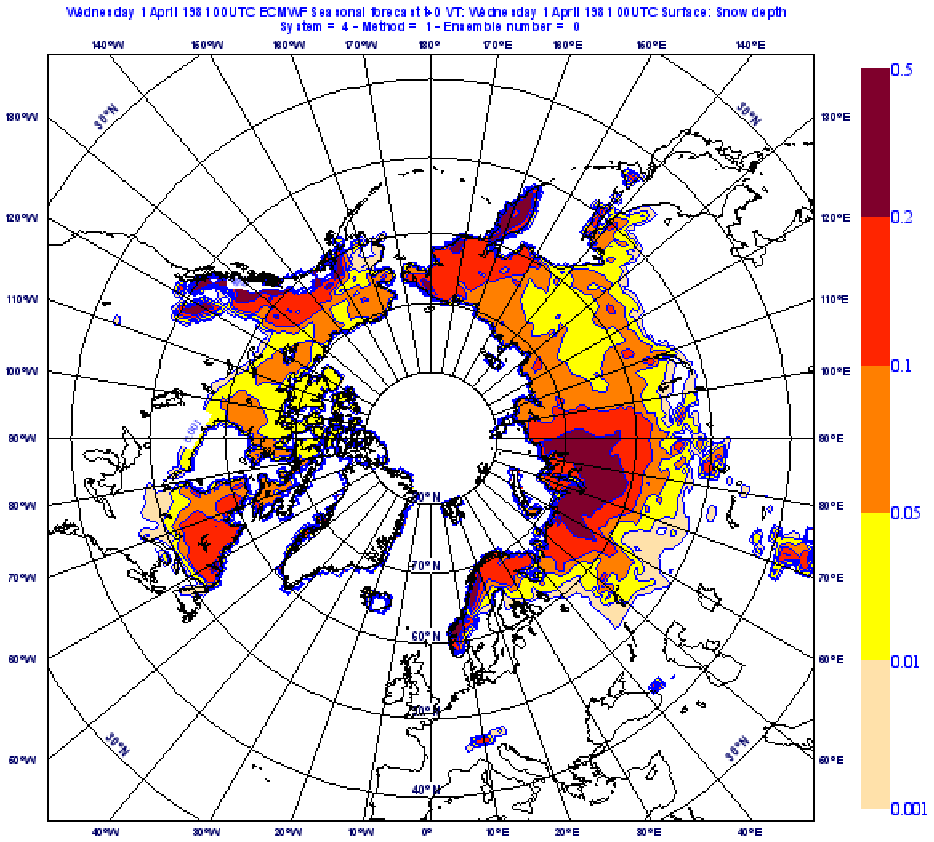
- Re-forecasts from 1981-2010 (30 years)
  - 15 member ensemble every month
  - 15 members extended to 13 months once per quarter

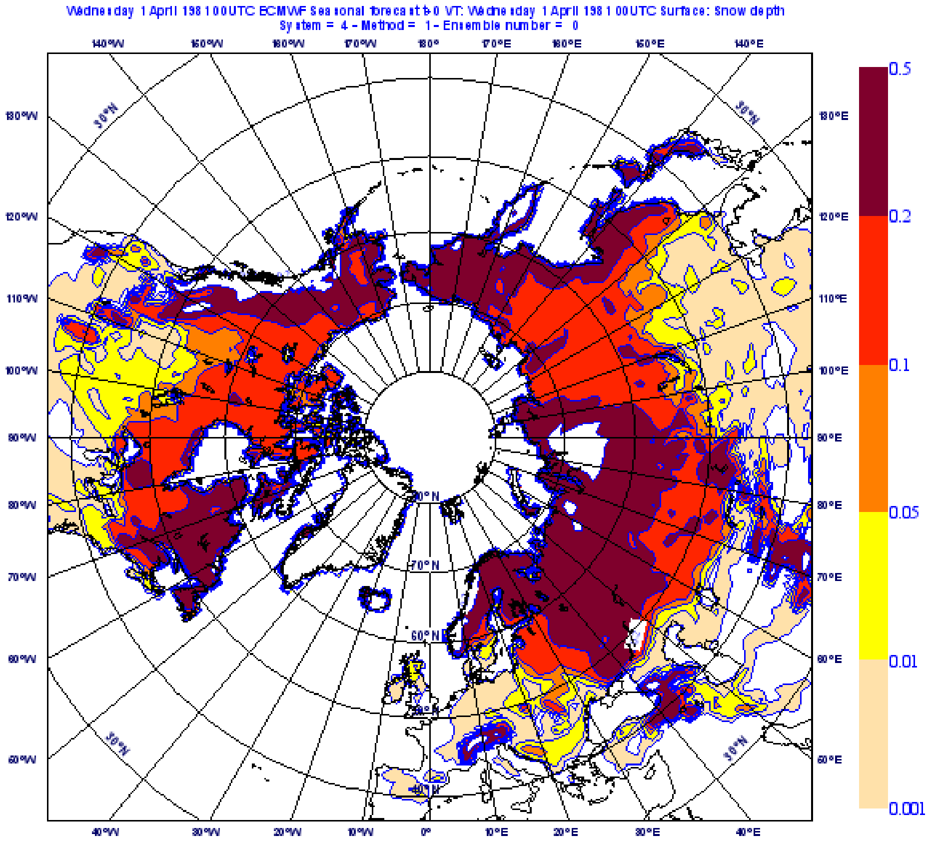
  - Extended to 51 members for Feb, May, Aug and Nov starts

How many re-forecasts?

- Re-forecasts dominate total cost of system
  - System 4: 5400 back integrations (must be in first year)
  - 612 real-time integrations (per year)
- Re-forecasts define model climate
  - Need both climate mean and the pdf, latter needs large sample
  - May prefer to use a “recent” period (30 years? Or less??)
  - System 2 had a 75 member “climate”, S3 had 275, S4 has 450.
  - Sampling is basically OK
- Re-forecasts provide information on skill
  - A forecast cannot be used unless we know (or assume) its level of skill
  - Observations have only 1 member, so large ensembles are less helpful than large numbers of cases.
  - Care needed e.g. to estimate skill of 51 member ensemble based on past performance of 15 member ensemble
  - For regions of high signal/noise, System 4 gives adequate skill estimates
  - For regions of low signal/noise (eg <= 0.5), need hundreds of years

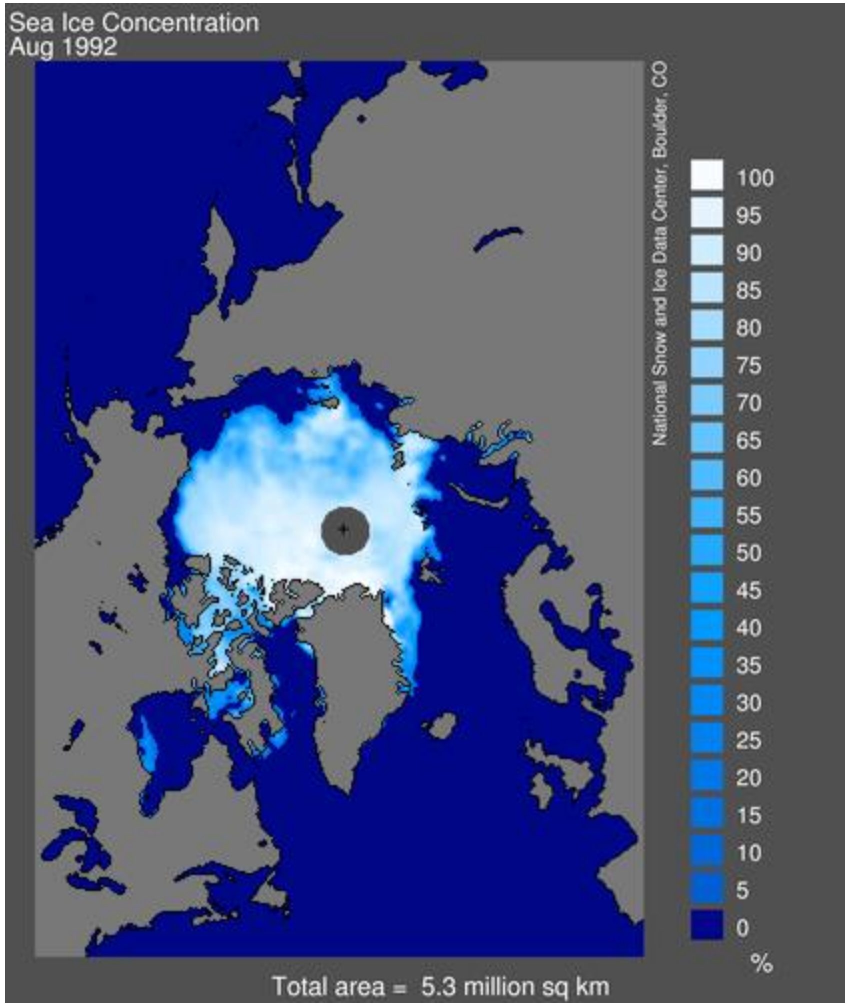
Land surface

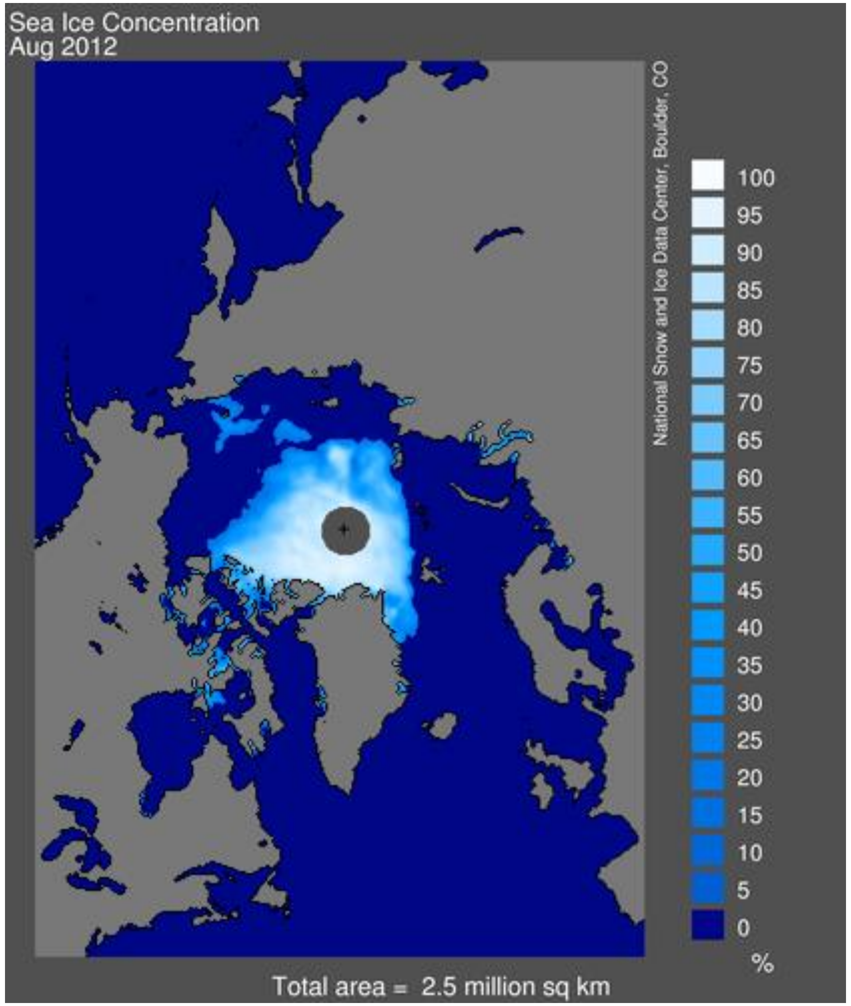




Snow depth limits, 1st April

Sea ice

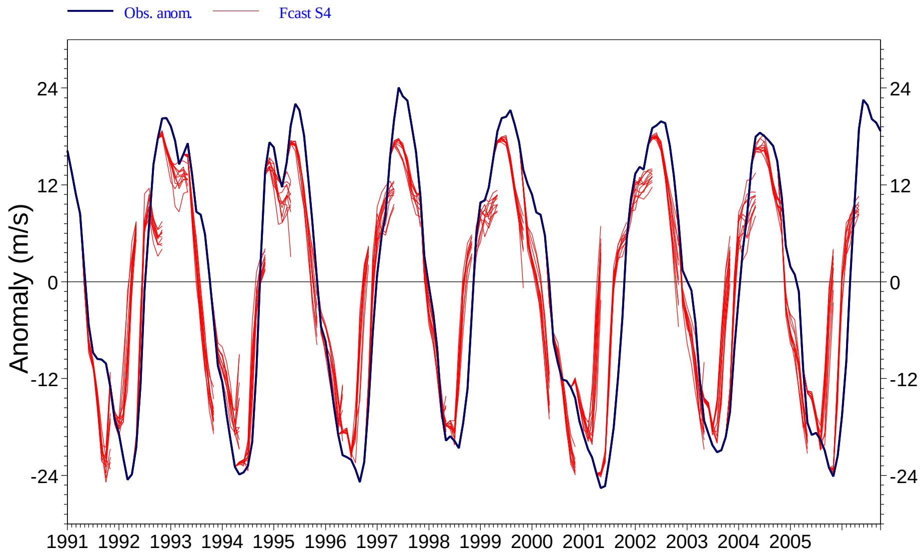




S4 does not have a physical ice model, but each ensemble member uses ice cover from one of the previous 5 years. This captures the downwards trend in arctic ice, and also samples uncertainty in the ice cover to be expected during the forecast.

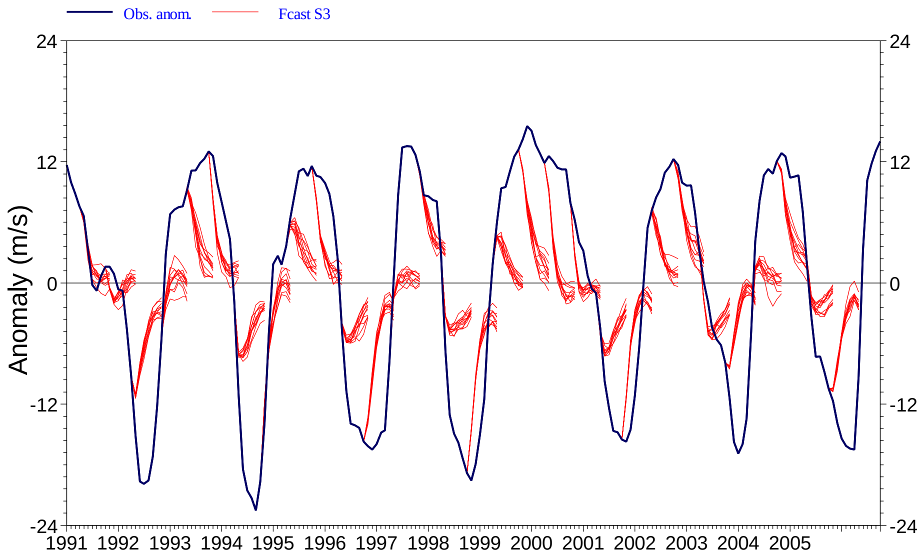
QBO

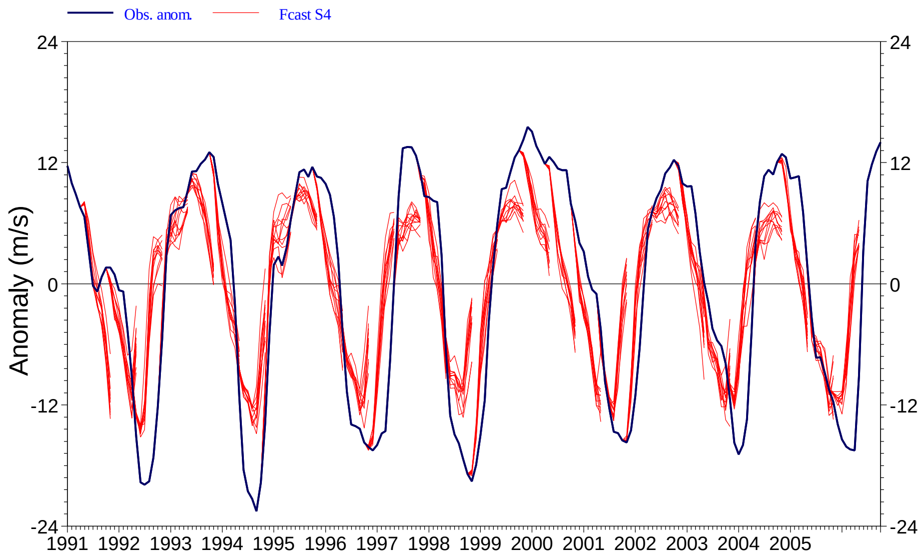
System 4



30hPa

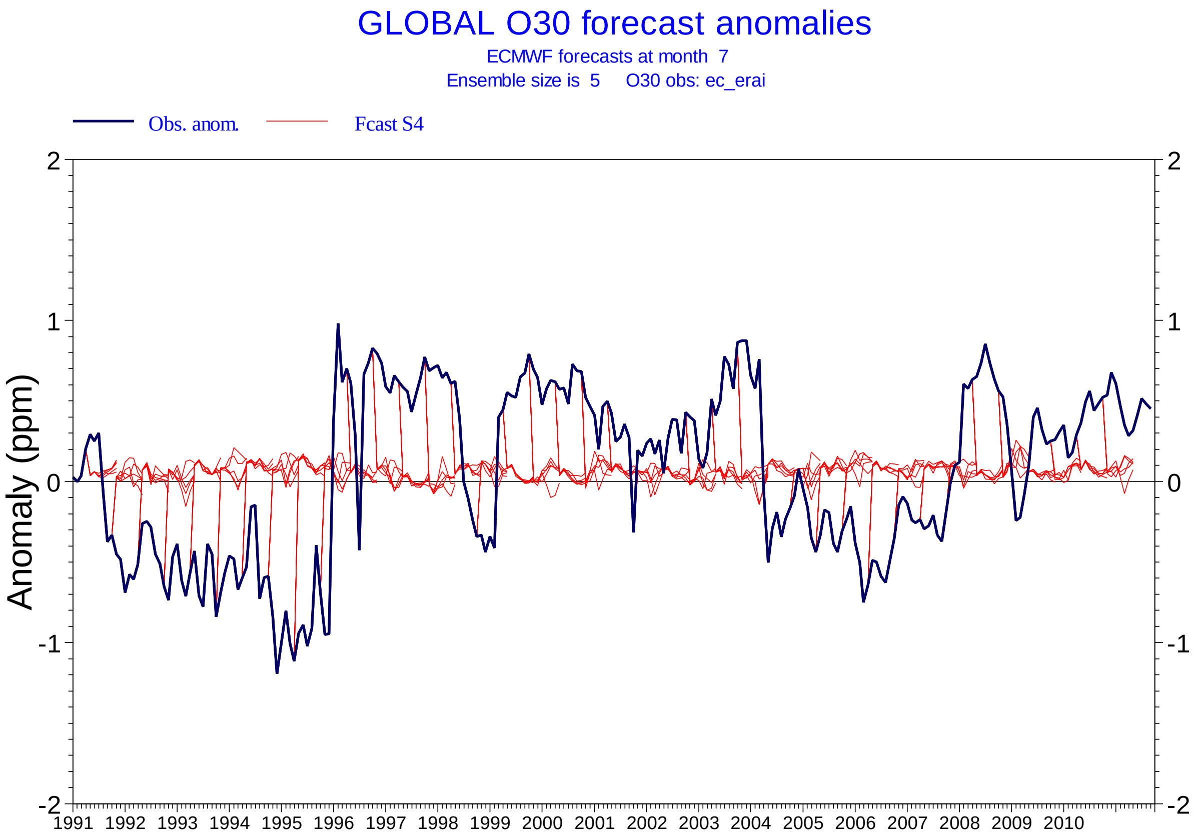
System 3





50hPa

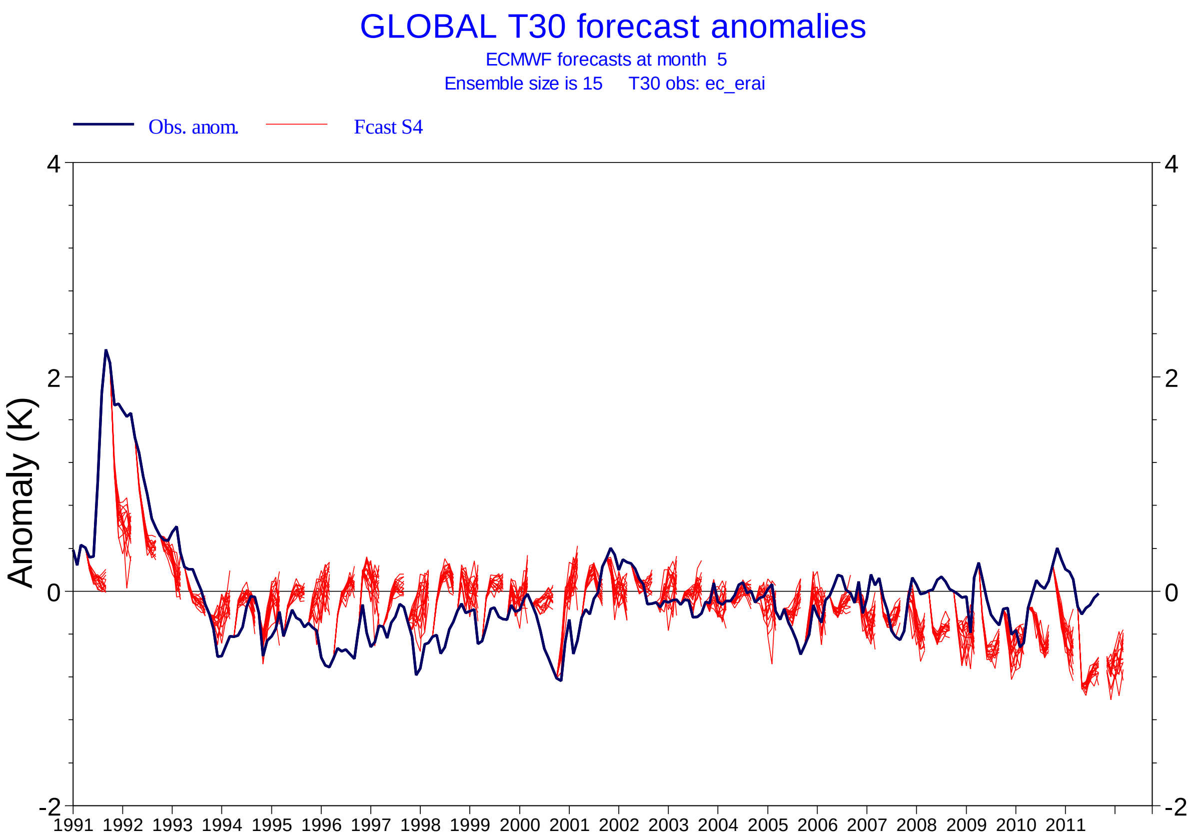
Ozone



S4 uses interactive ozone, which is able to improve temperature forecasts in the stratosphere.

Ozone re-analyses are dominated by spurious changes, and cannot be used to initialize forecasts. For S4, we were forced to use a climatological initial condition instead.

Stratospheric trends - problems



Spurious cooling in recent re-forecasts/forecasts. This is due to an erroneous trend in initial conditions of stratospheric water vapour, which in turn is due to changes in the observing system.

This affects both ERA interim and operational analyses.

(Independent of the trend issue, lower stratosphere humidity is problematic in both ECMWF analyses and forecasts)

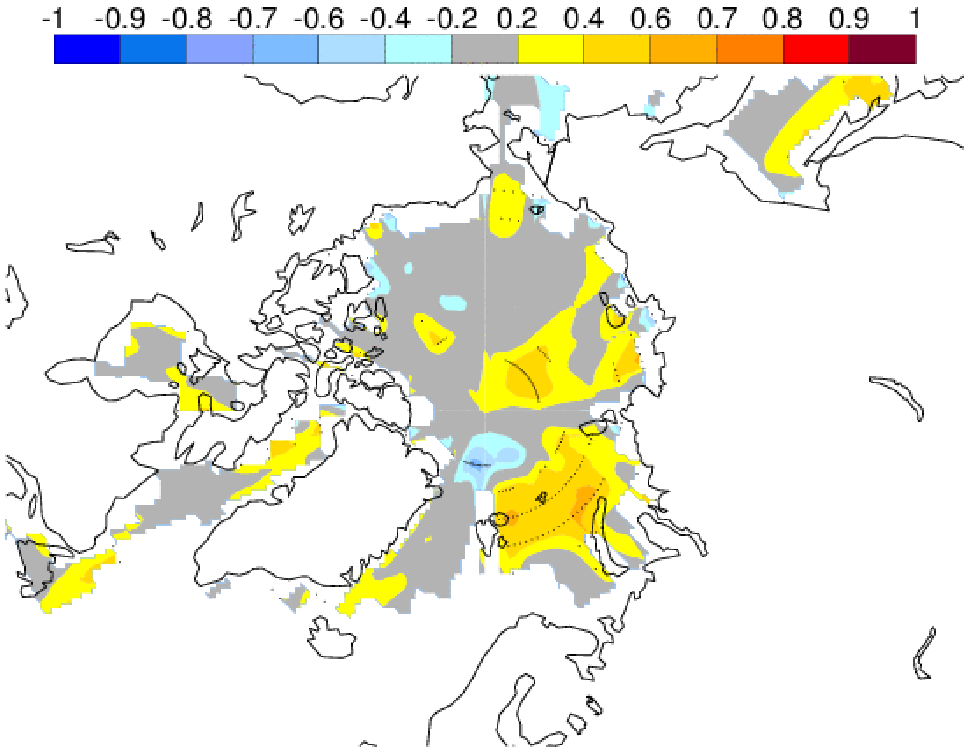
SEAS5

- New high resolution system, SEAS5, will be introduced later this year (expected October)
- Re-forecasts have just started

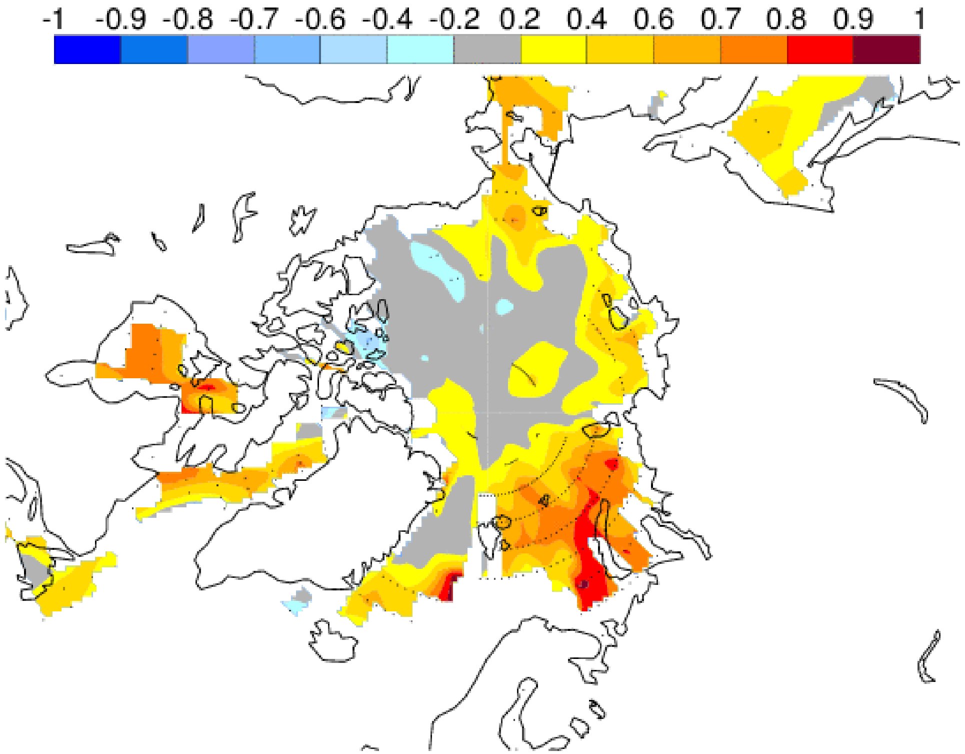
	System 4 (S4)	SEAS5 (S5)
IFS Cycle	36r4	43r1
Resolution (grid spacing)	TL255 (80 km)	TCo319 (35 km)
Atmosphere levels	L91	L91
Ocean resolution	ORCA1 (1°)	ORCA025 (0.25°)
Ocean levels	L42	L75
Sea-ice	Prescribed (last 5 years)	LIM2
Re-forecast years	1981-2010 (30y)	1981-2016 (36y)
Re-forecast ensemble size	15	25

Sea ice cover - DJF anomaly correlations

SEAS4



SEAS5

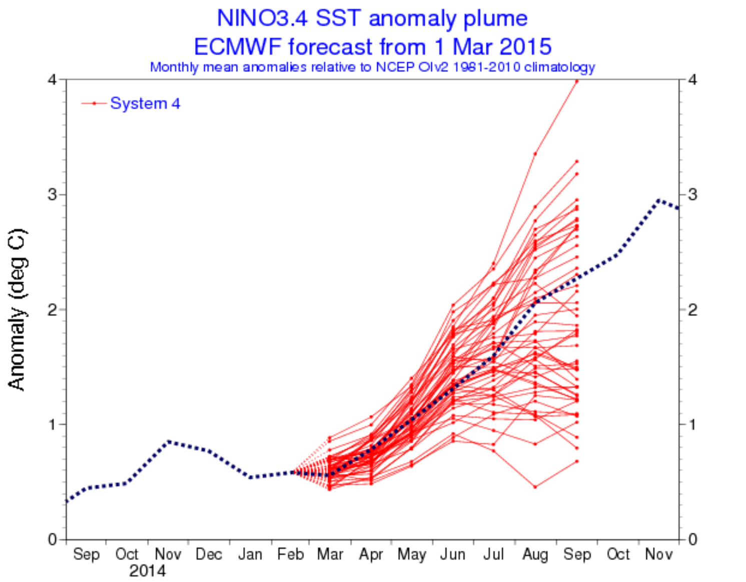


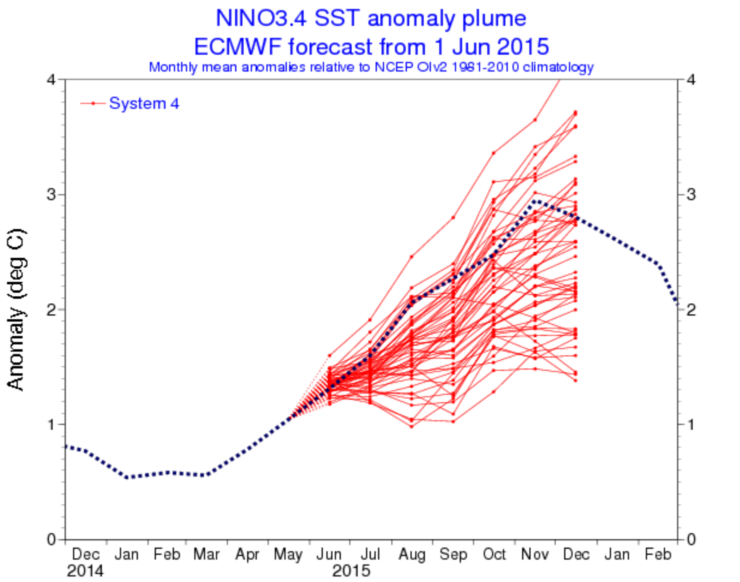
Sea ice cover predictability is improved when we include the interactive sea ice model

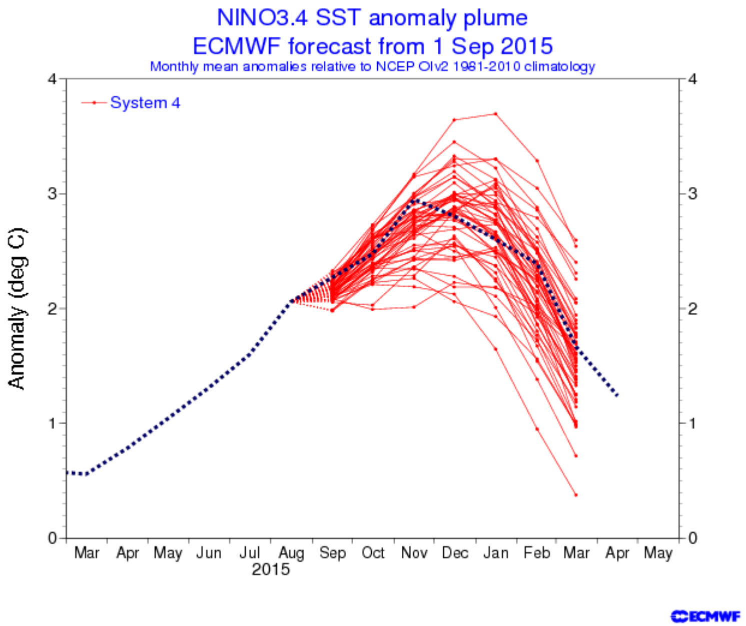
Example forecast products

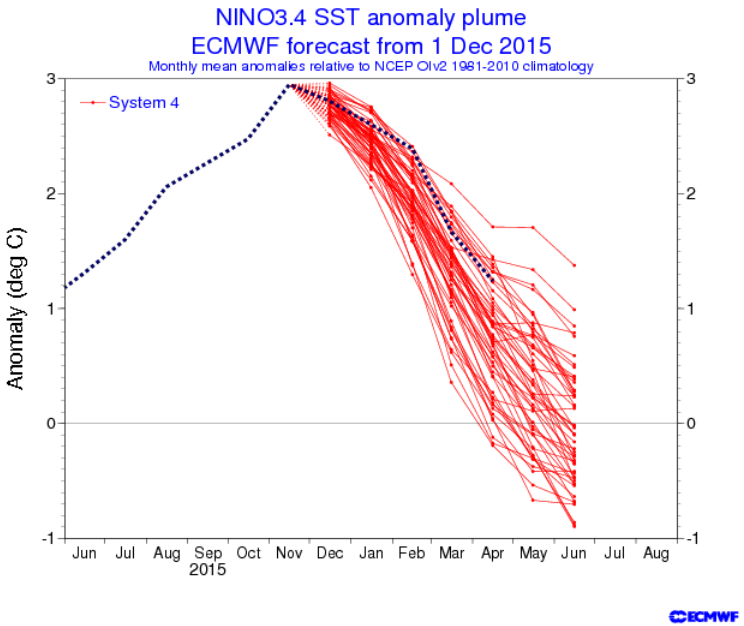
- A few examples only – see web pages for full details and assessment of skill
- All graphical products come with corresponding skill estimate

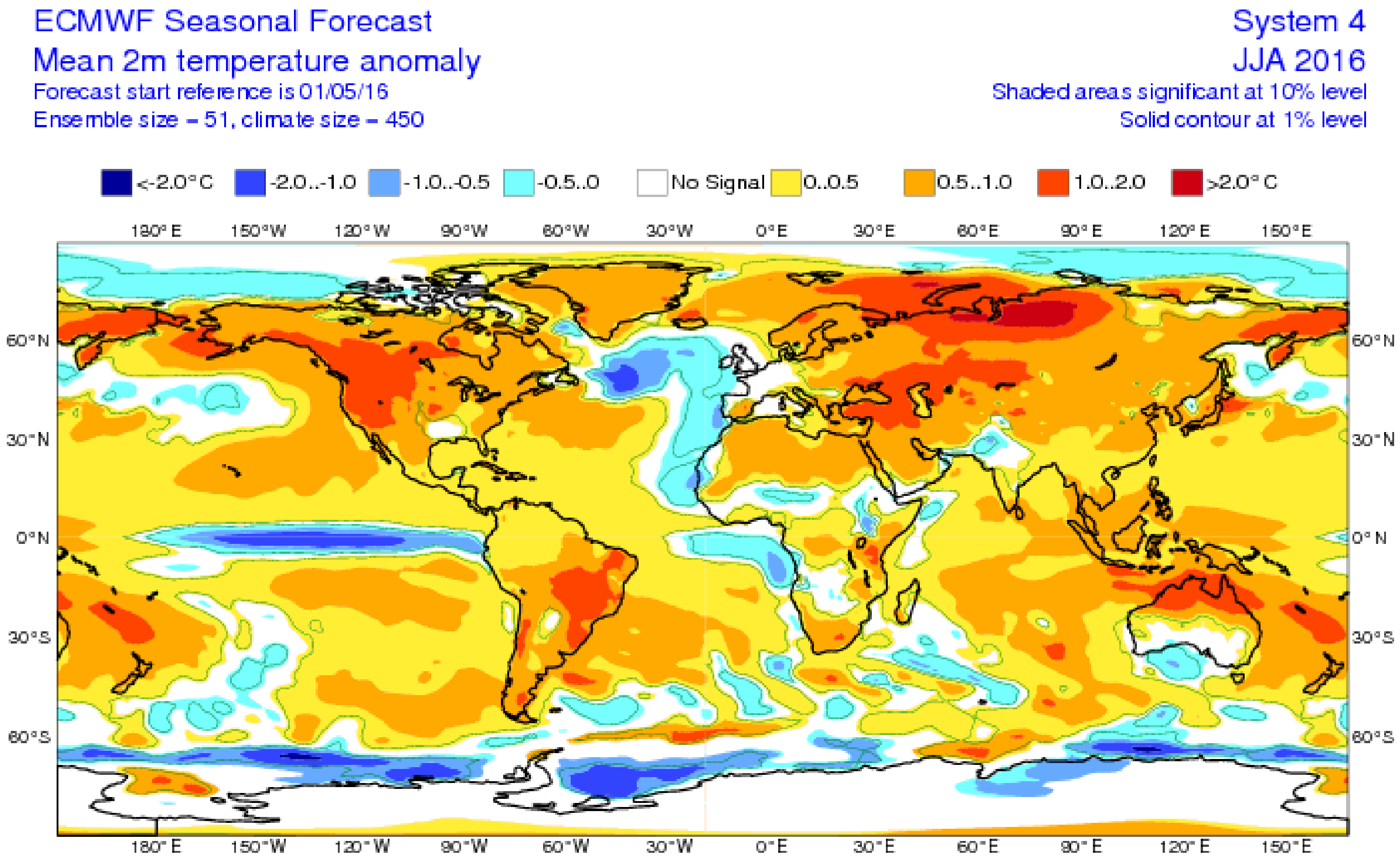
- Note: Significance values on plots
  - A lot of variability in seasonal mean values is due to chaos
  - Ensembles are large enough to test whether any apparent signals are real shifts in the model pdf
  - We use the w-test, non-parametric, based on the rank distribution
  - NOT related to past levels of skill

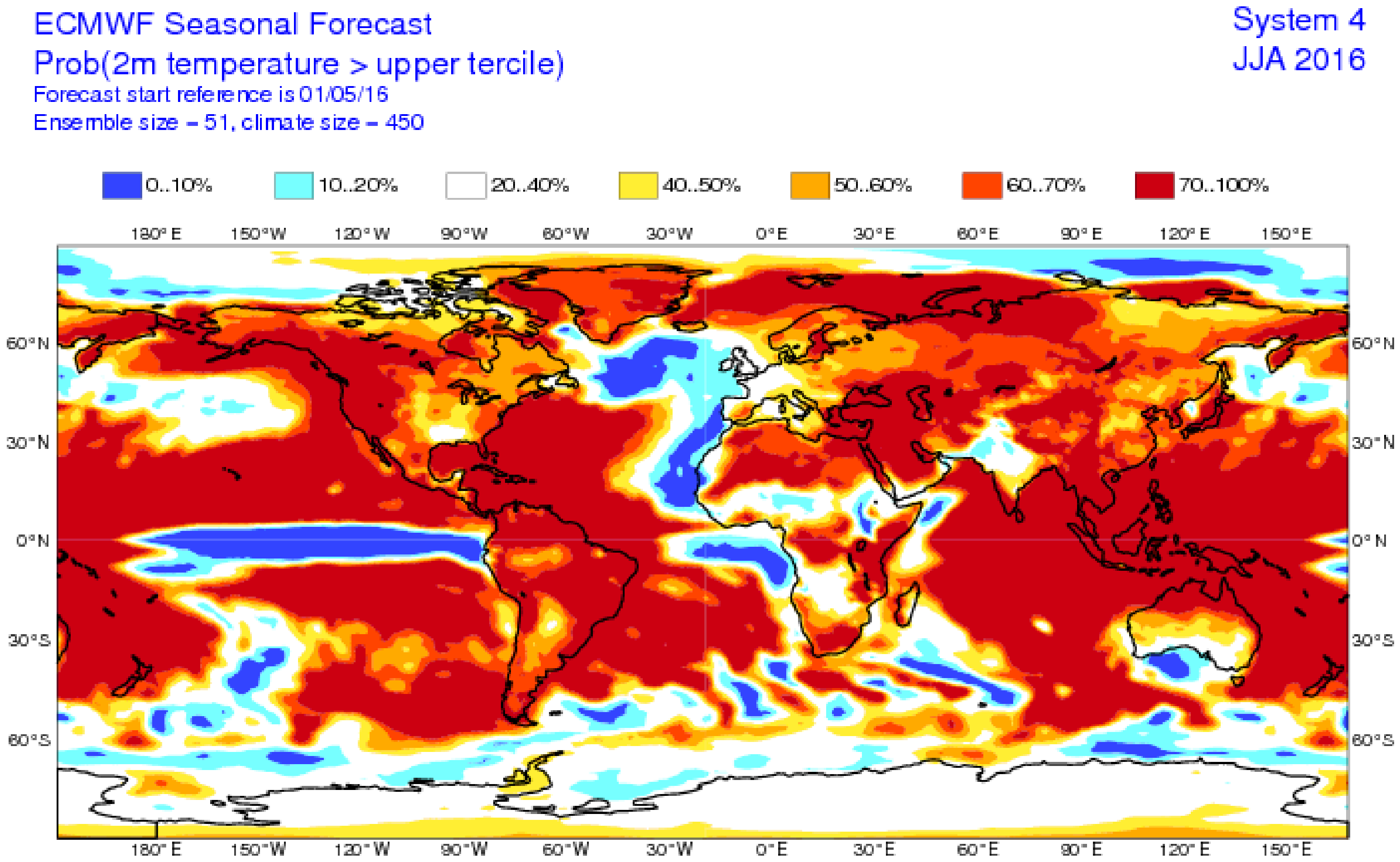

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Mar 2015
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology
System 4
Anomaly (deg C)
Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov
2014


NINO3.4 SST anomaly plume
ECMWF forecast from 1 Jun 2015
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology
System 4
Anomaly (deg C)
Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb
2014 2015

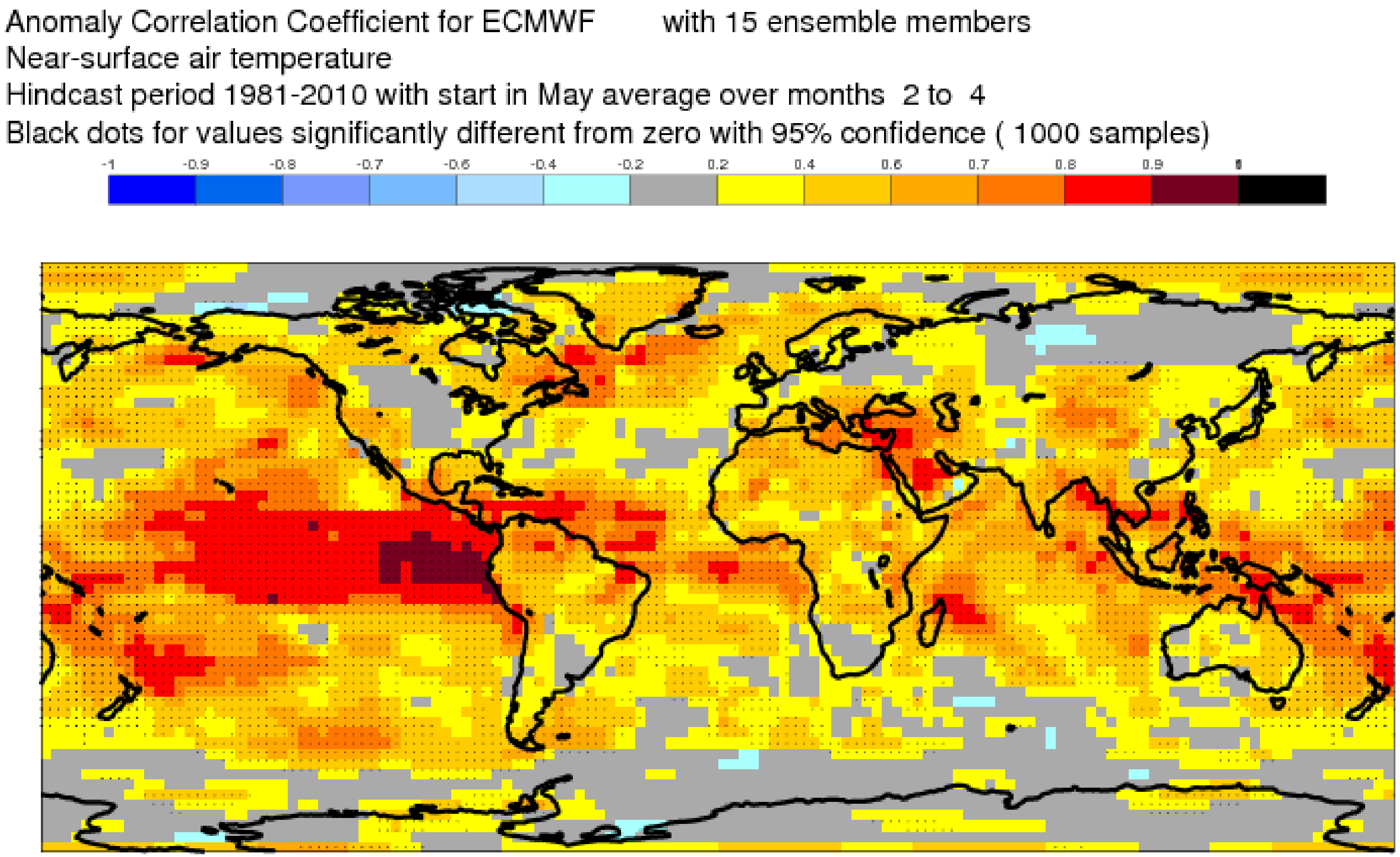

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Sep 2015
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology
System 4
Anomaly (deg C)
Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May
2015
ECMWF


NINO3.4 SST anomaly plume
ECMWF forecast from 1 Dec 2015
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology
System 4
Anomaly (deg C)
Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug
2015
ECMWF

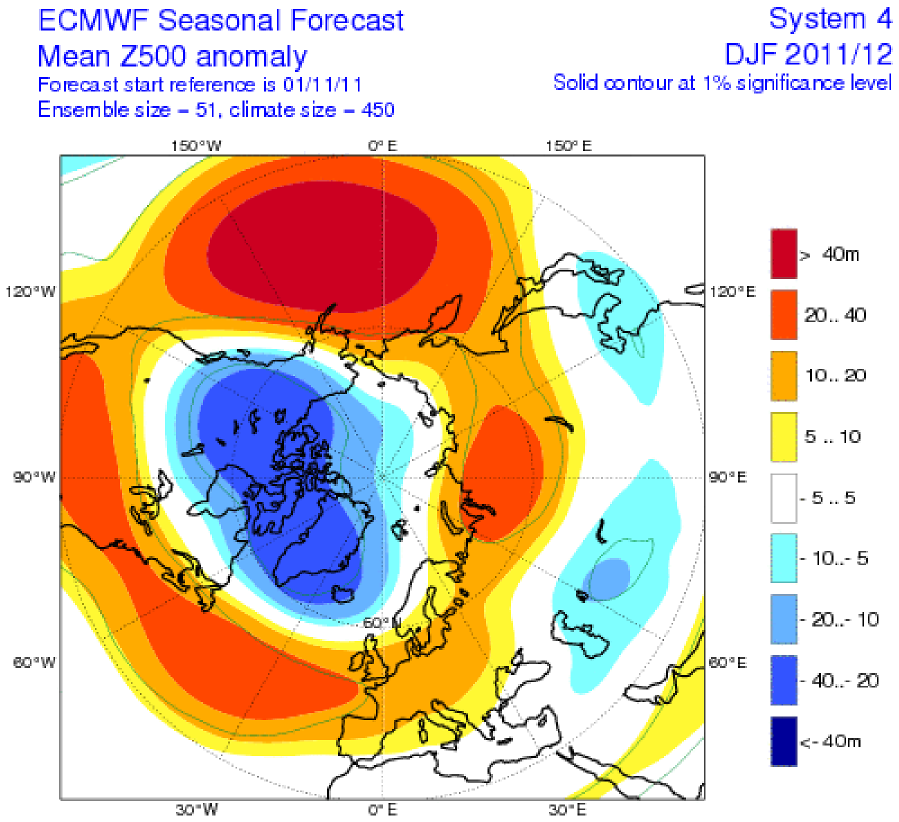

ECMWF Seasonal Forecast
Mean 2m temperature anomaly
Forecast start reference is 01/05/16
Ensemble size = 51, climate size = 450
System 4
JJA 2016
Shaded areas significant at 10% level
Solid contour at 1% level
<-2.0°C
-2.0..-1.0
-1.0..-0.5
-0.5..0
No Signal
0..0.5
0.5..1.0
1.0..2.0
>2.0°C
180° E
150°W
120°W
90°W
60°W
30°W
0°E
30°E
60°E
90°E
120°E
150°E
60°N
30°N
0°N
30°S
60°S

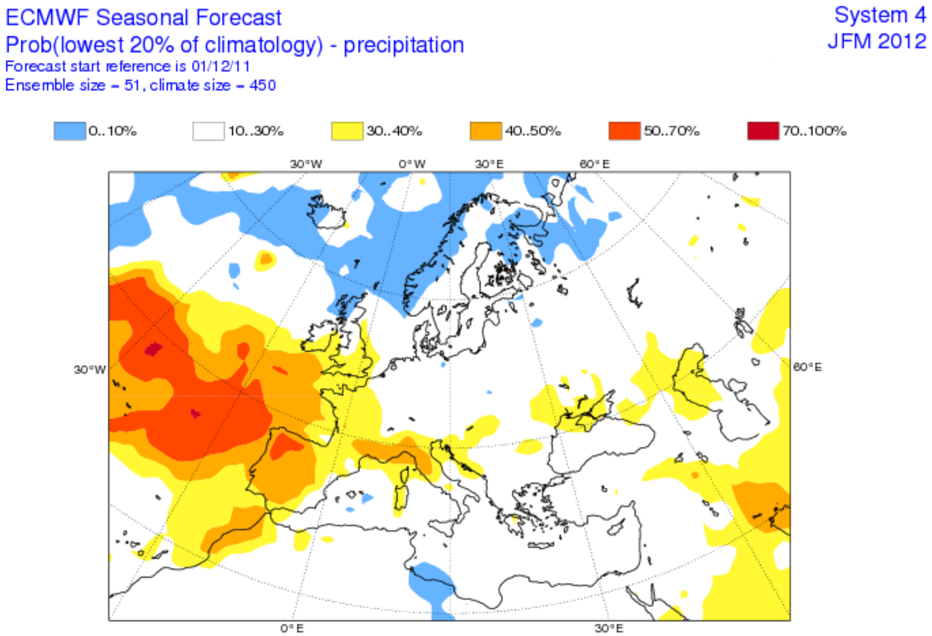

ECMWF Seasonal Forecast
Prob(2m temperature > upper tercile)
Forecast start reference is 01/05/16
Ensemble size = 51, climate size = 450
System 4
JJA 2016
0..10%
10..20%
20..40%
40..50%
50..60%
60..70%
70..100%
180° E
150°W
120°W
90°W
60°W
30°W
0°E
30°E
60°E
90°E
120°E
150°E
60°N
30°N
0°N
30°S
60°S

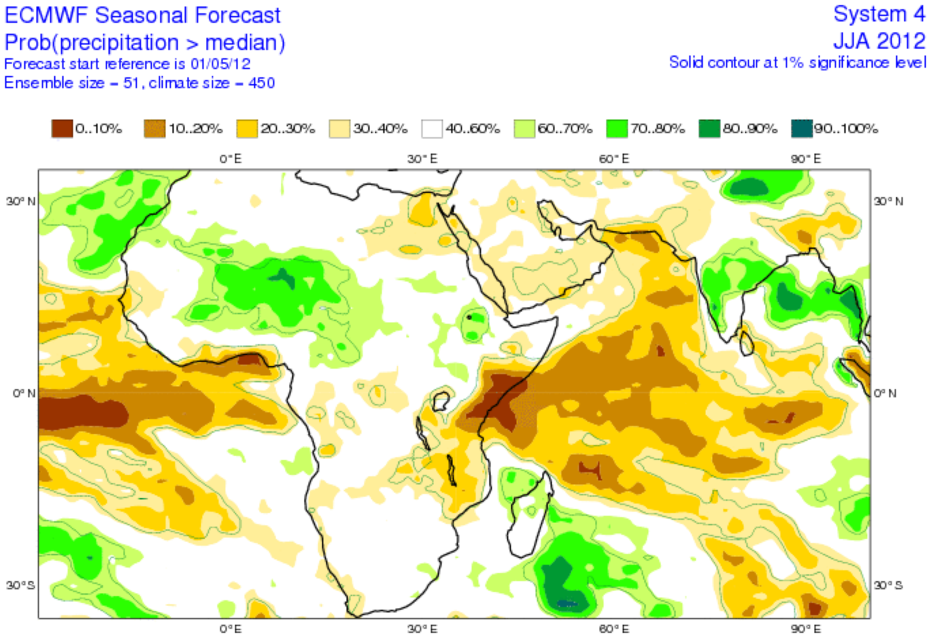
Measure of past skill (ACC)



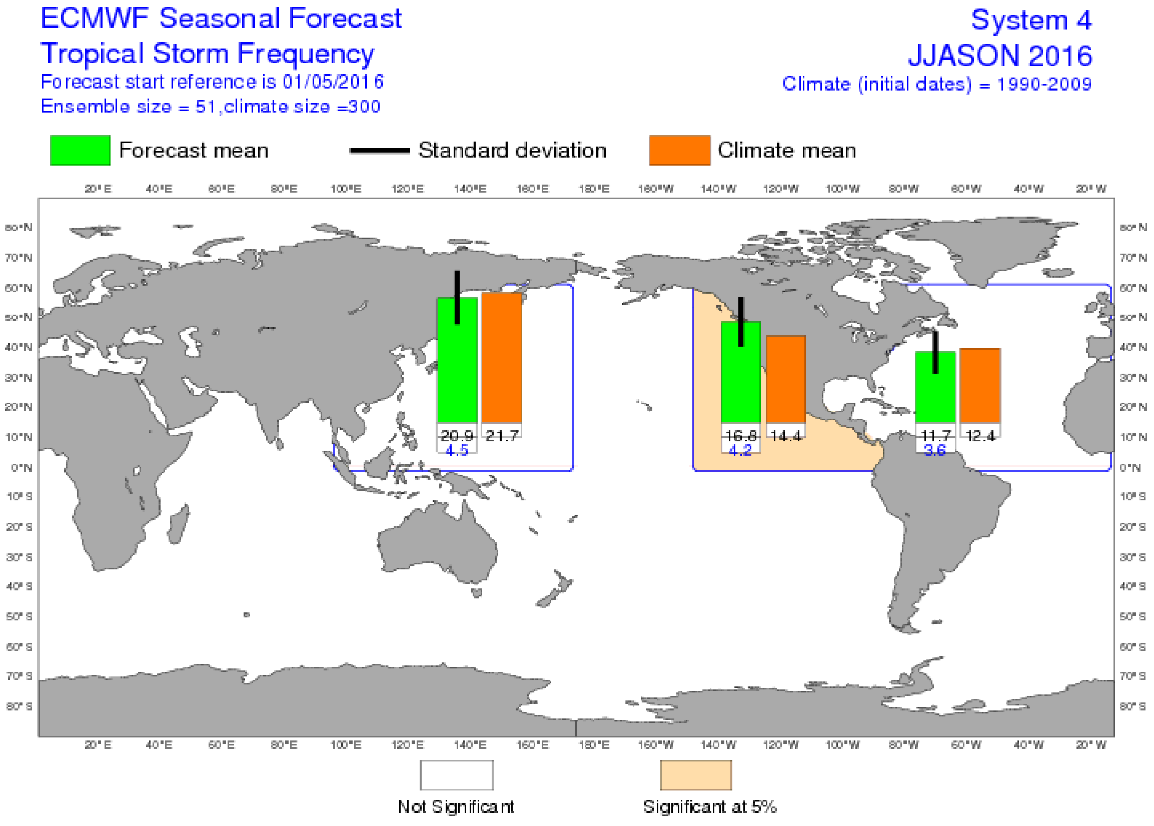
Other operational plot examples

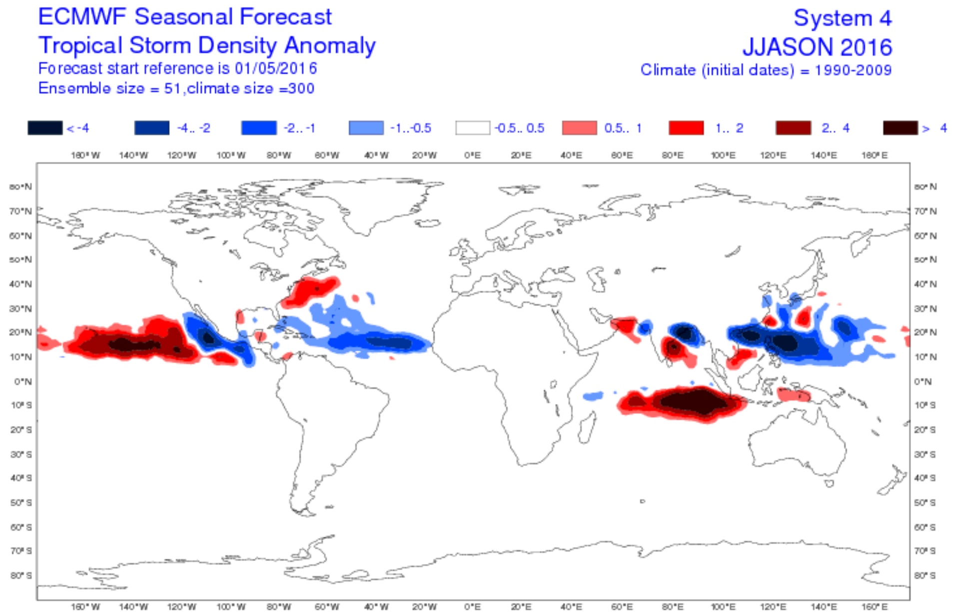


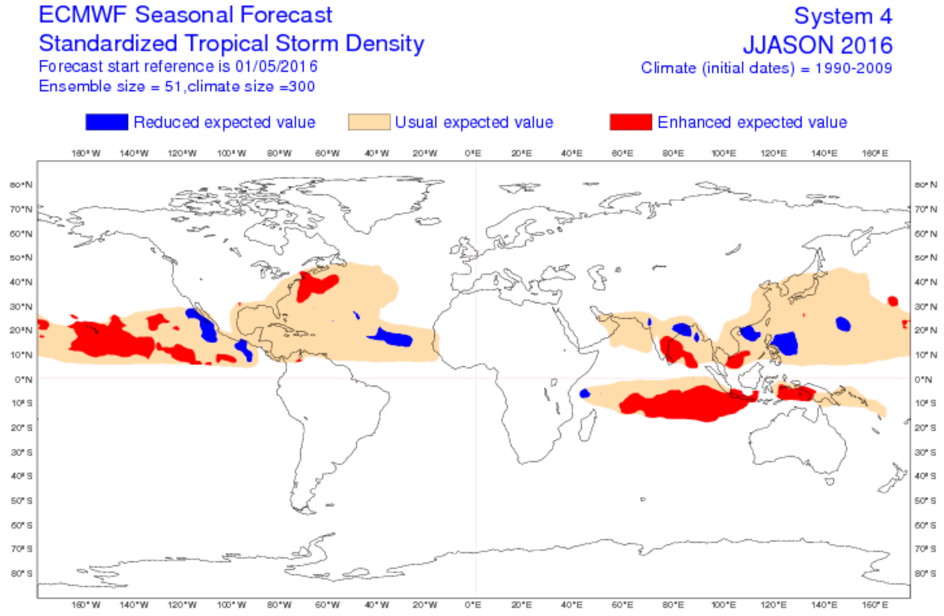




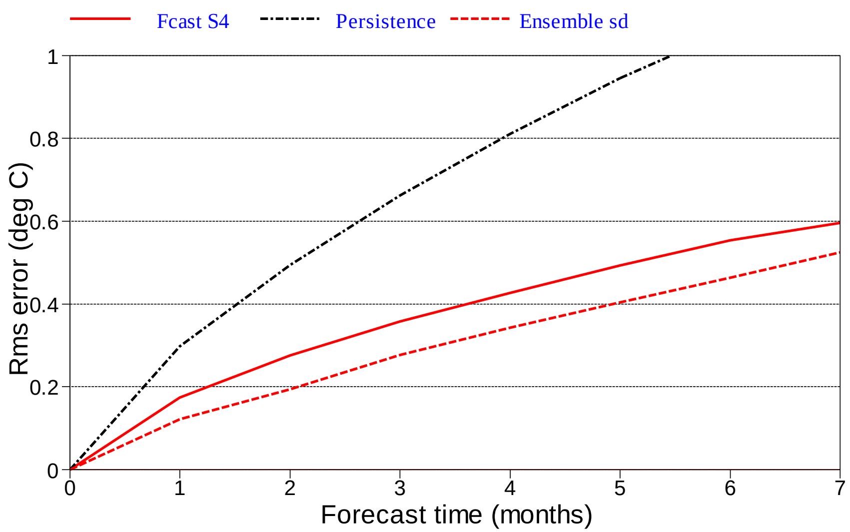
Tropical storm forecasts



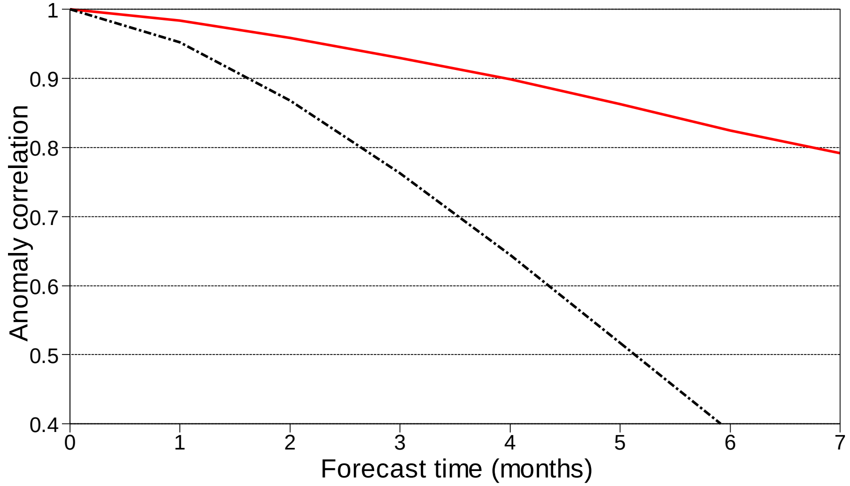




SST forecast performance

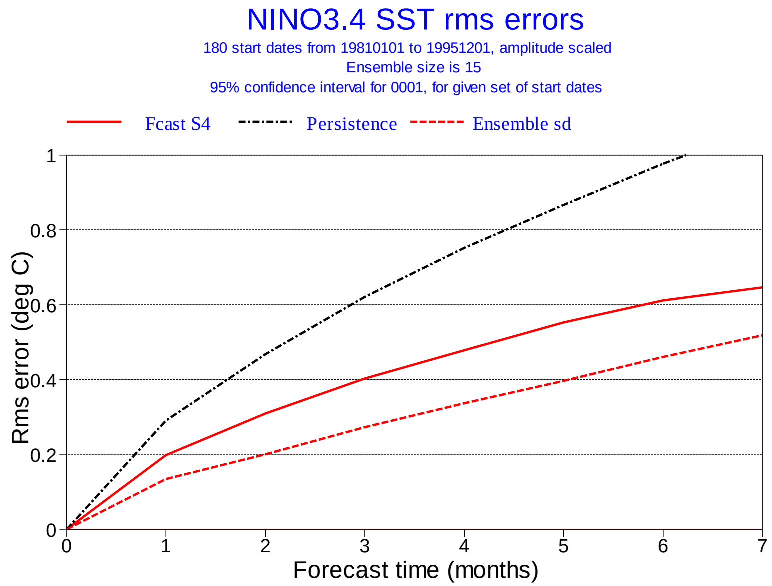


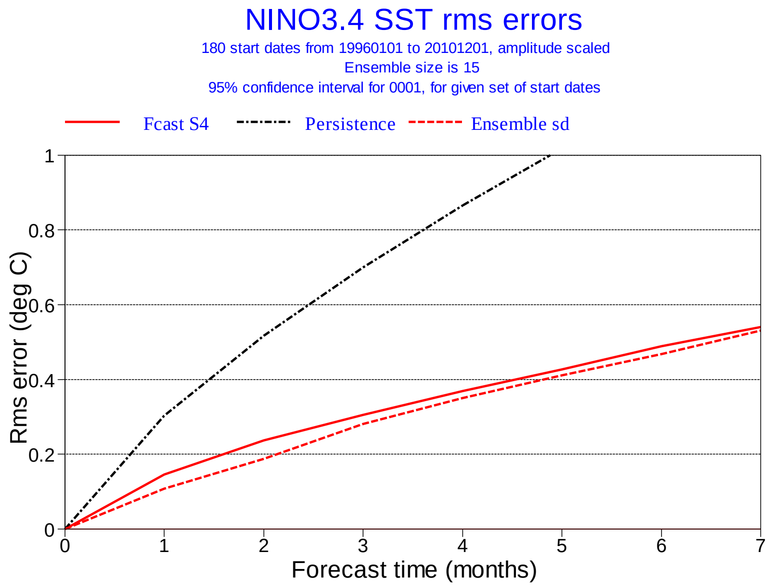
Actual rms errors > model estimate of "perfect model" errors



NOTE: In System 4, stochastic physics gives substantially increased spread to Nino SSTs, due to representation of low-frequency model error. This gives better probabilistic scores, but means the ensemble spread is not a predictability limit: in future systems, we can reduce the amplitude of the stochastic "noise" as model errors are reduced.

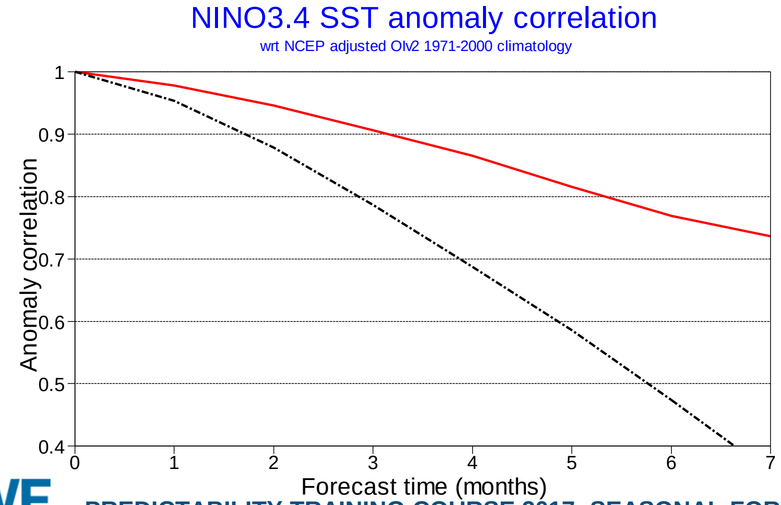
More recent SST forecasts are better

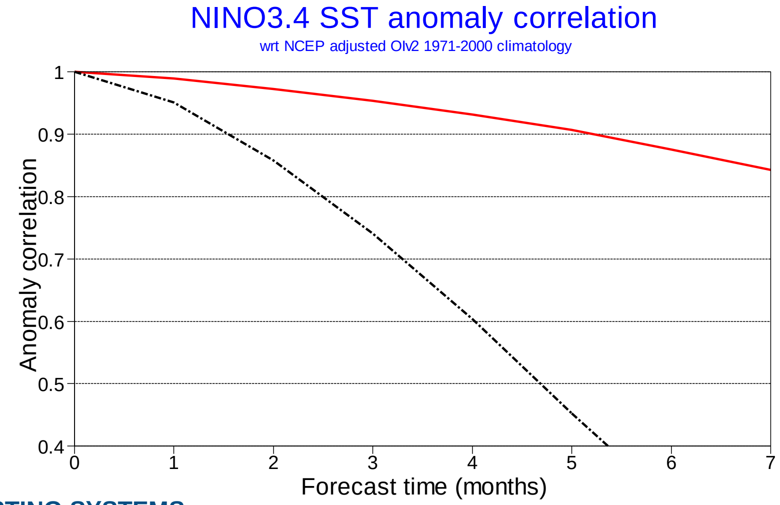




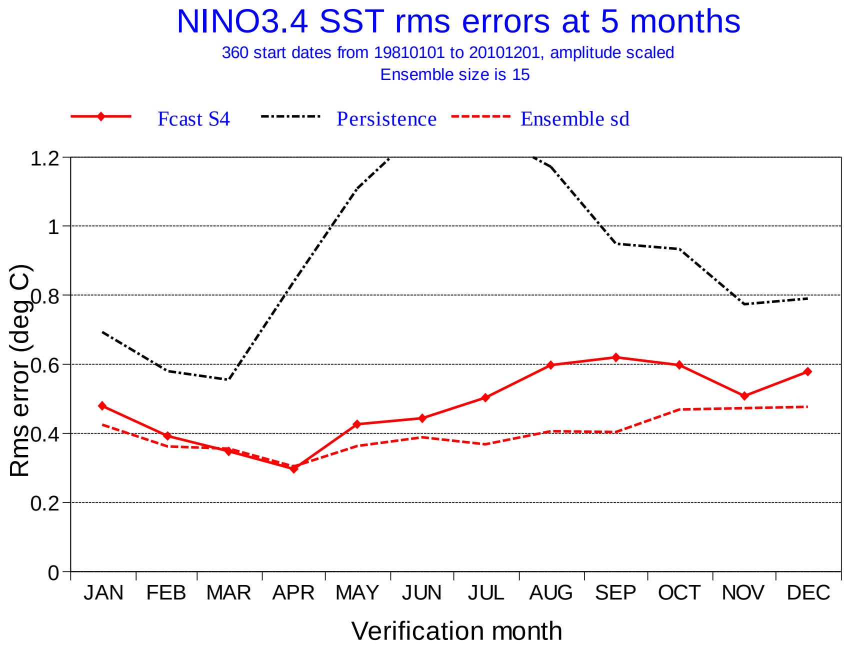
1981-1995

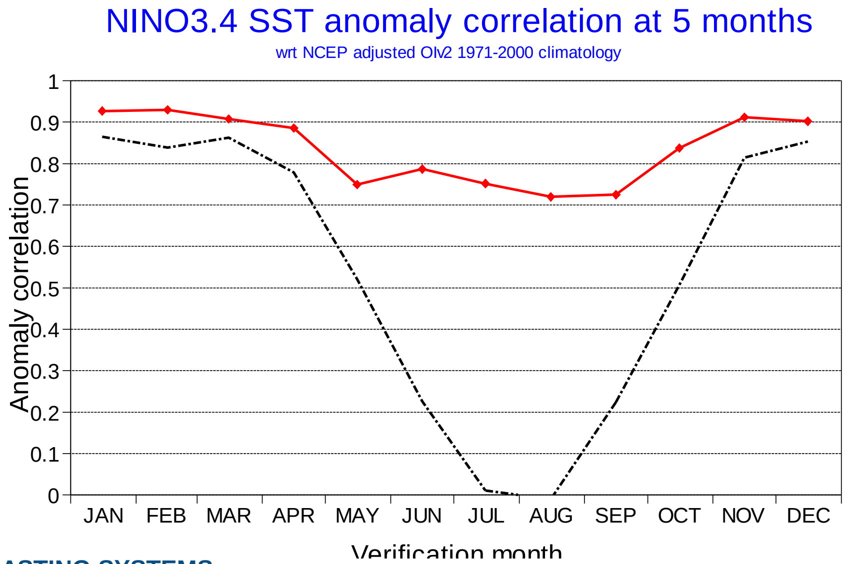
1996-2010



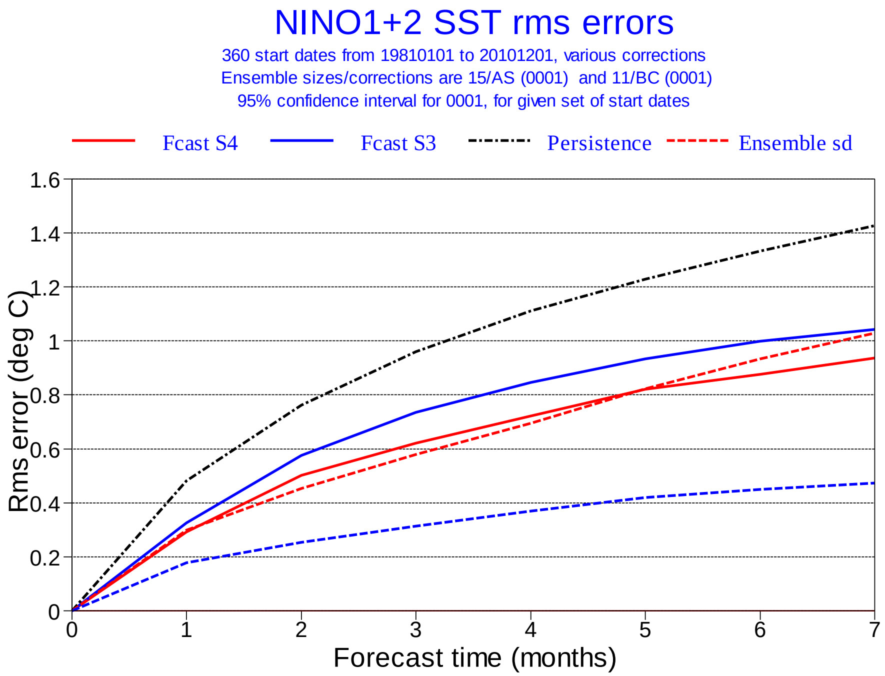


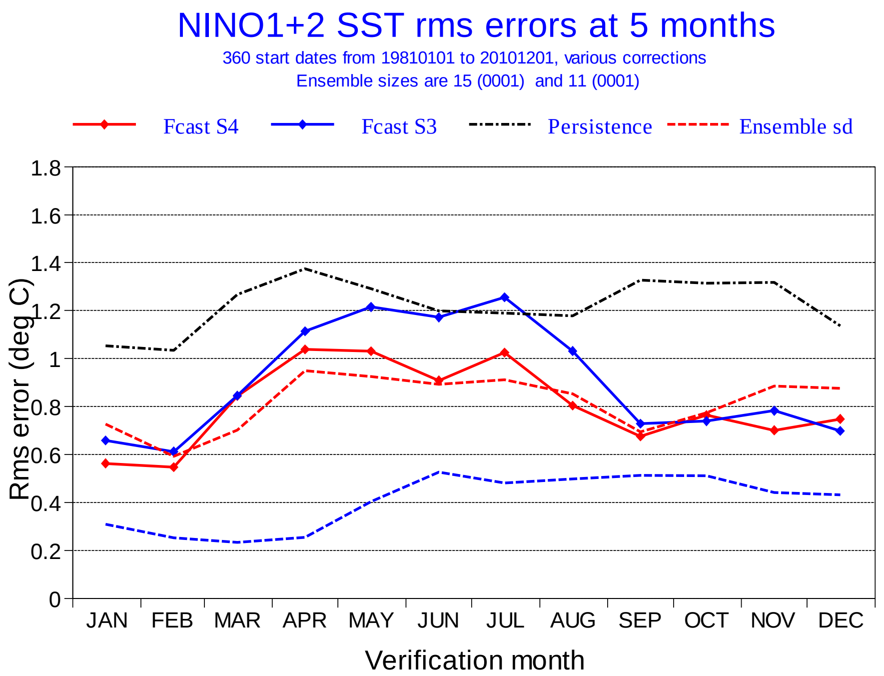
Seasonal dependence

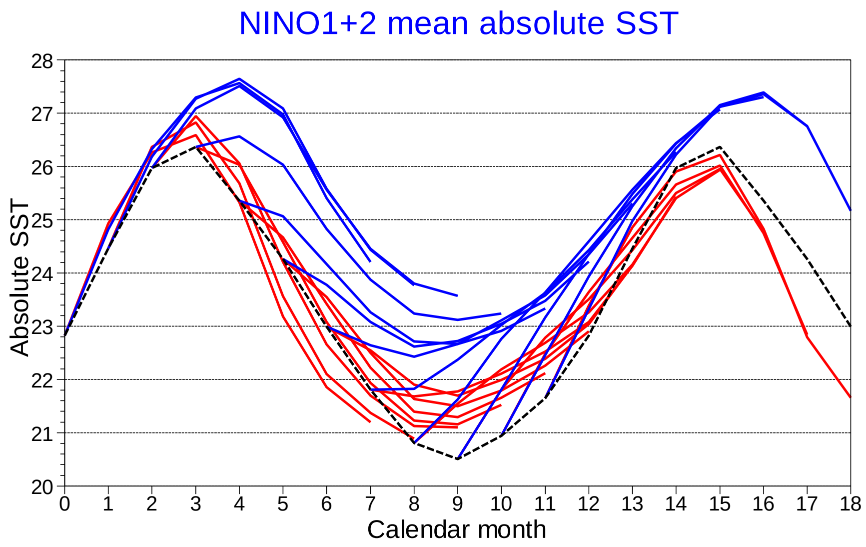




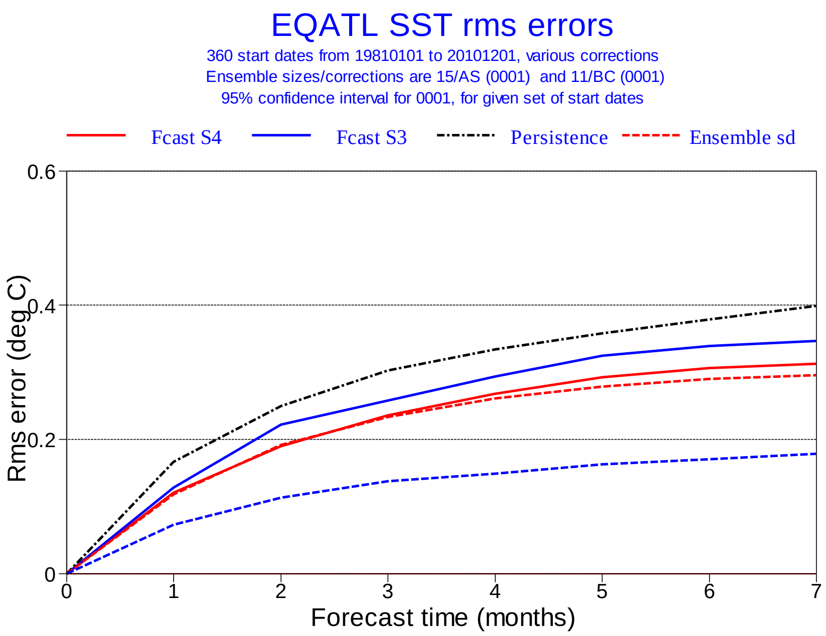
Nino 1+2

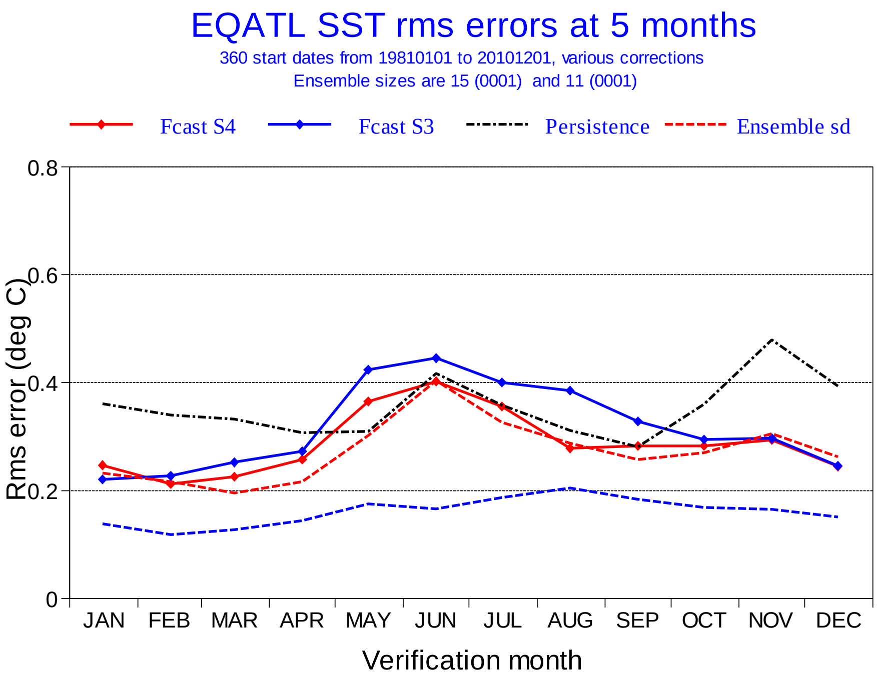

NINO1+2 SST rms errors
360 start dates from 19810101 to 20101201, various corrections
Ensemble sizes/corrections are 15/AS (0001) and 11/BC (0001)
95% confidence interval for 0001, for given set of start dates
Fcast S4
Fcast S3
Persistence
Ensemble sd
Rms error (deg C)
Forecast time (months)

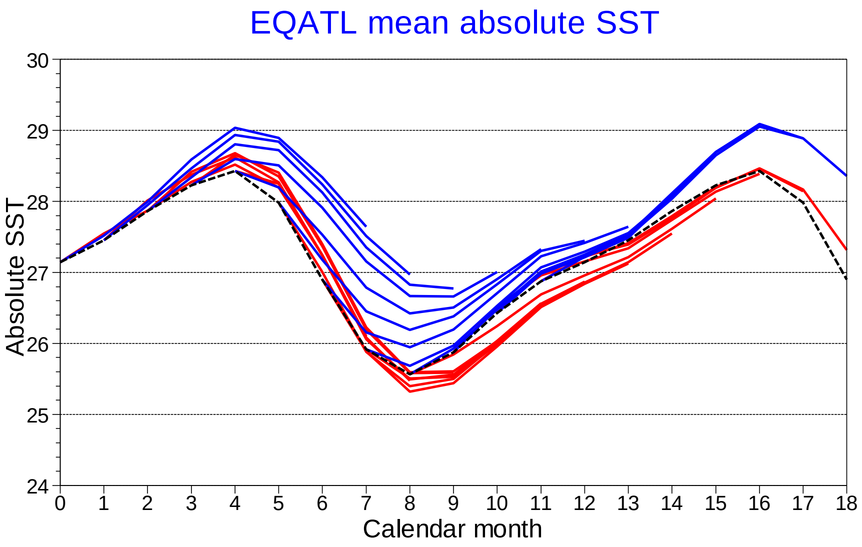

NINO1+2 SST rms errors at 5 months
360 start dates from 19810101 to 20101201, various corrections
Ensemble sizes are 15 (0001) and 11 (0001)
Fcast S4
Fcast S3
Persistence
Ensemble sd
Rms error (deg C)
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
Verification month


NINO1+2 mean absolute SST
Absolute SST
Calendar month

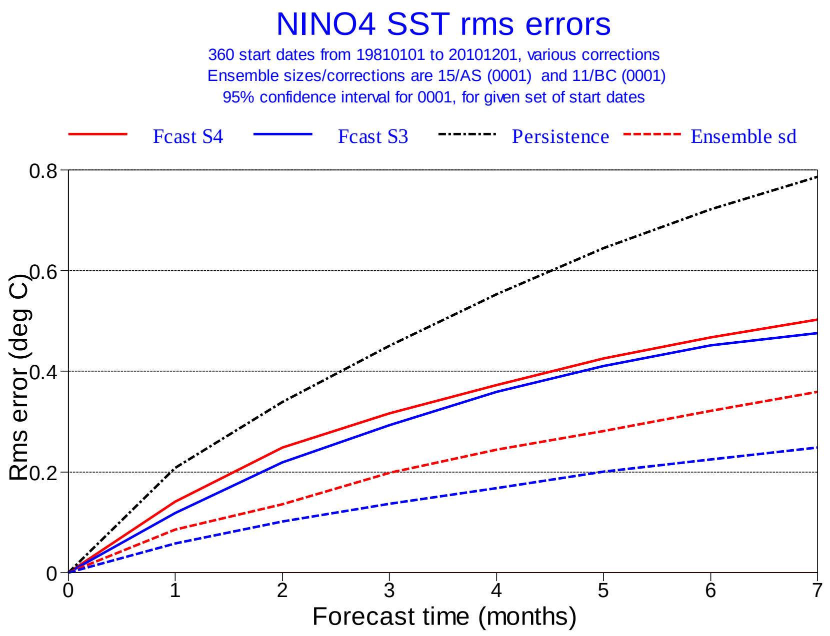
Equatorial Atlantic

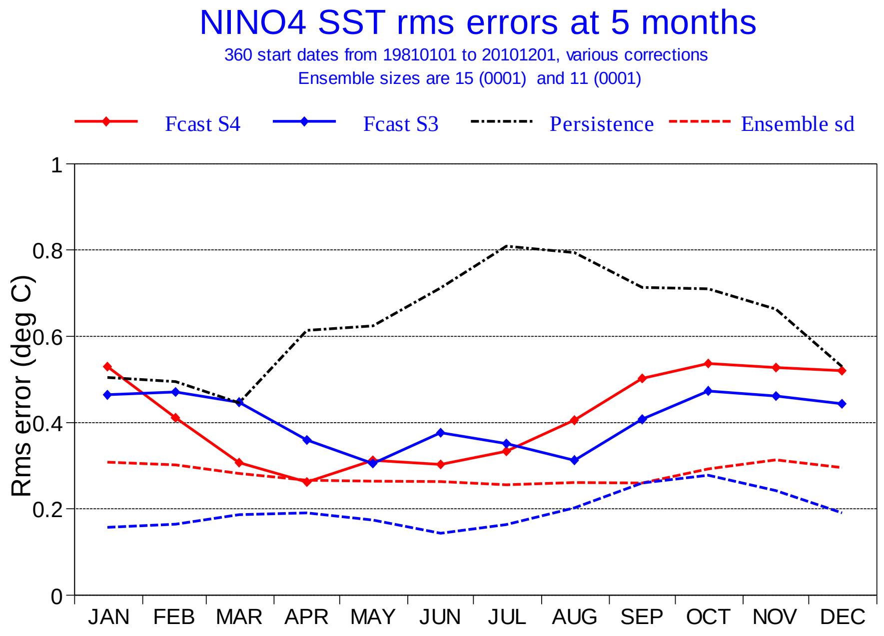


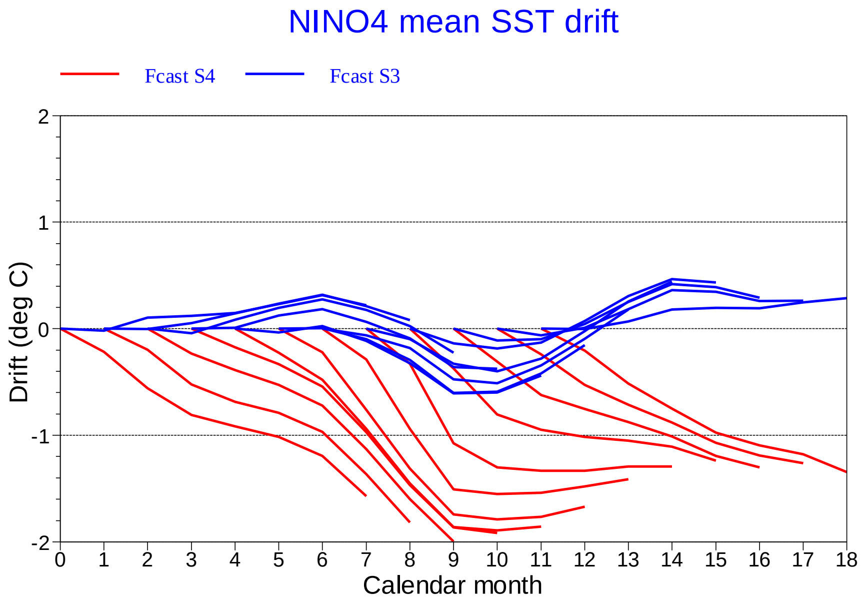




NINO 4


NINO4 SST rms errors
360 start dates from 19810101 to 20101201, various corrections
Ensemble sizes/corrections are 15/AS (0001) and 11/BC (0001)
95% confidence interval for 0001, for given set of start dates
Fcast S4
Fcast S3
Persistence
Ensemble sd
Rms error (deg C)
Forecast time (months)


NINO4 SST rms errors at 5 months
360 start dates from 19810101 to 20101201, various corrections
Ensemble sizes are 15 (0001) and 11 (0001)
Fcast S4
Fcast S3
Persistence
Ensemble sd
Rms error (deg C)
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC


NINO4 mean SST drift
Fcast S4
Fcast S3
Drift (deg C)
Calendar month

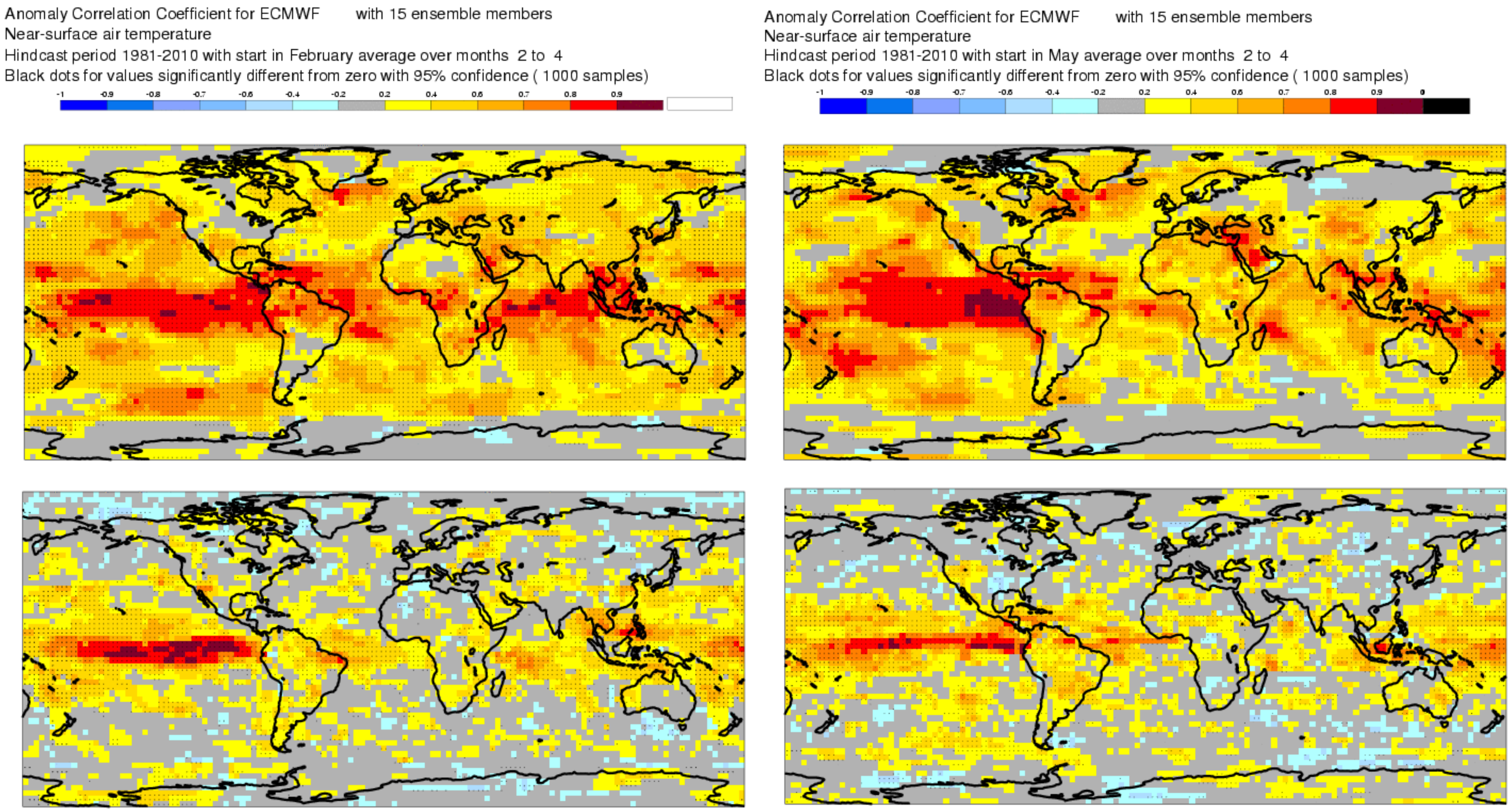
How good are the atmospheric forecasts?

Deterministic skill: ACC

MAM **JJA**

T2m

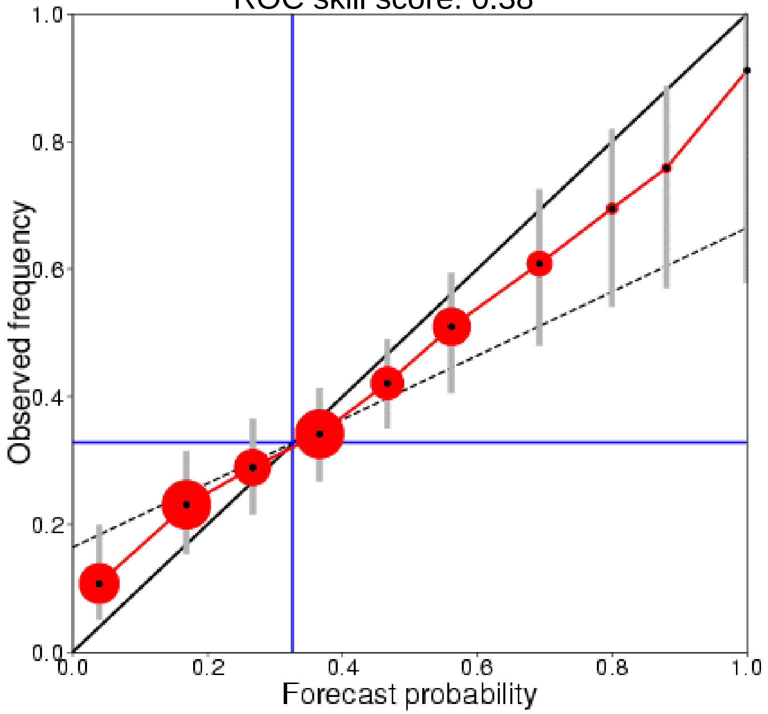
Precip



Scores for Europe: JJA

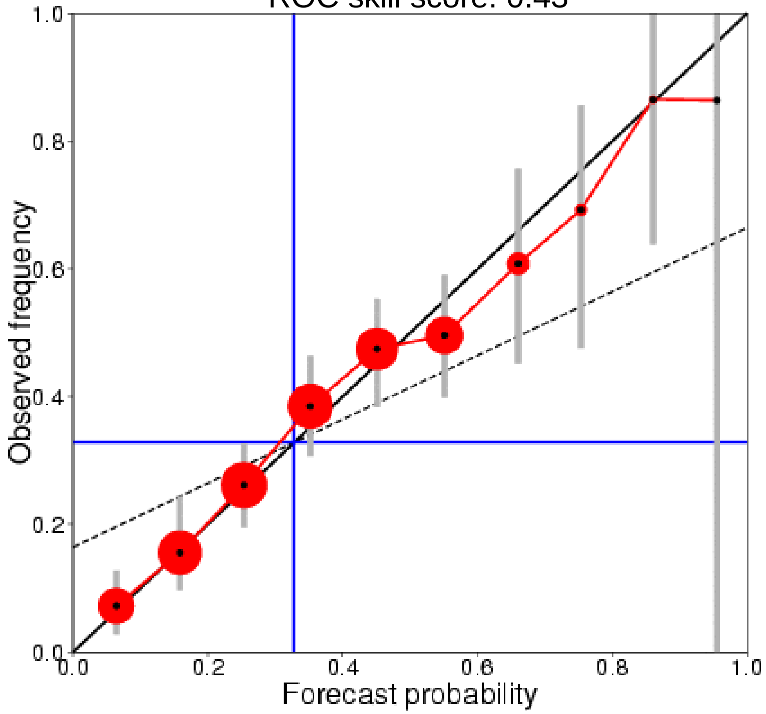
15 members

JJA Europe T2m>upper tercile
Re-forecasts from 1 May, 1981-2010
Reliability score: 0.987
ROC skill score: 0.38



51 members

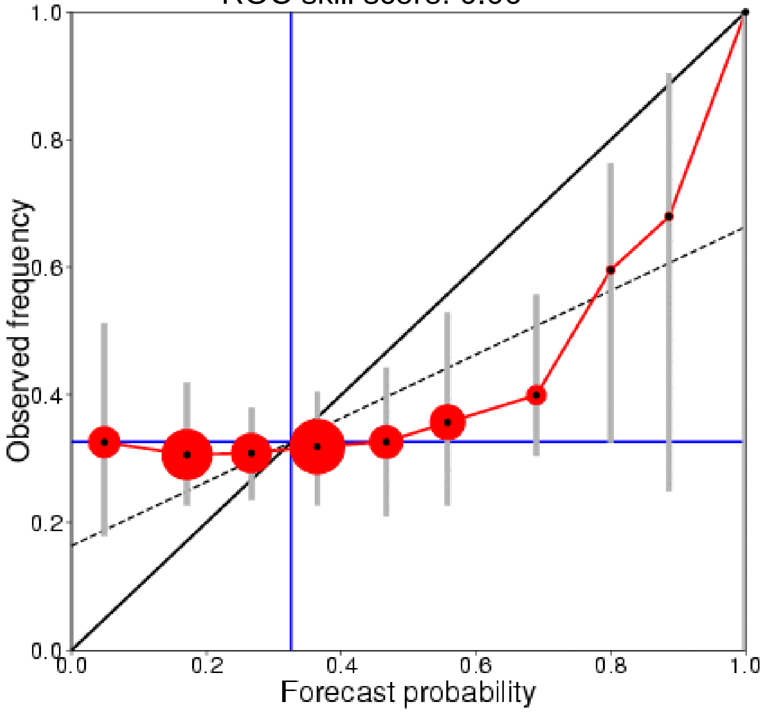
JJA Europe T2m>upper tercile
Re-forecasts from 1 May, 1981-2010
Reliability score: 0.996
ROC skill score: 0.43



Scores for Europe: DJF

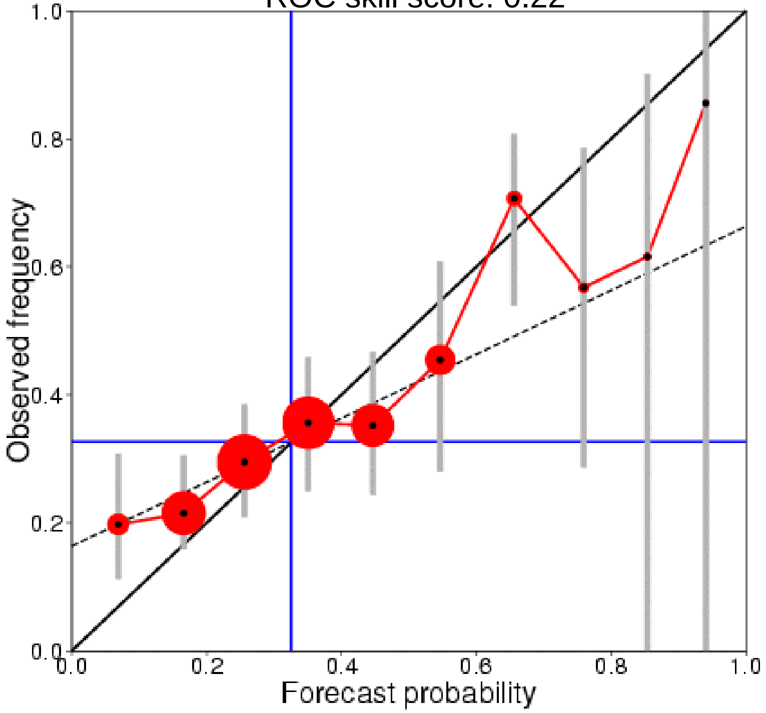
15 members

DJF Europe T2m>upper tercile
Re-forecasts from 1 Nov, 1981-2010
Reliability score: 0.902
ROC skill score: 0.06



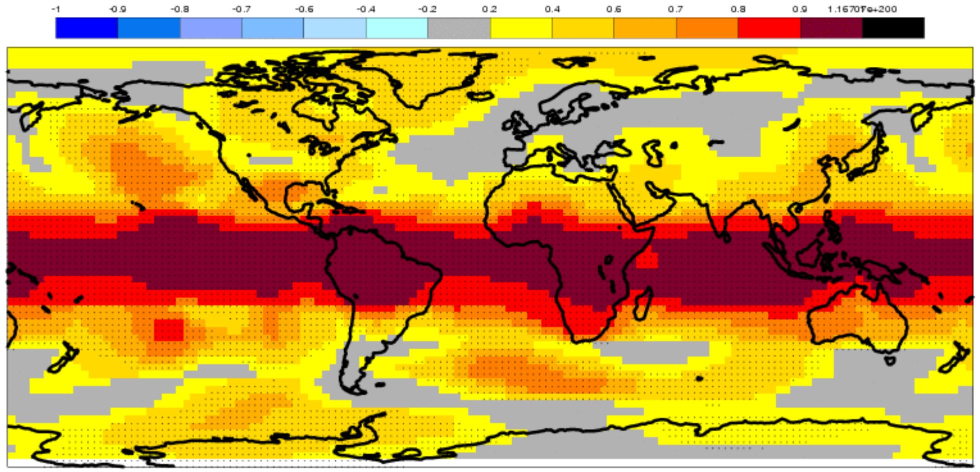
51 members

DJF Europe T2m>upper tercile
Re-forecasts from 1 Nov, 1981-2010
Reliability score: 0.981
ROC skill score: 0.22



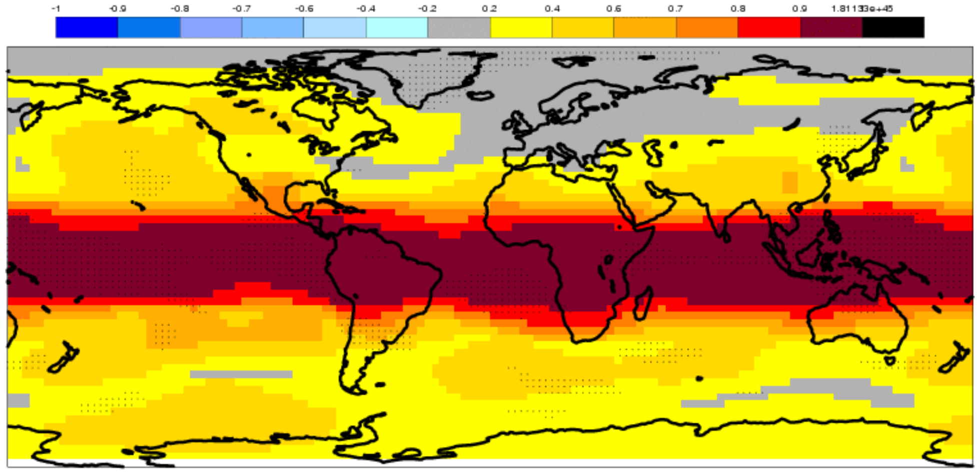
S4 ACC
DJF Z500

Anomaly Correlation Coefficient for ECMWF S4 with 51 ensemble members
500 hPa geopotential height
Hindcast period 1981-2010 with start in November average over months 2 to 4
Black dots for values significantly different from zero with 95% confidence (1000 samples)



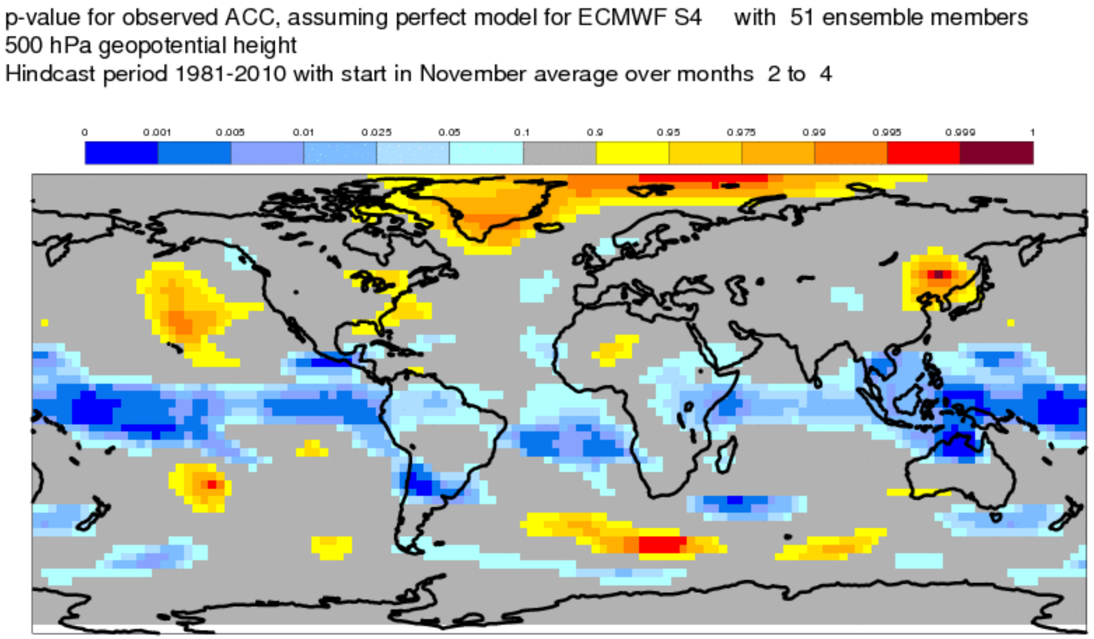
S4 ACC
perfect model limit

Perfect-model Anomaly Correlation Coefficient for ECMWF S4 with 51 ensemble members
500 hPa geopotential height
Hindcast period 1981-2010 with start in November average over months 2 to 4
Black dots where perfect model assumption is violated with 95% confidence (1000 samples)

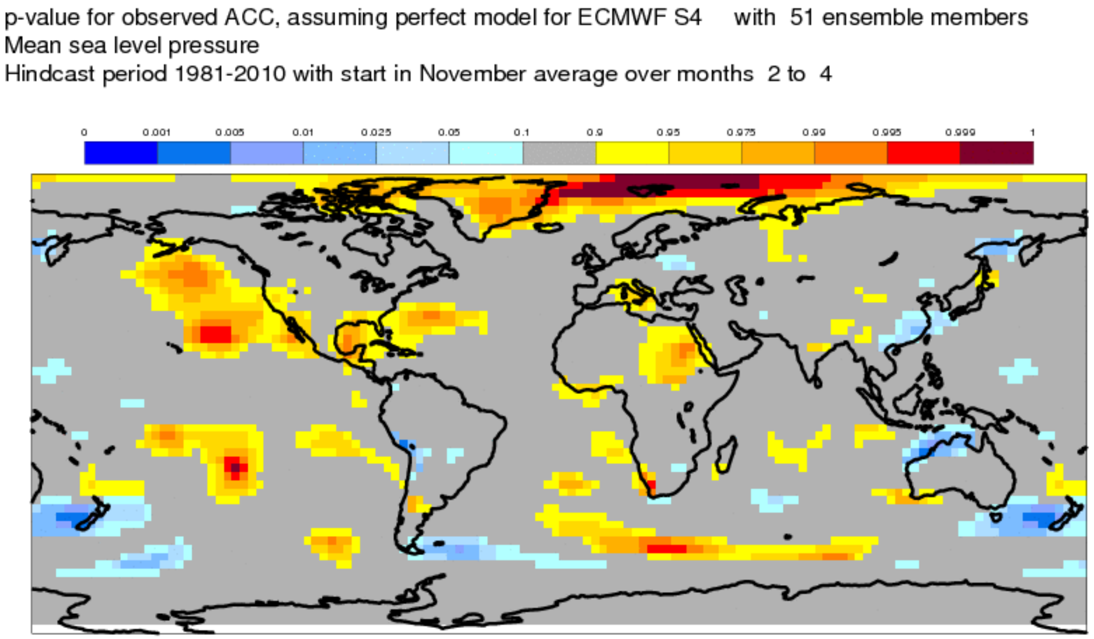


Predictive skill vs. Predictability limit

DJF Z500



DJF MSLP

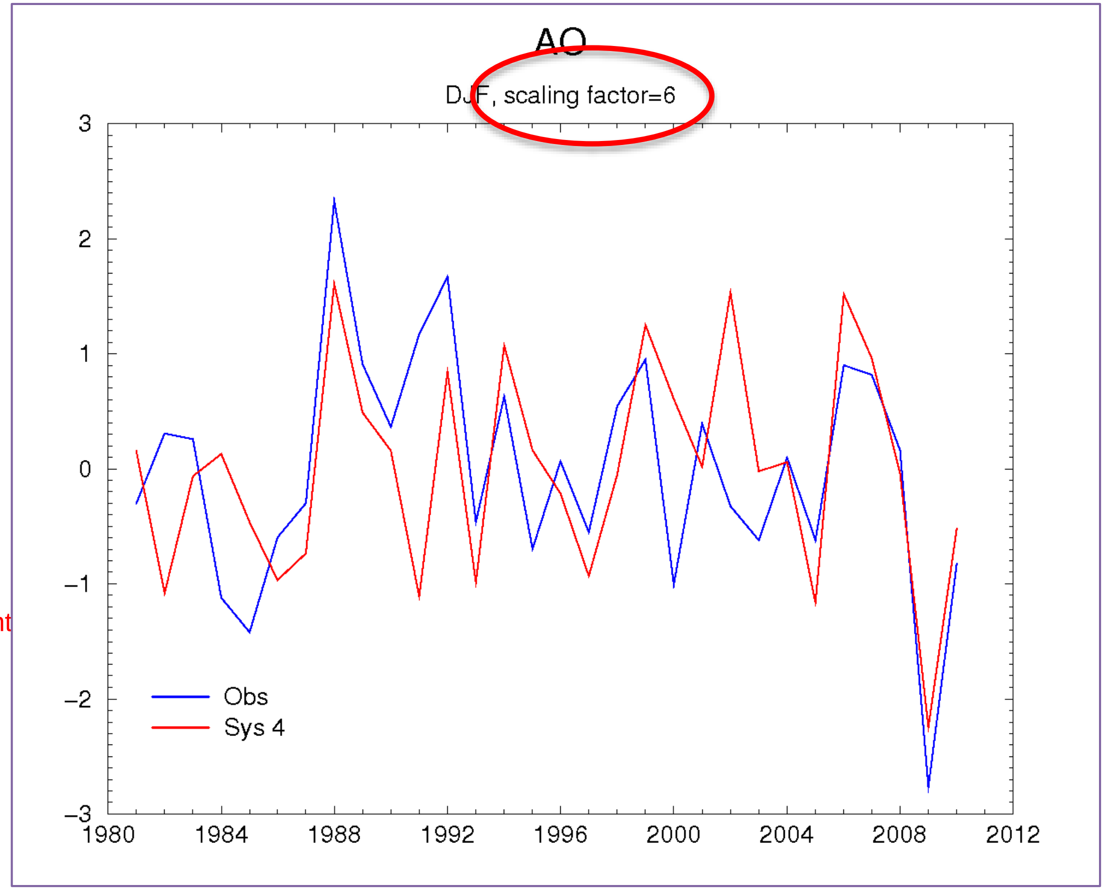


Indistinguishable from perfect
Worse than perfect
Better than perfect

Predictability of the Arctic Oscillation

Predictability can be under-estimated if we miss or under-represent important processes

Scope for improvement



Correlation (30y) =0.608

```
Unbiased variance estimates: Obs/Tot/Int/Ext:      1.0000      0.8390      0.8316

Model/obs stddev ratio:   0.9159
Model/obs stddev ratio interval:       0.693   1.129
Bootstrap over nens, pval for ratio=1:  0.7960

==========================================
SNR actual                      :   0.0941
SNR jackknife over nens         :   0.0202  0.1029  0.1857
==========================================

==========================================
ACC actual                      :   0.6085
ACC basic bootstrap over nens   :   0.5568  0.7121  0.8144
ACC basic bootstrap over nyears:   0.2052  0.6069  0.8326
==========================================

ACP from internal sampling: -0.2947  0.0583  0.4010
Mean ACC for nens-1:   0.6049
p val of measured acc if model perfect:   0.9996
```

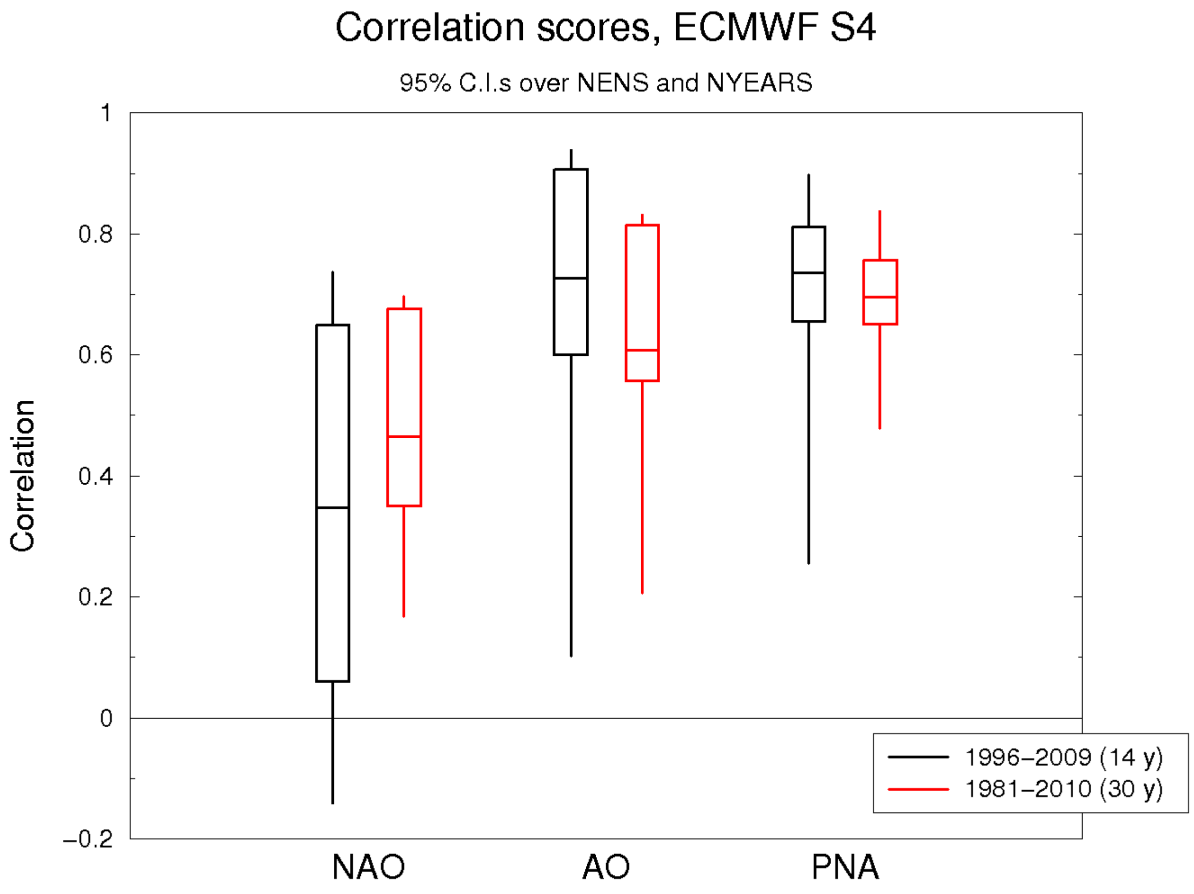
← model variability consistent

← 95% interval due to ensemble size

← bigger uncertainty range here

← only a 0.0004 chance we could get this correlation

Challenge: sampling errors are large!



Box = 95% interval, bootstrapping on ensemble size

Whiskers = 95% interval, bootstrapping on years included

How good are the forecasts?

- Skill (ACC, BSS, ROC, …) relative to climate is typically moderate to high in the tropics, moderate to low in mid-latitudes.
- Reliability is on average moderately good; large ensemble sizes are needed to measure this in low-predictability areas.
- Even with large ensemble sizes, the limited number of years means that skill assessments have large uncertainties in mid-latitude regions.
- Can average skill over many gridpoints, seasons etc, but trade resolution to gain accuracy.
- Indications that in some cases the forecast spread is too large.

Model error and forecast interpretation

- Model error is still quite large, relative to requirements
  - It still dominates some SST forecast errors (e.g. west Pacific)
  - Mean state and variability errors are very significant
  - Errors cannot be easily fixed

- Products typically account for sampling error only
  - Don’t take model probabilities as true probabilities

- Estimating forecast skill can be difficult
  - In many cases, data is insufficient to produce sensible estimates
  - This problem will not go away

- In the end we need trustworthy models
  - (Multi-model ensembles are helpful, but only partially span the space of model errors)

Some final comments

- Plenty of scope for improving model forecasts
  - Nino SST forecasts, while good, are still worse than predictability limits
  - Model errors still obvious in many cases, some processes poorly treated
  - Ocean initial conditions ~OK in Pacific since about 1993, recently improved elsewhere by ARGO
- Model output -> use of forecast
  - Calibration and presentation of forecast information
  - Potential for multi-model ensembles
  - Integration with decision making
- Timescale for improvements
  - Optimist: in 10 years, we'll have much better models, pretty reliable forecasts, confidence in our ability to handle climate variations
  - Pessimist: in 10 years, modelling will still be a hard problem, and progress will largely be down to improved calibration