Strategies for improving seasonal prediction

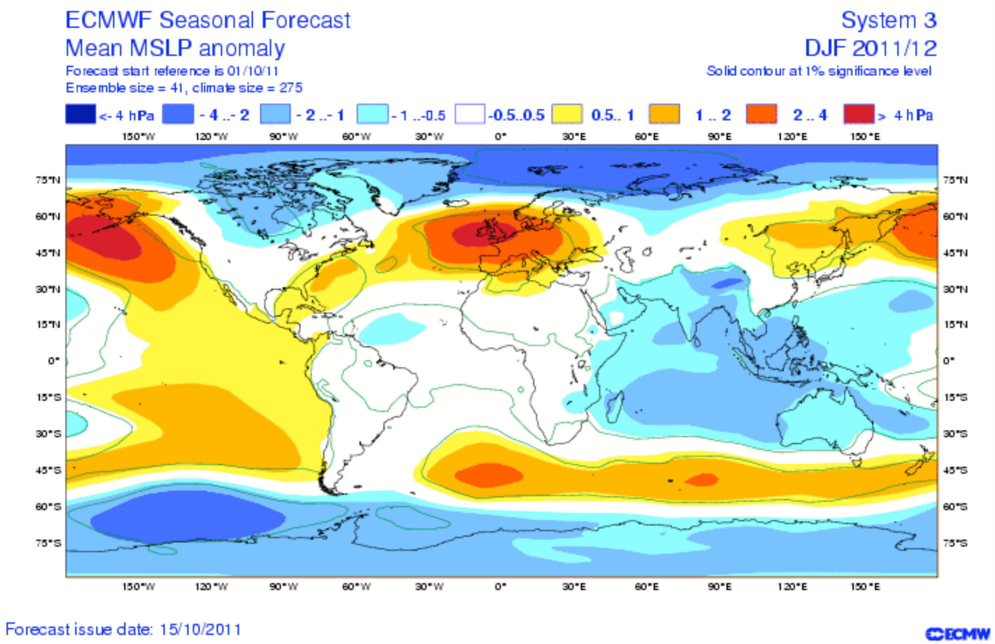
Tim Stockdale, Franco Molteni, Magdalena Balmaseda, Kristian Mogensen and Laura Ferranti

Outline

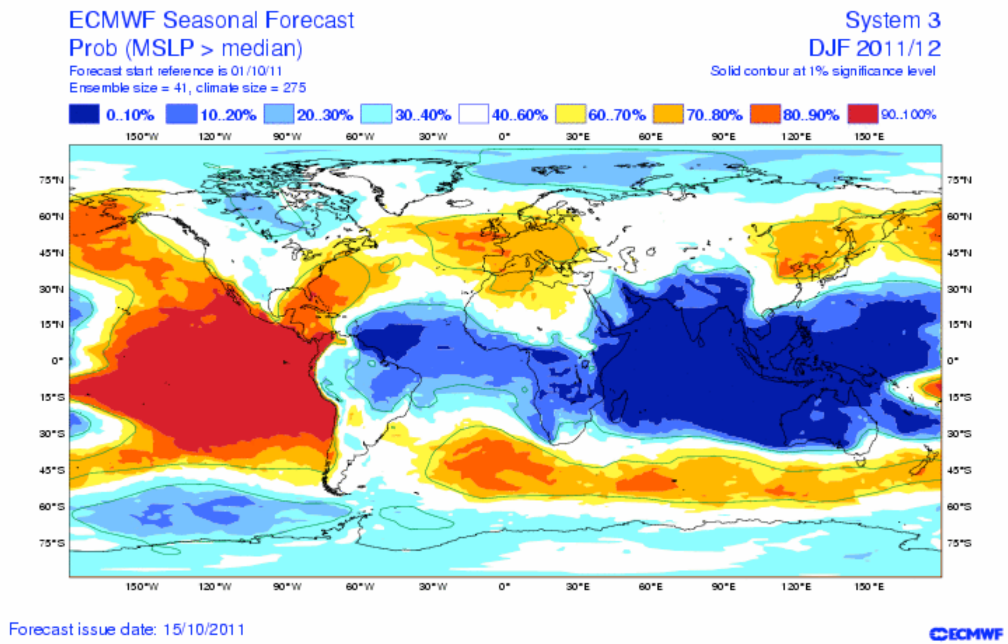
- **The seasonal prediction problem is tough**
  - The need for **accuracy**: DJF 2011/12
  - Sampling limitations
- **Forecast system improvements**
  - Example of ECMWF System 4
  - Benefits can be demonstrated, but challenges remain
- **Benefits of multi-model**
- **Conclusions**

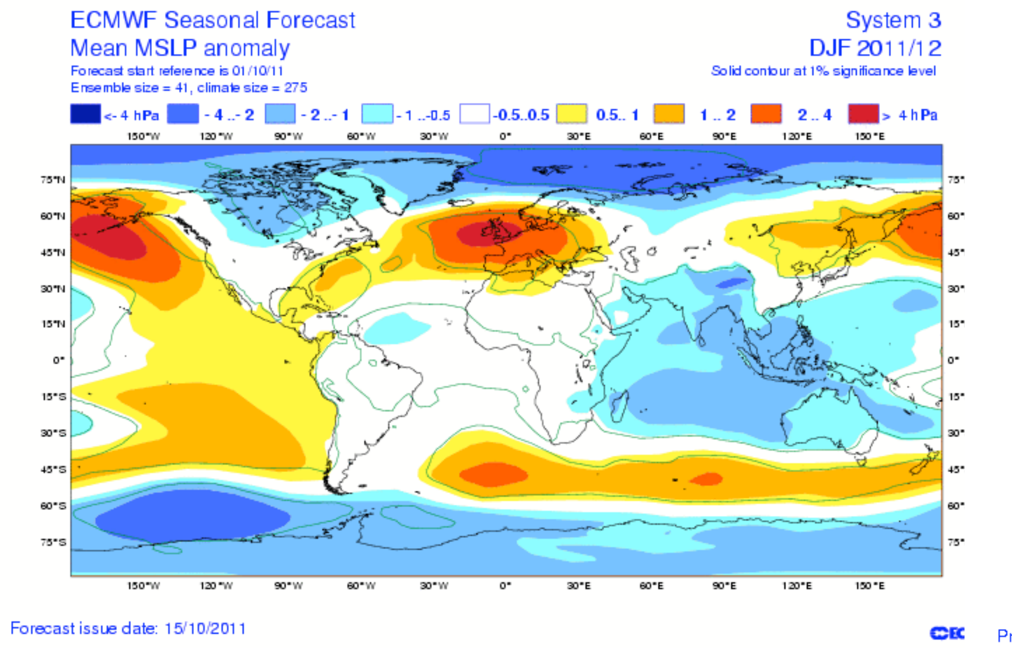
MSLP DJF 2011/12, ECMWF S3:

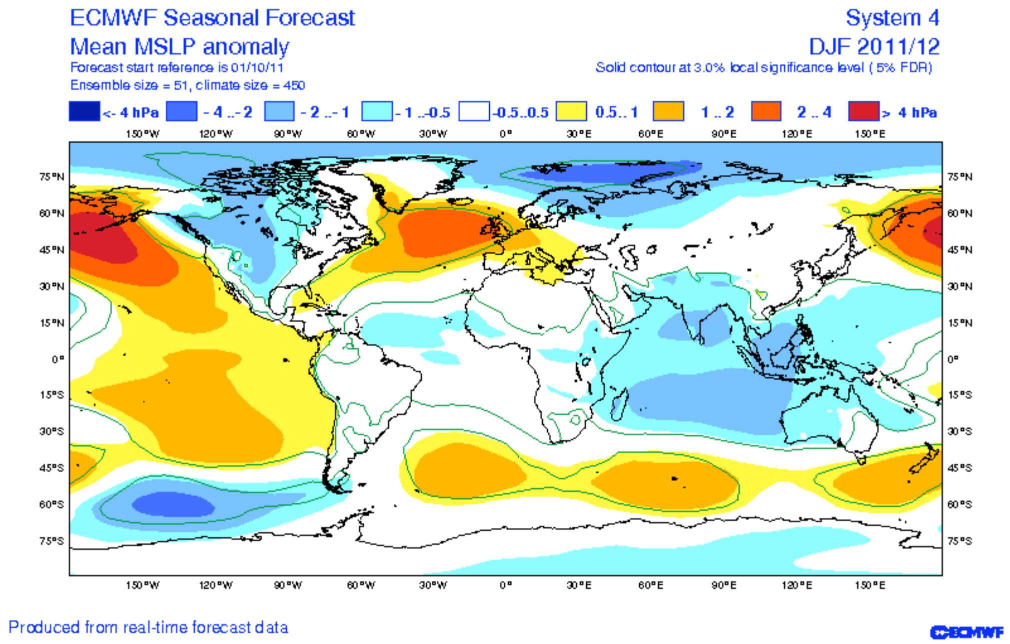
Ensemble mean

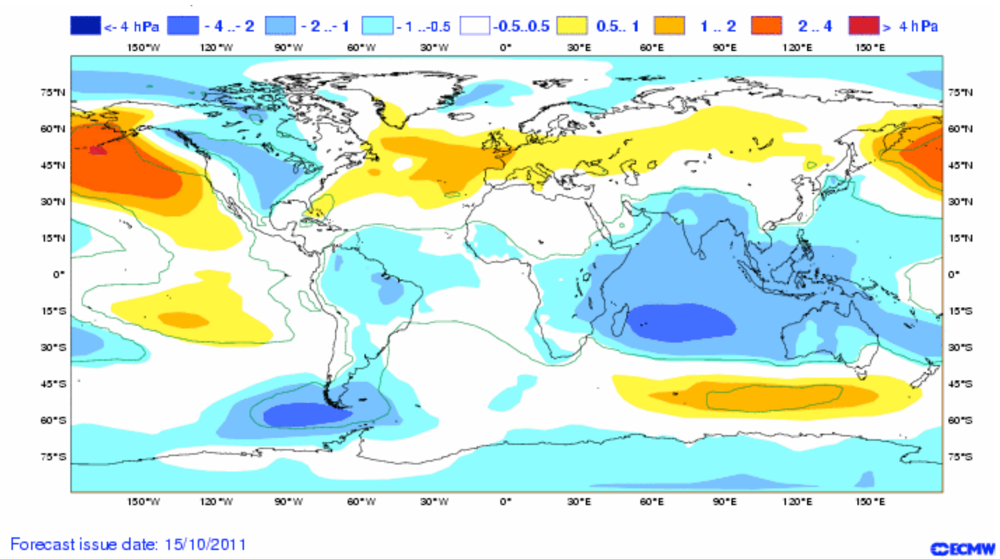


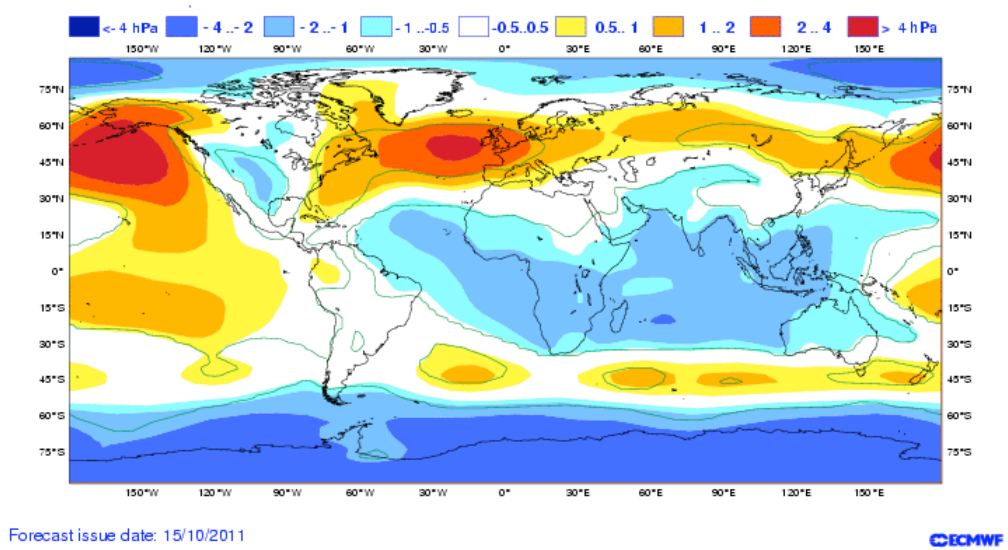
Prob MSLP > median

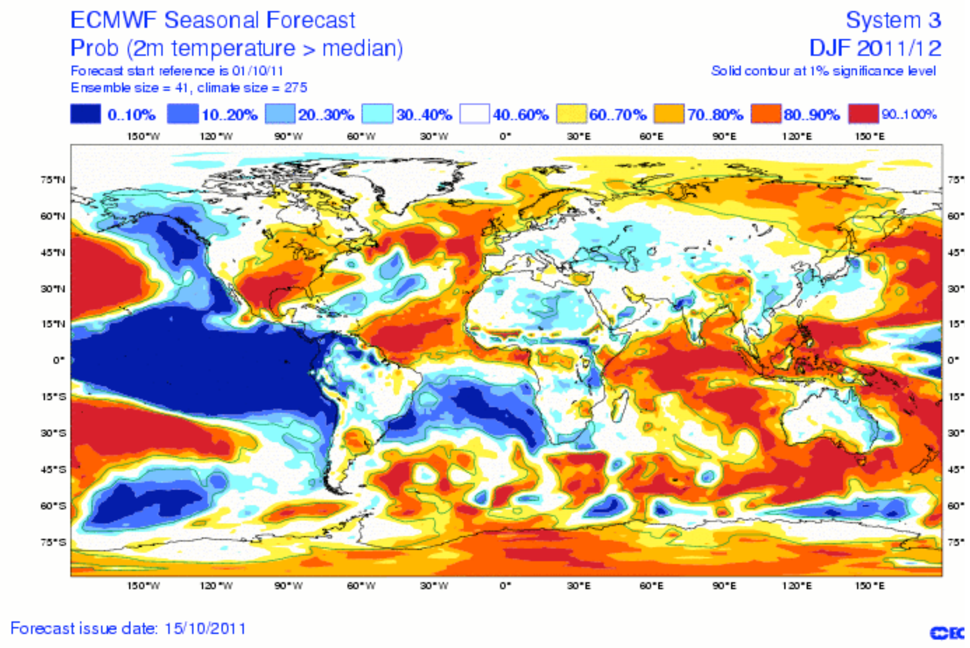


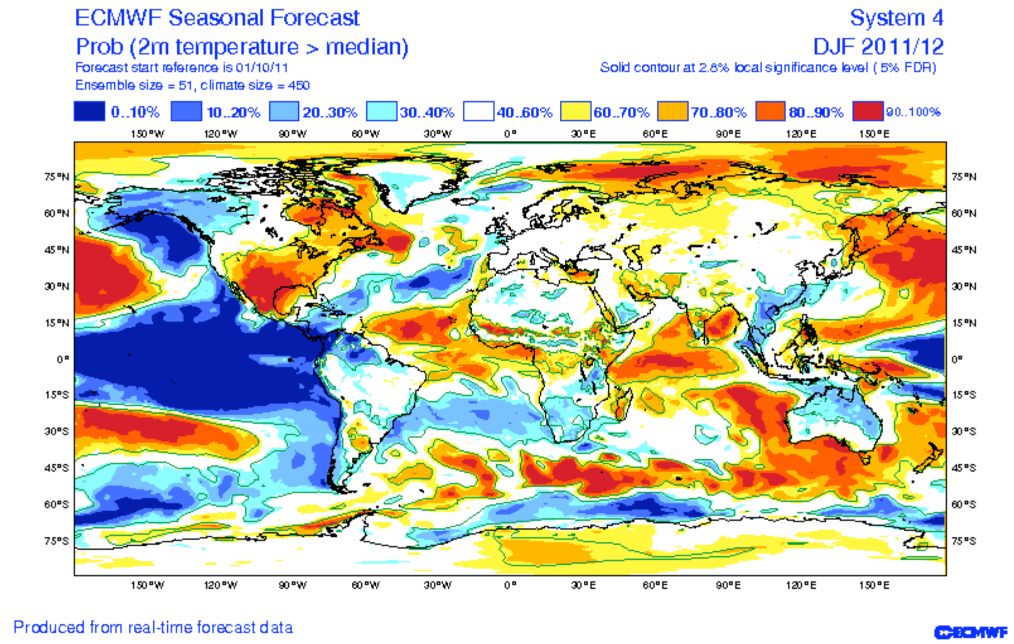

ECMWF Seasonal Forecast
Mean MSLP anomaly
Forecast start reference is 01/10/11
Ensemble size = 41, climate size = 275
System 3
DJF 2011/12
Solid contour at 1% significance level
<- 4 hPa
- 4 ..- 2
- 2 ..- 1
- 1 ..-0.5
-0.5..0.5
0.5.. 1
1 .. 2
2 .. 4
> 4 hPa
Forecast issue date: 15/10/2011

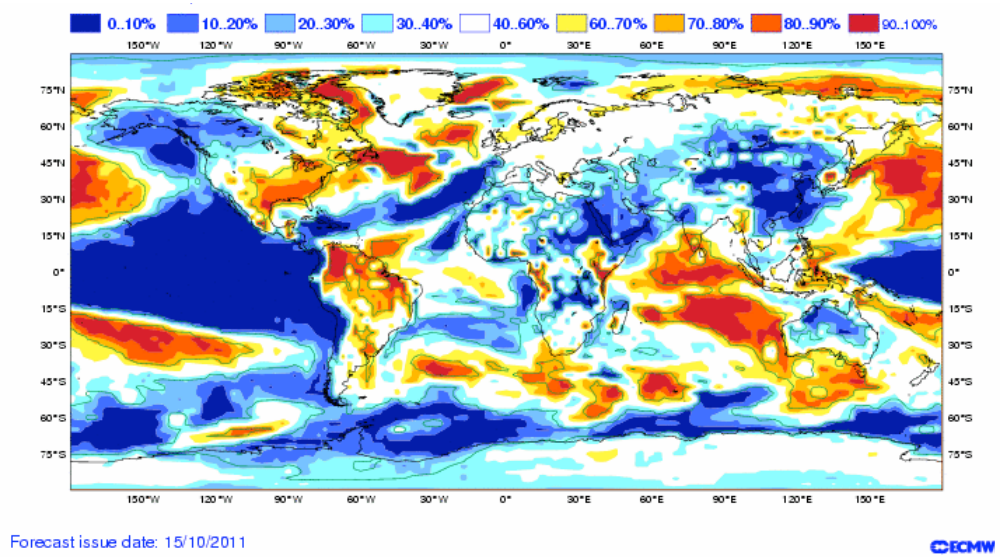

ECMWF Seasonal Forecast
Mean MSLP anomaly
Forecast start reference is 01/10/11
Ensemble size = 51, climate size = 450
System 4
DJF 2011/12
Solid contour at 3.0% local significance level (5% FDR)
<- 4 hPa
- 4 ..- 2
- 2 ..- 1
- 1 ..-0.5
-0.5..0.5
0.5.. 1
1 .. 2
2 .. 4
> 4 hPa
Produced from real-time forecast data

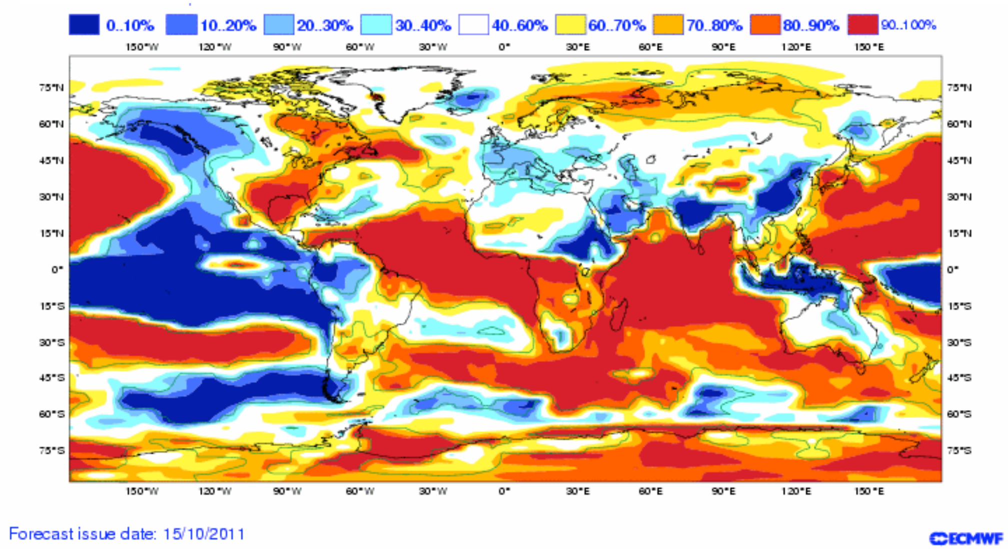

<- 4 hPa
- 4 ..- 2
- 2 ..- 1
- 1 ..-0.5
-0.5..0.5
0.5.. 1
1 .. 2
2 .. 4
> 4 hPa
Forecast issue date: 15/10/2011


<- 4 hPa
- 4 ..- 2
- 2 ..- 1
- 1 ..-0.5
-0.5..0.5
0.5.. 1
1 .. 2
2 .. 4
> 4 hPa
Forecast issue date: 15/10/2011


ECMWF Seasonal Forecast
Prob (2m temperature > median)
Forecast start reference is 01/10/11
Ensemble size = 41, climate size = 275
System 3
DJF 2011/12
Solid contour at 1% significance level
0..10% 10..20% 20..30% 30..40% 40..60% 60..70% 70..80% 80..90% 90..100%
Forecast issue date: 15/10/2011


ECMWF Seasonal Forecast
Prob (2m temperature > median)
Forecast start reference is 01/10/11
Ensemble size = 51, climate size = 450
System 4
DJF 2011/12
Solid contour at 2.8% local significance level (5% FDR)
0..10% 10..20% 20..30% 30..40% 40..60% 60..70% 70..80% 80..90% 90..100%
Produced from real-time forecast data


0..10% 10..20% 20..30% 30..40% 40..60% 60..70% 70..80% 80..90% 90..100%
Forecast issue date: 15/10/2011


0..10% 10..20% 20..30% 30..40% 40..60% 60..70% 70..80% 80..90% 90..100%
Forecast issue date: 15/10/2011

Sampling limitations

- **Re-forecasts have small number of events**
  - Each forecast gives a pdf – obs could be anywhere in that pdf
  - For low or intermittent signal areas, 30 years is a very small sample!

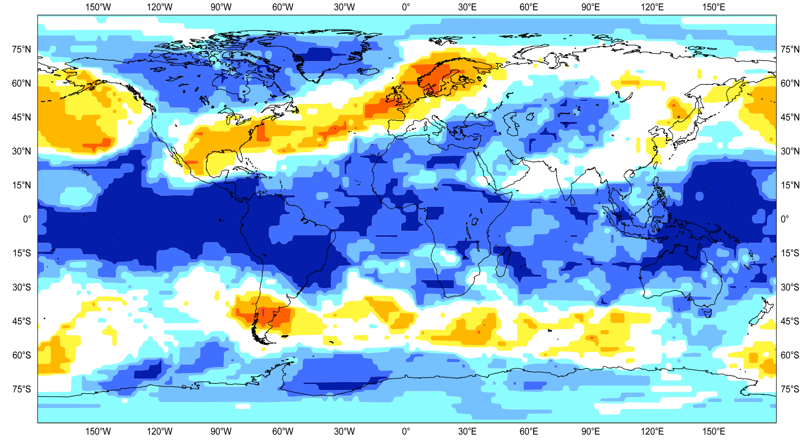
- **Re-forecasts are (usually) small ensembles**
  - Forecast pdfs are not that well sampled, especially in re-forecasts
  - Easy to end up calculating scores by correlating "mostly noise" with "mostly noise" ….

  - One practical benefit of multi-model – spreads the cost of producing large hindcast ensembles (eg 100 members, 30 years, 12 start dates, 7 months = 21,000 years of model integration)

1989/90

Expt fg79
Prob (Z500 > median) DJF 1989/90
Forecast start reference is 01/11/89
Ensemble size = 11, climate size = 220
No field significance at 5% level (FDR test)

0..10% | 10..20% | 20..30% | 30..40% | 40..60% | 60..70% | 70..80% | 80..90% | 90..100%

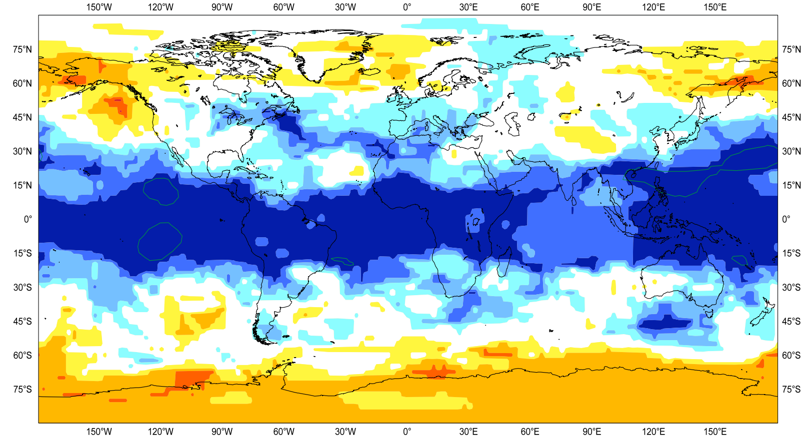


Produced from hindcast data

ECMWF

Expt fgcn
Prob (Z500 > median) DJF 1989/90
Forecast start reference is 01/11/89
Ensemble size = 11, climate size = 220
Solid contour at 0.08% local significance level (5% FDR)

0..10% | 10..20% | 20..30% | 30..40% | 40..60% | 60..70% | 70..80% | 80..90% | 90..100%

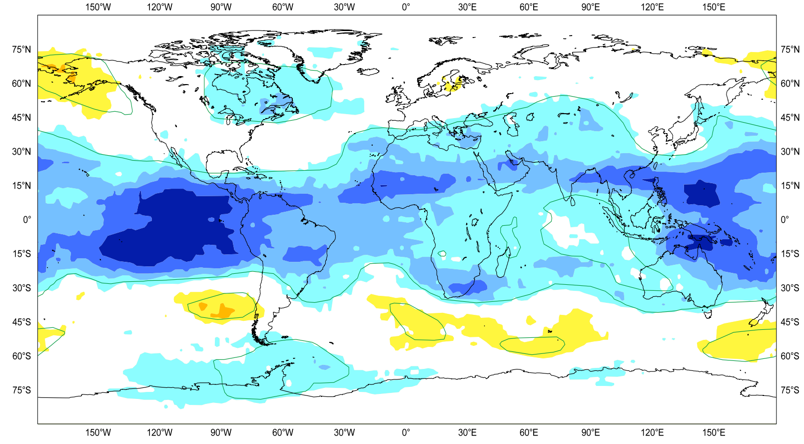


Produced from hindcast data

ECMWF

Expt fg79
Prob (Z500 > median) DJF 1989/90
Forecast start reference is 01/11/89
Ensemble size = 101, climate size = 220
Solid contour at 2.1% local significance level (5% FDR)

0..10% | 10..20% | 20..30% | 30..40% | 40..60% | 60..70% | 70..80% | 80..90% | 90..100%

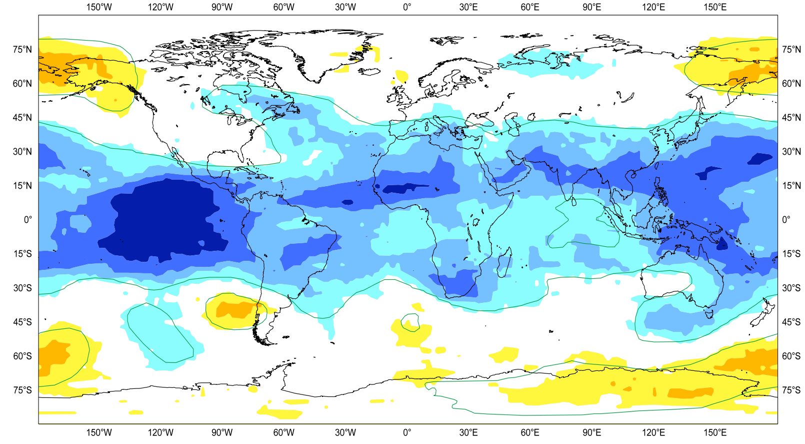


Produced from hindcast data

ECMWF

Expt fgcn
Prob (Z500 > median) DJF 1989/90
Forecast start reference is 01/11/89
Ensemble size = 101, climate size = 220
Solid contour at 2.5% local significance level (5% FDR)

0..10% | 10..20% | 20..30% | 30..40% | 40..60% | 60..70% | 70..80% | 80..90% | 90..100%



Produced from hindcast data

ECMWF

An improved forecast system:

- **ECMWF System 4**
  - Replaces System 3, operational since March 2007
  - Many changes, lots of testing, large re-forecast set now complete
- **Major model changes**
  - **NEMO ocean model** replaces HOPE. Similar resolution, but better mixed layer physics.
  - New **IFS cycle 36r4** (circa 5 years progress)
  - **T255** horizontal resolution (cf T159)
  - **L91**, and **enhanced stratospheric physics** (cf L62)
  - Stochastic physics: SPPT3 and stochastic backscatter instead of old SPPT: SPPT3 represents model uncertainty – **big spread** in ENSO forecasts
  - **Ice** sampled from preceding five years instead of fixed climatology

S4 initial conditions

- **Major initial condition changes**
  - NEMOVAR ocean analysis/re-analysis. New 3D-VAR system, incorporating all major elements of previous system, but many aspects of re-analyses are improved.
  - Land surface initial conditions: offline run of HTESSEL, with GPCP-corrected ERA interim forcing (re-forecasts); operational analyses (forecasts).
  - ERA Interim initial conditions for atmosphere to end 2010, then operations
  - Stratospheric ozone from climatology of selected ERA interim years (direct use of ozone analysis problematic).
  - Volcanic aerosol input as NH/TROPICS/SH zonal mean values at start of each integration

Benefits of an improved system

- **Much better mean state**
  - Mostly much better, but one important thing is worse
  - Progress is real, but not monotonic and not easy (experience of many modelling groups over the last 20 years)

- **Better ENSO forecasts**
  - Much better in NINO3, bit better in NINO34, bit worse in NINO4
  - Amplitude of ENSO too strong, mean state error problems

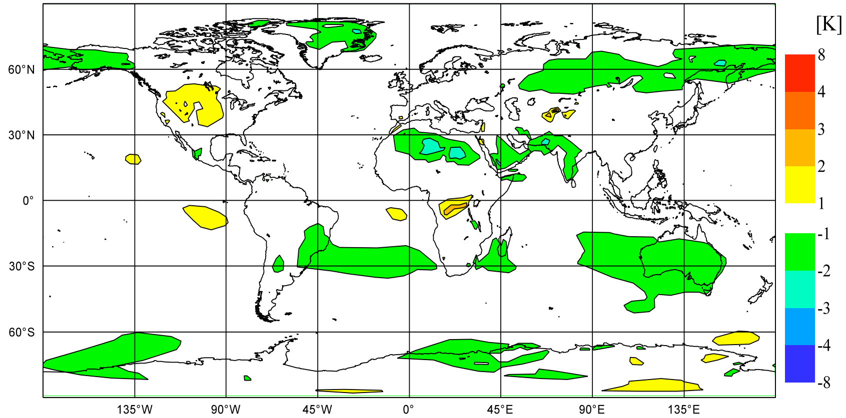
- **Better atmospheric forecasts**
  - Very strong consistent improvements in tropics, and strong improvements in NH scores also (but not all months, eg NH winter Z500 noisy)
  - Strong improvements both in ACC and in reliability scores

- **Big improvements, but not a "perfect" system yet!**

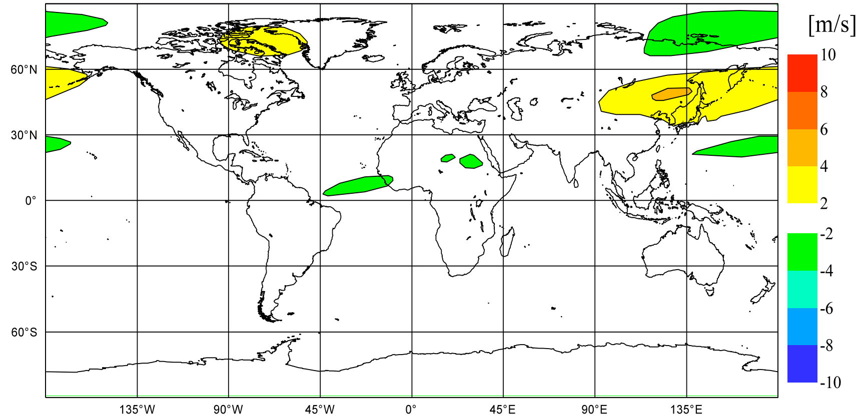
Mean state errors

850hPa temperature S4(15)-ERA Int 1991-2008 JJA
Global rms error: 0.663 NH:0.669 TR:0.662 SH:0.66

S4

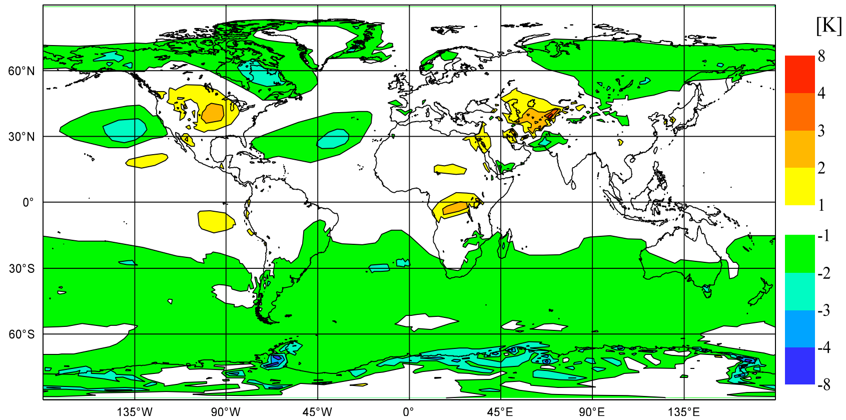


50hPa zonal wind S4(15)-ERA Int 1991-2008 DJF
Global rms error: 1 NH:1.43 TR:0.853 SH:0.72

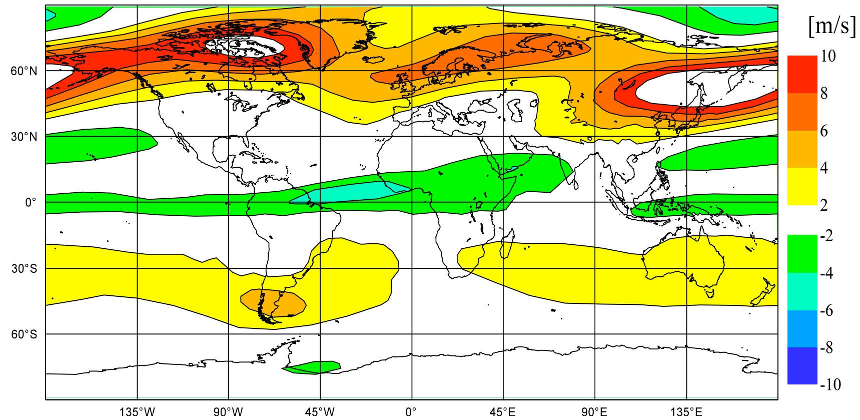


850hPa temperature S3(11)-ERA Int 1991-2008 JJA
Global rms error: 1.07 NH:1.06 TR:0.798 SH:1.48

S3



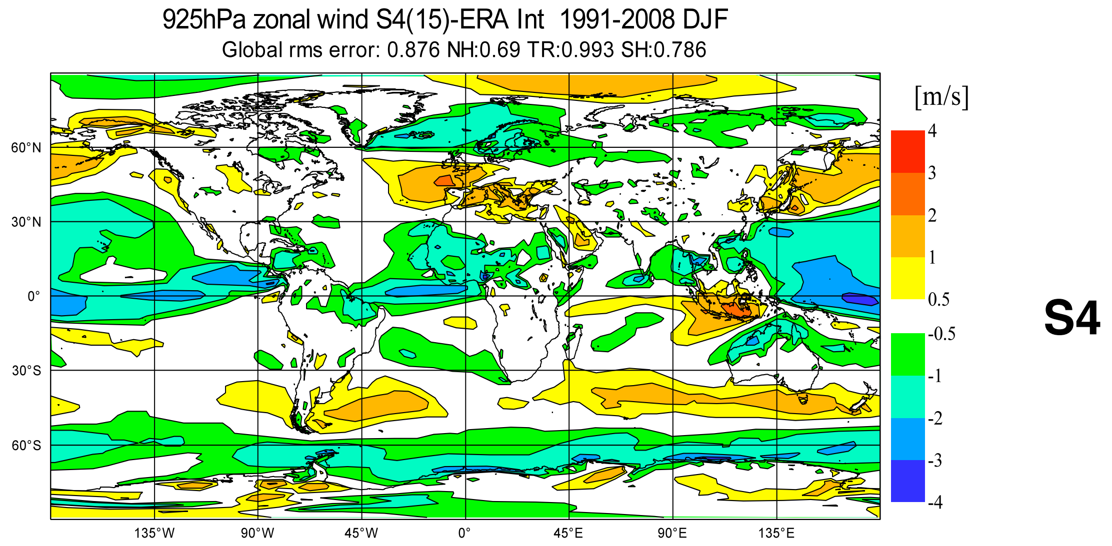
50hPa zonal wind S3(11)-ERA Int 1991-2008 DJF
Global rms error: 3.26 NH:5.53 TR:2.02 SH:2.03



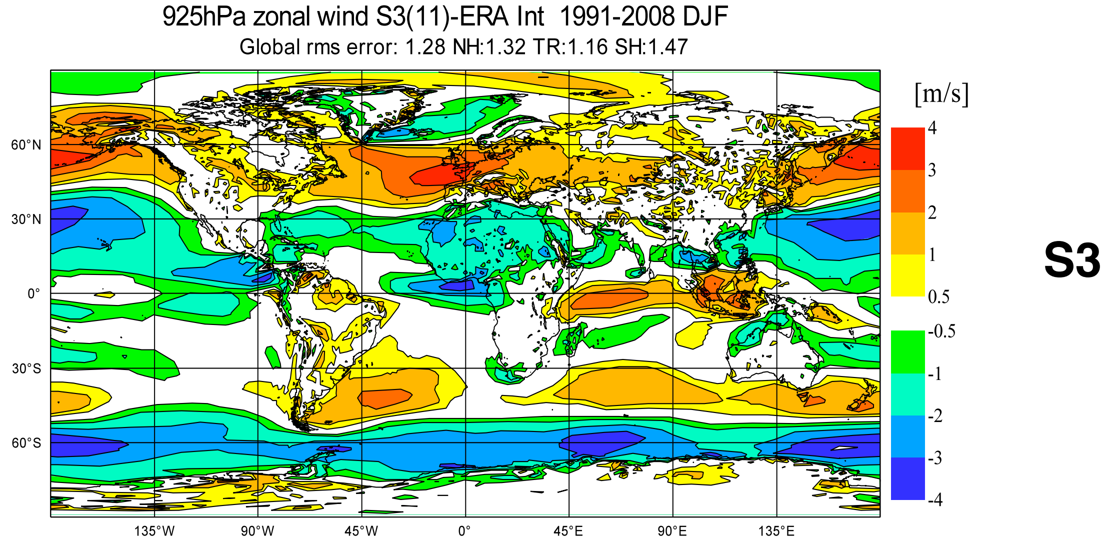
T850

U50

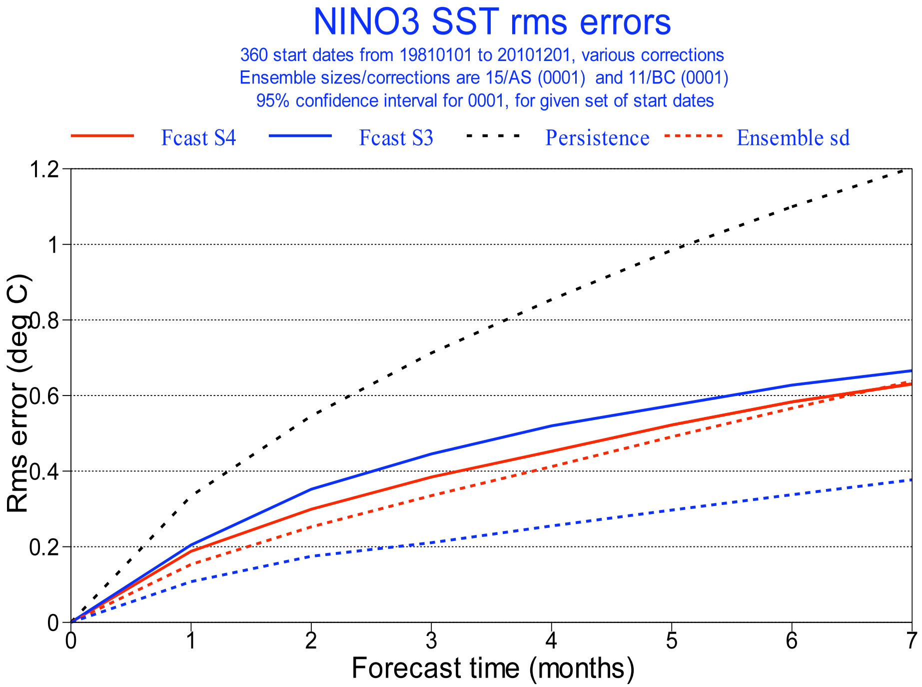
Mean state 925hPa winds

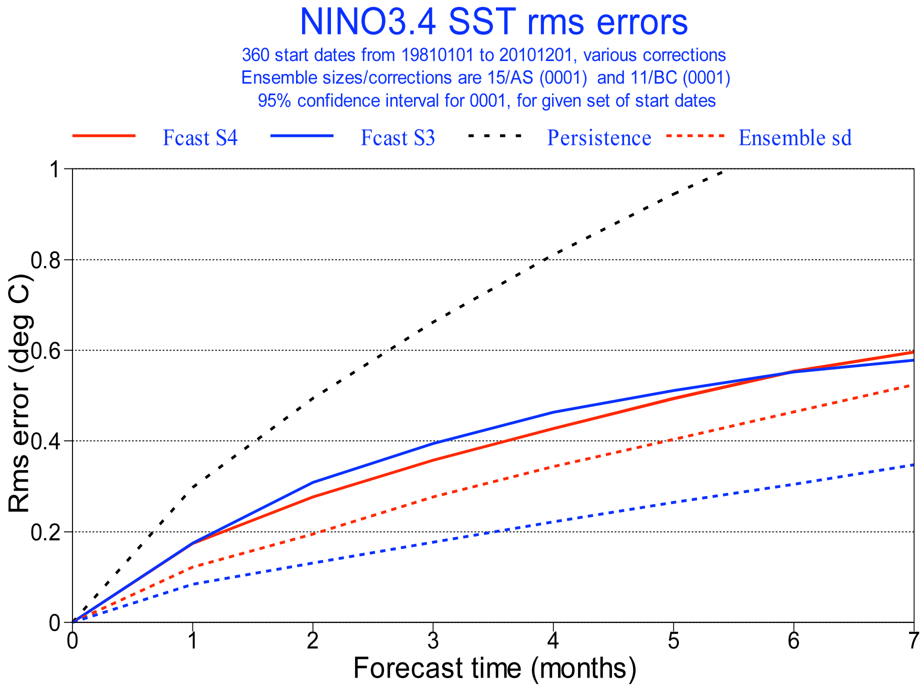


Overall biases are reduced, **but** wind bias in equatorial West Pacific is a problem

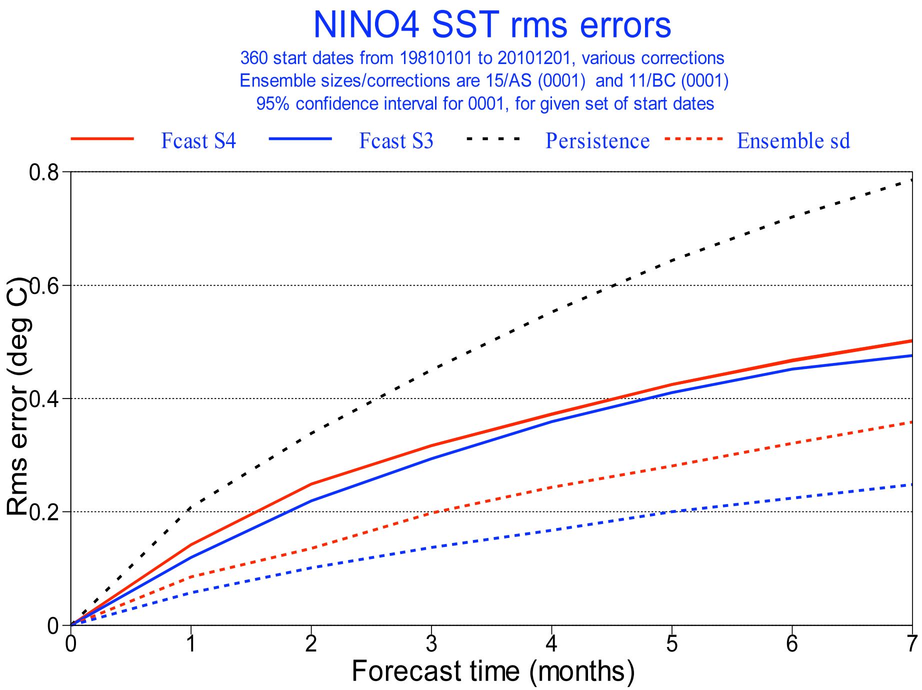


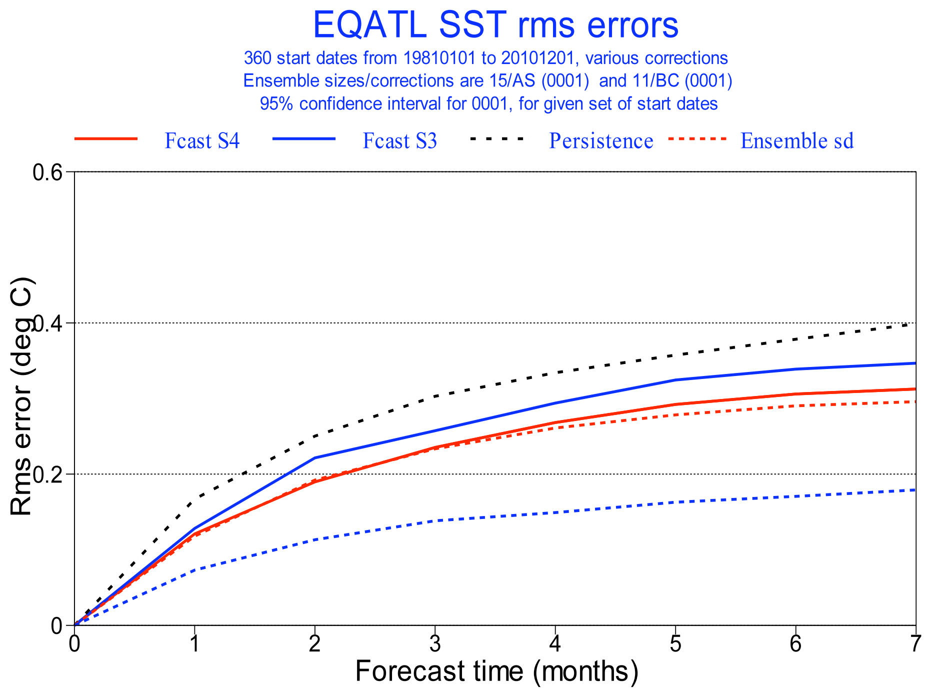
ENSO forecasts (S4, S3)





SST scores (S4, S3)





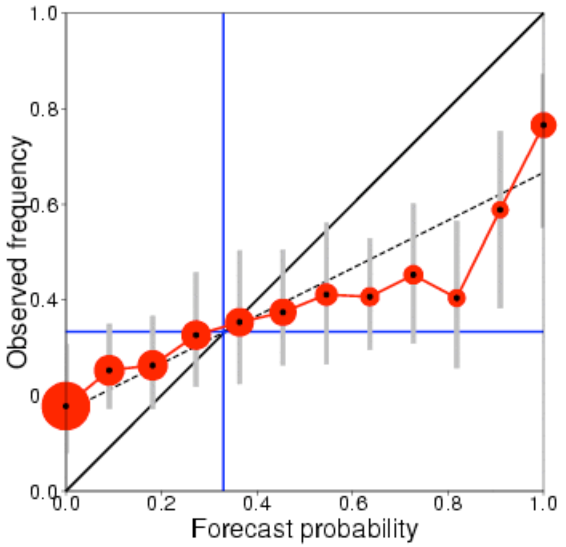
Tropospheric scores: ACC statistics (30y)

Field	Lead (months)	S3 mean	S4 mean	S4 wins
Tropics T850	1	0.573	**0.605**	12/12
Tropics T2m	1	0.601	**0.635**	12/12
NH Z500	1	0.246	**0.271**	7/12
NH T850	1	0.266	**0.307**	10.5/12
NH T2m	1	0.345	**0.376**	10/12
Tropics T850	4	0.471	**0.510**	11/12
Tropics T2m	4	0.462	**0.505**	12/12
NH Z500	4	0.167	**0.221**	11/12
NH T850	4	0.192	**0.249**	11/12
NH T2m	4	0.240	**0.287**	10/12

Statistic=z-transform spatial mean of ACC of 3 month forecast, 1981-2010
Assessed for each of 12 possible start months, and scores aggregated
“NH” is poleward of 30N, “Tropics” is 30N-30S

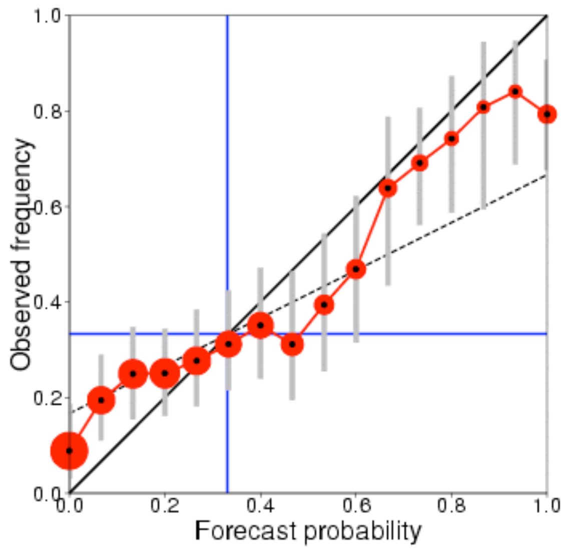
Probabilistic scores: reliability, S. America

Reliability diagram for ECMWF with 11 ensemble members
Near-surface air temperature anomalies above the upper tercile
Accumulated over South America (land points only)
Hindcast period 1981-2010 with start in May average over months 2 to 4
Skill scores and 95% conf. intervals (1000 samples)
Brier skill score: -0.030 (-0.221, 0.140)
Reliability skill score: 0.860 (0.726, 0.924)
Resolution skill score: 0.110 (0.042, 0.229)



S3

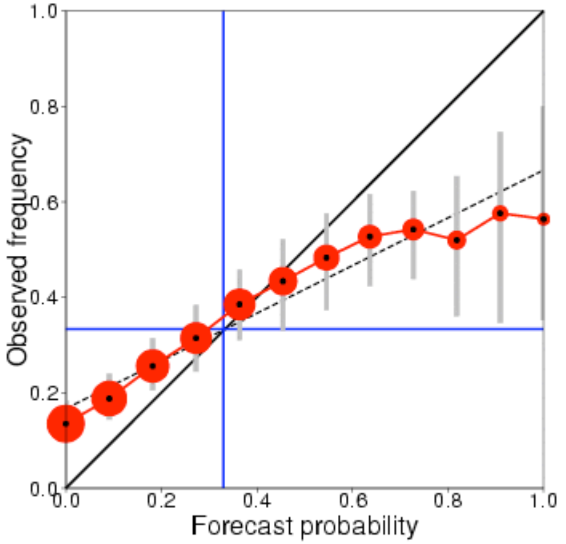
Reliability diagram for ECMWF with 15 ensemble members
Near-surface air temperature anomalies above the upper tercile
Accumulated over South America (land points only)
Hindcast period 1981-2010 with start in May average over months 2 to 4
Skill scores and 95% conf. intervals (1000 samples)
Brier skill score: 0.147 (0.012, 0.252)
Reliability skill score: 0.957 (0.888, 0.977)
Resolution skill score: 0.189 (0.106, 0.294)



S4

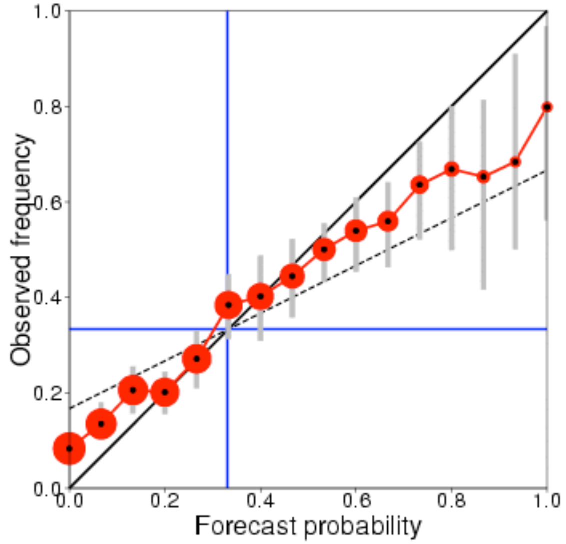
Probabilistic scores: reliability, Africa

Reliability diagram for ECMWF with 11 ensemble members
Near-surface air temperature anomalies above the upper tercile
Accumulated over Africa (land points only)
Hindcast period 1981-2010 with start in May average over months 2 to 4
Skill scores and 95% conf. intervals (1000 samples)
Brier skill score: 0.018 (-0.120, 0.109)
Reliability skill score: 0.923 (0.821, 0.960)
Resolution skill score: 0.095 (0.053, 0.150)



S3

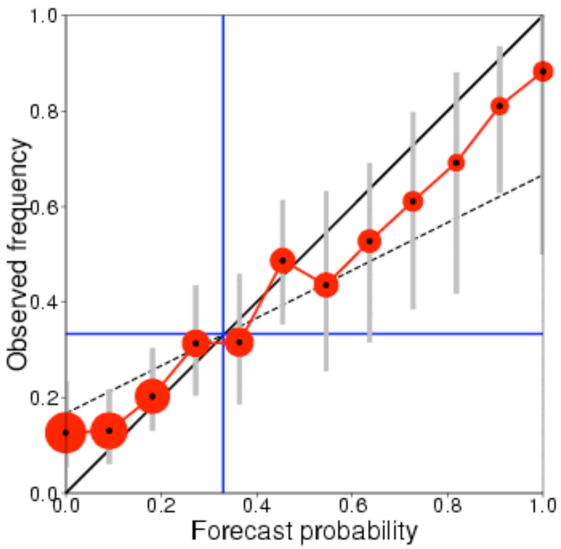
Reliability diagram for ECMWF with 15 ensemble members
Near-surface air temperature anomalies above the upper tercile
Accumulated over Africa (land points only)
Hindcast period 1981-2010 with start in May average over months 2 to 4
Skill scores and 95% conf. intervals (1000 samples)
Brier skill score: 0.129 (0.023, 0.202)
Reliability skill score: 0.975 (0.925, 0.988)
Resolution skill score: 0.154 (0.093, 0.219)



S4

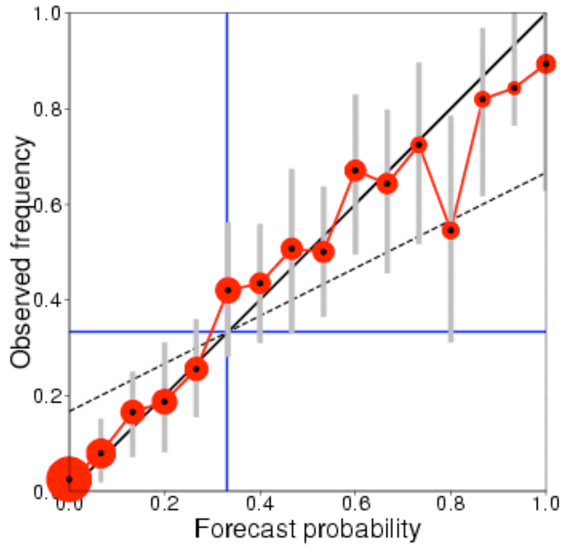
Probabilistic scores: reliability, SE Asia

Reliability diagram for ECMWF with 11 ensemble members
Near-surface air temperature anomalies above the upper tercile
Accumulated over Southeast Asia (land points only)
Hindcast period 1981-2010 with start in May average over months 2 to 4
Skill scores and 95% conf. intervals (1000 samples)
Brier skill score: 0.190 (-0.020, 0.353)
Reliability skill score: 0.967 (0.866, 0.985)
Resolution skill score: 0.222 (0.101, 0.373)



S3

Reliability diagram for ECMWF with 15 ensemble members
Near-surface air temperature anomalies above the upper tercile
Accumulated over Southeast Asia (land points only)
Hindcast period 1981-2010 with start in May average over months 2 to 4
Skill scores and 95% conf. intervals (1000 samples)
Brier skill score: 0.328 (0.158, 0.451)
Reliability skill score: 0.982 (0.921, 0.987)
Resolution skill score: 0.346 (0.226, 0.474)



S4

(Some) Future ECMWF developments

- **Better atmosphere/ocean models**
  - Reduction of equatorial wind bias, plus other improvements. Evidence suggests higher resolution atmosphere will play a role.
  - Tropospheric aerosol variations
  - Higher resolution ocean
- **Land surface**
  - Full offline re-analysis of land surface initial conditions, esp snow
  - Fully consistent real-time initialization
  - Improvements: vegetation response, hydrology
- **Stratosphere**
  - Spectrally resolved UV radiation, to allow proper impact of solar variability
  - Increased vertical resolution, to allow better QBO dynamics
  - Better (post eruption) volcanic aerosol specification, better ozone
- **Sea-ice**
  - Actually having a model ….

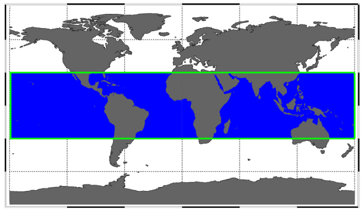
Multi-model approach

- **Operational multi-model system at ECMWF**
  - Called EUROSIP, initially ECMWF/Met Office/Meteo-France
  - NCEP have now joined
  - Others intending to join
- **Multi-model likes high quality models**
  - Automatically benefit
  - May be some issues if there is a mix of excellent models and poor ones
  - Ideally like long re-forecast set and skill estimates for each model
- **Past research has shown multi-model hard to beat**

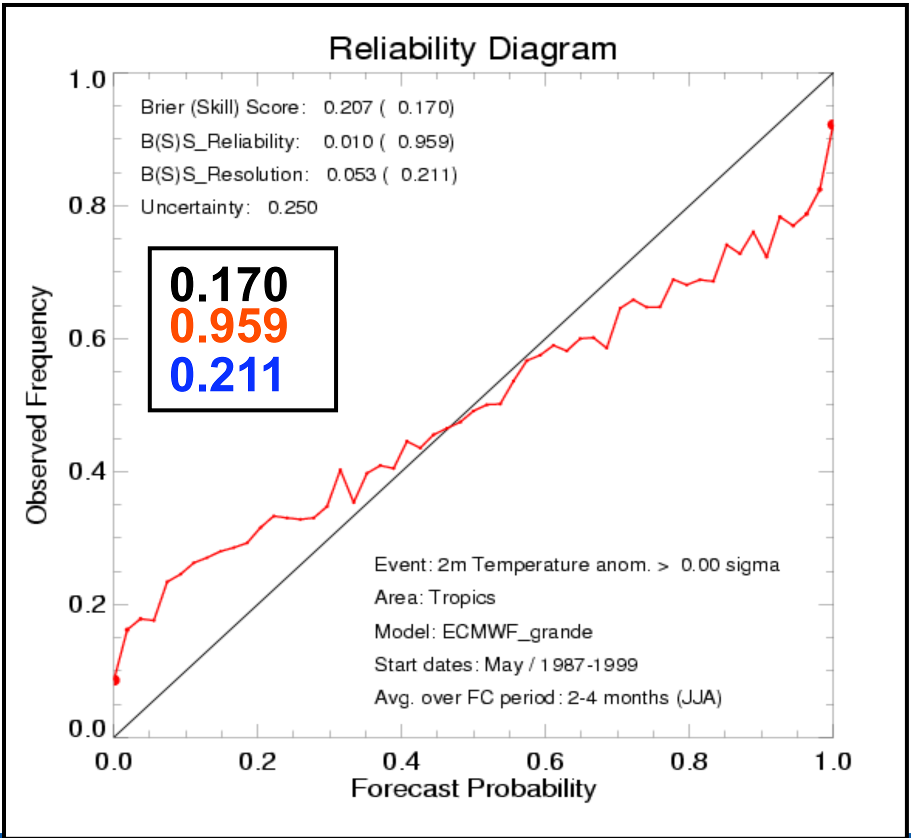
DEMETER: impact of ensemble size

BSS
Rel-Sc
Res-Sc

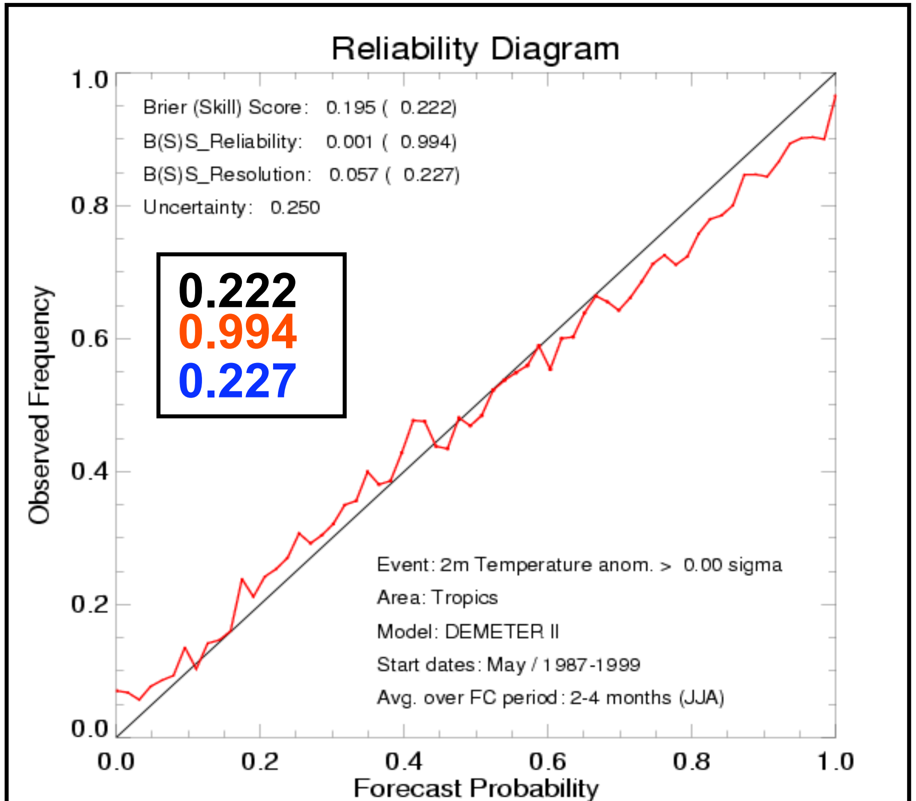
Reliability diagrams (T2m > 0)
1-month lead, start date May, 1987 - 1999



single-model [54 members]



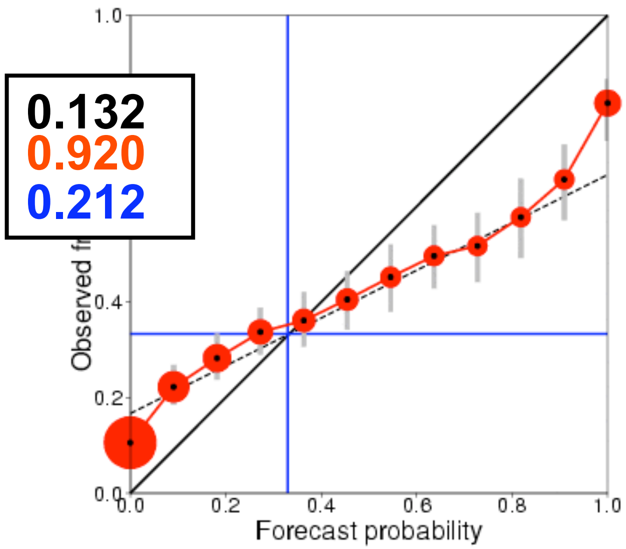
multi-model [54 members]



Hagedorn et al. (2005)

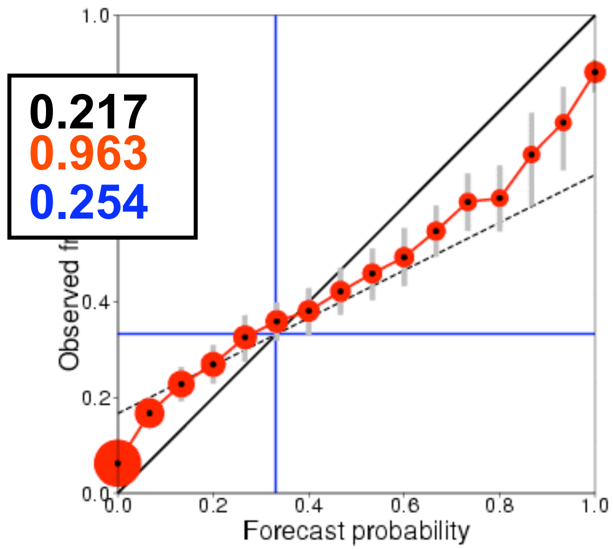
Cf benefit from model improvement

Reliability diagram for ECMWF with 11 ensemble members
Near-surface air temperature anomalies above the upper tercile
Accumulated over tropical band (land and sea points)
Hindcast period 1981-2010 with start in May average over months 2 to 4
Skill scores and 95% conf. intervals (1000 samples)
Brier skill score: 0.132 (0.026, 0.223)
Reliability skill score: 0.920 (0.875, 0.947)
Resolution skill score: 0.212 (0.149, 0.279)



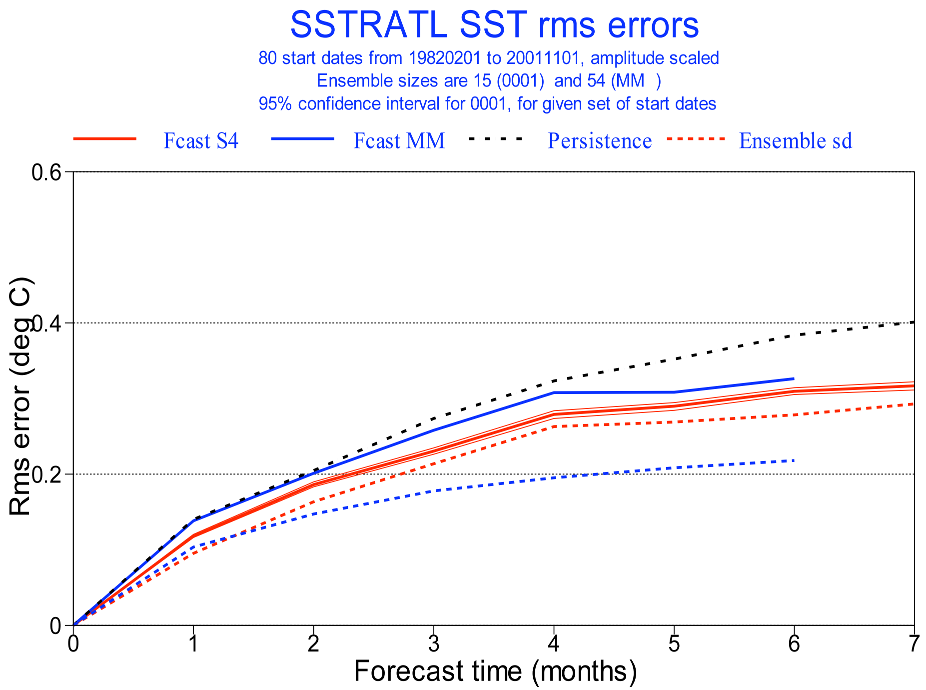
S3

Reliability diagram for ECMWF with 15 ensemble members
Near-surface air temperature anomalies above the upper tercile
Accumulated over tropical band (land and sea points)
Hindcast period 1981-2010 with start in May average over months 2 to 4
Skill scores and 95% conf. intervals (1000 samples)
Brier skill score: 0.217 (0.133, 0.296)
Reliability skill score: 0.963 (0.937, 0.975)
Resolution skill score: 0.254 (0.192, 0.324)



S4

Example: better model vs multi-model



DEMETER 6-model multi-model ensemble

ECMWF System 4

cf Stockdale et al, J. Clim 2006

Conclusions

- **Producing good forecasts is hard**
  - Models need to include relevant processes to a **high accuracy**
  - Models need to be **complete,** including all main sources of variability
- **Verifying forecasts is hard**
  - Large ensemble sizes needed to properly characterize pdfs
  - Limited number of events to look at modest shifts in pdfs
- **Multi-model forecasts are very useful**
  - They always give a sanity check
  - They can be combined to give more reliable and usually better forecasts
- **Keep up the work on the forecast systems …**
  - To produce most informative forecasts possible
  - Need to aim at being intrinsically reliable