Supplementary Material

Significance tests

We test for significance by computing the probability that any particular statistic (e.g. skill measure, RPC or difference in skill) could be accounted for by uncertainties arising from a finite ensemble size (*M*) and a finite number of validation points (*T*). This is achieved using a non-parametric block boot-strap approach [*Wilks*, 2006; *Goddard et al.*, 2012; *Smith et al.*, 2013], in which an additional 1000 hindcasts (or pairs of hindcasts when testing differences, for example before and after RPC correction) are created as follows:

1. Randomly sample with replacement *T* validation cases (over time). In order to take autocorrelation into account this is done in blocks of five consecutive cases (years) for the decadal hindcasts.

2. For each of these, randomly sample without replacement *M*-3 ensemble members. Replacement is not used over ensemble members because repeatedly resampling the same members reduces the number of independent data points in the ensemble mean, and so reduces the correlation unfairly. We use *M*-3 members to retain a large enough ensemble size to maintain representative estimates of the skill measure.

3. Compute the required statistic (e.g. skill measure, RPC or difference in skill) for the ensemble mean of the bootstrapped sample (or samples if testing differences).

4. Repeat from step (1) 1000 times to create a probability distribution (PDF) of the required statistic.

5. Obtain the significance level based on a two-tailed test against the null hypothesis. For example, the null hypothesis 'RPC is not different to one' is rejected at the 90% level if the 5 to 95% confidence interval obtained from the bootstrapped pdf distribution does not span one.

For RPC, MSSS and correlation this method is performed on individual time-series of grid points or area averages separately, while for the reliability diagram in Figure 2 it is performed by re-sampling entire fields (for the regions of interest). For example, for the correlation of the NAO in Figure 3a, this method leads to a 99% confidence interval of [0.20, 0.91] such that the null hypothesis that the correlation is not different to zero is rejected.

Table S1 Gridded observation datasets[a]

Variable	Dataset	References
SAT	Average from:	
	HadCRUT4	*Morice et al.* [2012]
	NASA	*Hansen et al.* [2010]
	NCDC	*Smith and Reynolds* [2005]
MSLP	HadSLP2	*Allan and Ansell* [2006]
PREC	GPCP[b]	*Adler et al.* [2003]
PREC	GPCC[c]	*Schneider et al.* [2011]

[a] Details of gridded observation datasets used.

[b] For assessing seasonal hindcasts.

[c] For assessing decadal hindcasts which precede the coverage of GPCP (noting that GPCC observations only cover land).

Table S2. General Circulation Models[a]

Modelling Centre	Model	Ensemble Size; start dates	References
Met Office Hadley Centre (seasonal forecasts)	GloSea5	24 members; around 1st Nov 1992 to 2011	*MacLachlan et al.* [2014]
Met Office Hadley Centre	DePreSys, HadCM3	37 members; 1st Nov 1960 to 2005	*Smith et al.* [2010; 2013]
Met Office Hadley Centre	DePreSys2, HadGEM3	4 members; 1st Nov 1960 to 2005	*Knight et al., submitted*[c] [2014]
Canadian Centre for Climate Modelling and Analysis [b]	CanCM4	10 members; 1st Jan 1961 to 2006	*Merryfield et al.* [2013]
NOAA Geophysical Fluid Dynamics Laboratory [b]	GFDL-CM2.1	10 members; 1st Jan 1961 to 2006	*Delworth et al.* [2006]
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology [b]	MIROC5	6 members; 1st Jan 1961 to 2006	*Watanabe et al.* [2010]
Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology) [b]	MPI-ESM-LR	3 members; 1st Jan 1961 to 2006	*Jungclaus et al.* [2006]

[a] Details of General Circulation Models used (decadal systems except for GloSea5).

[b] Model output from the Coupled Model Intercomparison Project Phase 5 (CMIP5) [*Taylor et al.*, 2012].

[c] *Knight et al.*, Predictions of climate several years ahead using an improved decadal prediction system, submitted to J. Clim., 2014.

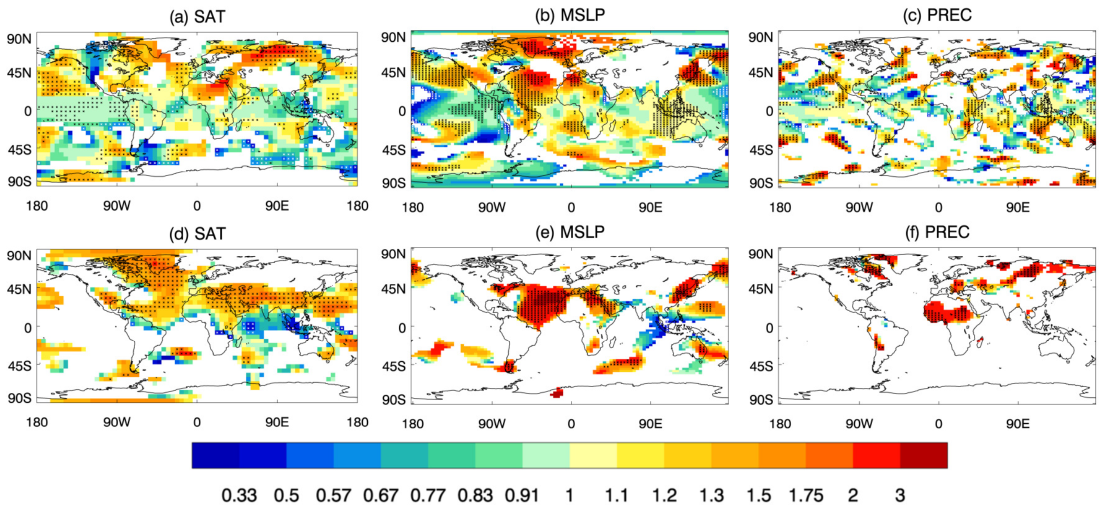

Figure S1: As Figure 1 but with the additional constraint that regions of correlation not *significantly* greater than zero are masked out, leading to the masking of regions with RPC below one that correspond to regions of insignificant skill.

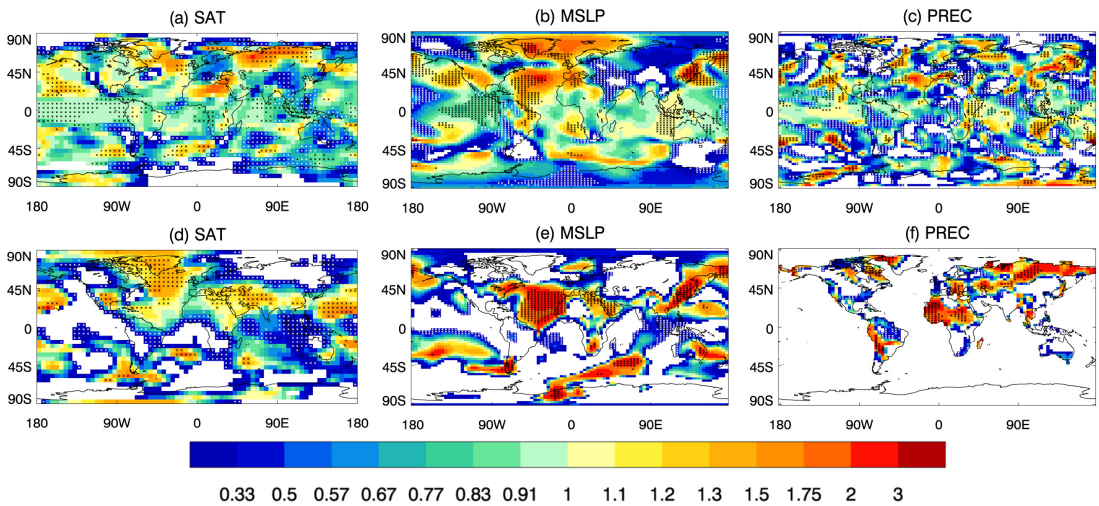

Figure S2: As Figure 1 but with all bias corrections applied in cross-validation mode (ignoring current year and, for decadal hindcasts the four years either side), leading to similar conclusions but with a slight reduction in the strength of the high RPC values as cross-validation is known to underestimate correlation [*Smith et al.*, 2013; *Gangsto et al.*, 2013].

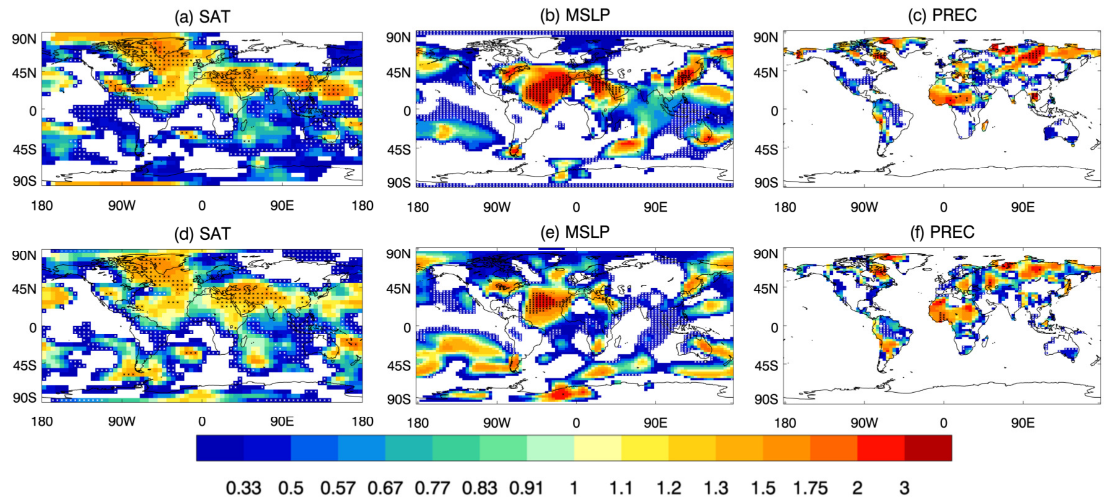

Figure S3: As Figure 1 but for HadCM3 only (row 1, 37 members) and for the four non-Met Office Hadley Centre CMIP5 (row 2, 29 members) to see the results for a single model versus that for the remaining models. This splitting of the models leads to the same conclusions, but with a slight reduction in the strength of the high RPC value, likely due to the reduced ensemble size.

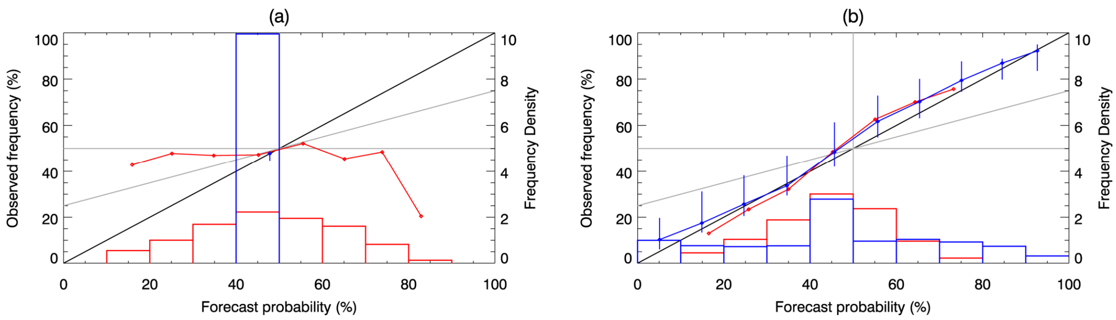

Figure S4: As Figure 2 but with all bias corrections applied in cross-validation mode (ignoring current year and four years either side).

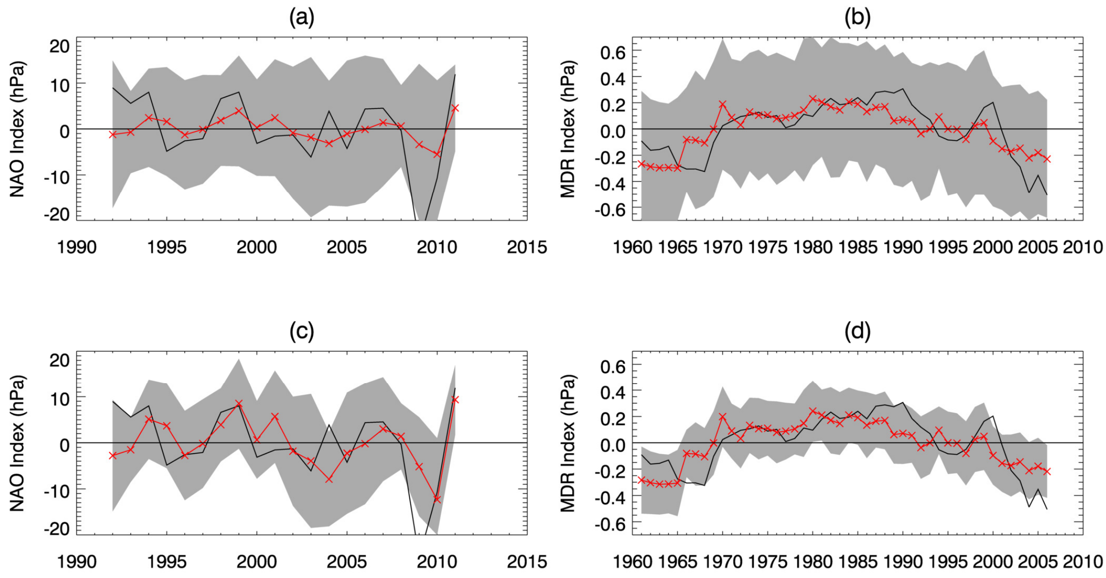

Figure S5: As Figure 3 but with all bias corrections applied in cross-validation mode (ignoring current year, and for decadal hindcasts the four years either side).

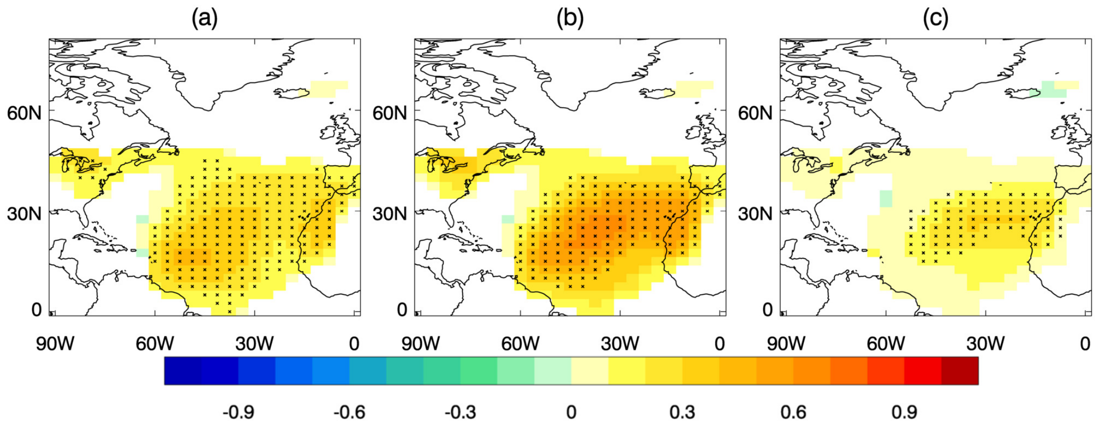

Figure S6: As Figure 4 but with the additional constraint that regions of correlation not *significantly* greater than zero are masked out as they imply zero skill.

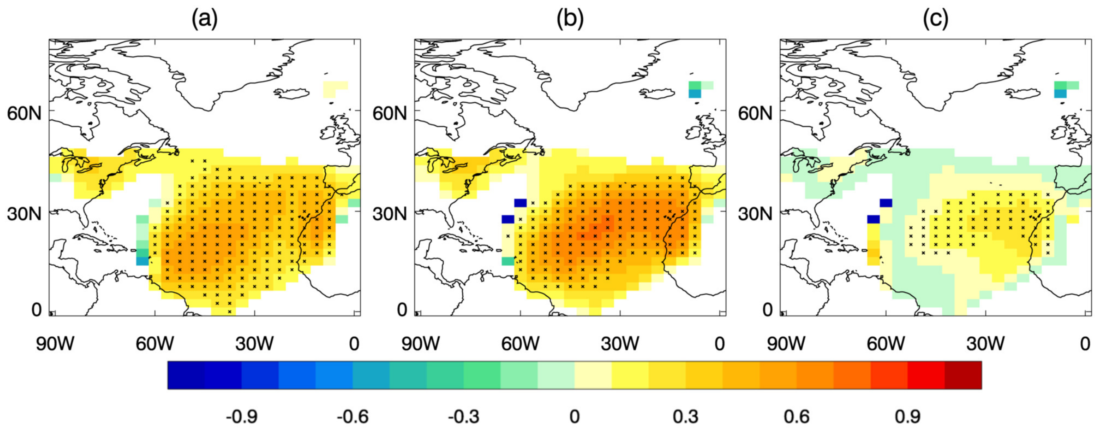

Figure S7: As Figure S6 but with all bias and variance corrections applied in cross-validation mode (ignoring current year and four years either side).

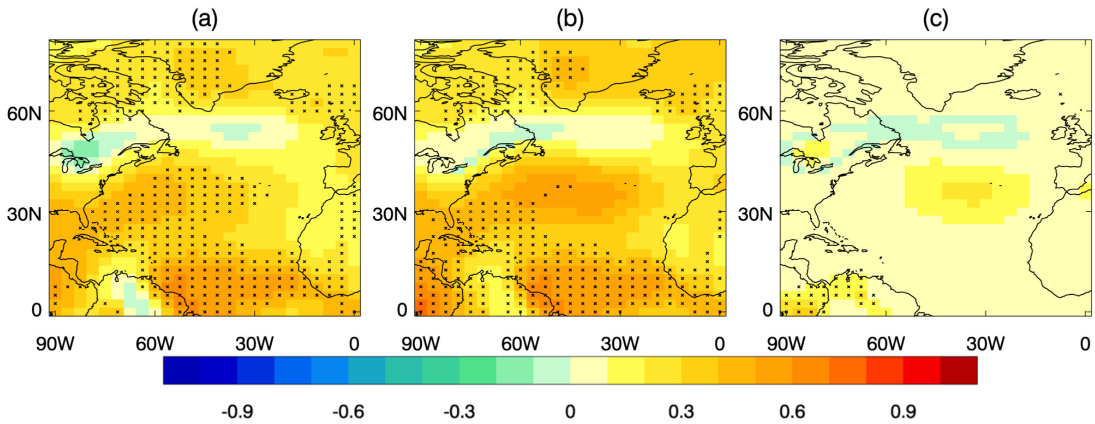

Figure S8: As Figure 4 but from the Met Office Hadley Centre seasonal forecasting system for DJF seasonal mean, showing a slight improvement after the variance correction but not significant (noting that there are only 20 years of model output from this system, while 46 years are analysed from the decadal systems).

References

Adler, R.F., G.J. Huffman, A. Chang, R. Ferraro, P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, P. Arkin (2003), The Version 2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979-Present). J. Hydrometeor., 4, 1147-1167.

Allan, R. J., and T. J. Ansell (2006), A new globally complete monthly historical mean sea level pressure data set (HadSLP2): 1850-2004, J. Clim., 19(22), 5816-5842.

Delworth, T. L., and Coauthors (2006), GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics. J. Clim., 19, 643–674, doi:10.1175/JCLI3629.1.

Gangsto, R., A. P. Weigel, M.A. Liniger, and C. Appenzeller (2013), Methodological aspects of the validation of decadal predictions, Clim. Res., 55, pp. 181-200, doi:10.3354/cr01135.

Goddard, L. et al. (2012), A Verification Framework for Interannual-to-Decadal Predictions Experiments, Clim. Dyn., 40, 245-272, doi:10.1007/s00382-012-1481-2.

Hansen, J., R. Ruedy, M. Sato, and K. Lo (2010), Global surface temperature change, Rev. Geophys., 48, RG4004, doi:10.1029/2010RG000345.

Jungclaus, J. H., and Coauthors (2006), Ocean circulation and tropical variability in the coupled model ECHAM5/MPI-OM, J. Clim., 19, 3952–3972, doi:10.1175/JCLI3827.1.

MacLachlan, C., and Coauthors (2014), Global Seasonal Forecast System version 5 (GloSea5): a high resolution seasonal forecast system, Q. J. R. Met. Soc., accepted online version, doi:10.1002/qj.2396

Merryfield, W. J., and Coauthors (2013), The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization, Mon. Wea. Rev. e-View, 141, 2910–2945, doi:10.1175/MWR-D-12-00216.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset, J. Geophys. Res., 117, D08101, doi:10.1029/2011JD017187.

Schneider, U., A. Becker, A. Meyer-Christoffer, M. Ziese, and B. Rudolf (2011), Global precipitation analysis products of the GPCC. Global Precipitation Climatology Centre (GPCC), DWD, Internet Publikation, pp. 1–13.

Smith, T. M., and R. W. Reynolds (2005), A global merged land air and sea surface temperature reconstruction based on historical observations (1880-1997), J. Clim., 18, 2021-2036, doi:10.1175/JCLI3362.1.

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife (2010), Skilful multi-year predictions of Atlantic hurricane frequency, Nat Geosci., 3, 846-849, doi: 10.1038/NGEO1004.

Smith, D. M., R. Eade, and H. Pohlmann (2013), A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction, Clim. Dyn., 41, 3325-3338, doi:10.1007/s00382-013-1683-2.

Taylor K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, Bull. Am. Met. Soc., 93(4), 485 - 498 , doi:10.1175/BAMS-D-11-00094.1.

Watanabe, M., et al. (2010), Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity, J. Clim., 23, 23, 6312–6335, doi:10.1175/2010JCLI3679.1.

Wilks, D. S. (2006), Statistical Methods in the Atmospheric Sciences, 2nd ed., Academic, San Diego, Calif.