Clim Dyn manuscript No.

Evaluating the potential for statistical decadal predictions of sea surface temperatures with a perfect model approach

Ed Hawkins · Jon Robson · Rowan Sutton · Doug Smith · Noel Keenlyside

Resubmitted to Climate Dynamics, 23rd November 2010

Abstract We explore the potential for making statistical decadal predictions of sea surface temperatures (SSTs) in a perfect model analysis, with a focus on the Atlantic basin. Various statistical methods (Lagged correlations, Linear Inverse Modelling and Constructed Analogue) are found to have significant skill in predicting the internal variability of Atlantic SSTs for up to a decade ahead in control integrations of two different global climate models (GCMs), namely HadCM3 and HadGEM1. Statistical methods which consider non-local information tend to perform best, but which is the most successful statistical method depends on the region considered, GCM data used and prediction lead time. However, the Constructed Analogue method tends to have the highest skill at longer lead times. Importantly, the regions of greatest prediction skill can be very different to regions identified as potentially predictable from variance explained arguments. This finding suggests that significant local decadal variability is not necessarily a prerequisite for skillful decadal predictions, and that the statistical methods are capturing some of the dynamics of low-frequency SST evolution. In particular, using data from HadGEM1, significant skill at lead times of 6-10 years is found in the tropical North Atlantic, a region with relatively little decadal variability compared to interannual variability. This skill appears to come from reconstructing the SSTs in the far north Atlantic, suggesting that the more northern latitudes are optimal for SST observations to improve predictions. We additionally explore whether adding sub-surface temperature data improves these decadal statistical predictions, and find that, again, it depends on the region, prediction lead time and GCM data used. Overall, we argue that the estimated prediction skill motivates the further de-

E. Hawkins · J. Robson · R. Sutton
NCAS-Climate, Department of Meteorology, University of Reading, UK
E-mail: e.hawkins@reading.ac.uk

D. Smith
Met Office Hadley Centre, Exeter, UK

N. Keenlyside
IFM-GEOMAR, Kiel, Germany

velopment of statistical decadal predictions of SSTs as a benchmark for current and future GCM-based decadal climate predictions.

Keywords decadal prediction · Atlantic · SSTs · statistical methods · predictability

1 Introduction

Policymakers require quantitative regional climate predictions for the near-term (out to 30 years) to aid efforts in adapting to a changing climate, for example, for planning new infrastructure (e.g. Arnell and Delaney 2006) or for the insurance industry (e.g. Michaels et al. 1997). Changes in climate over these timescales occur for two reasons: the response of the climate to both historical and future radiative forcings, and because of internal climate variability. Until recently, climate projections have generally only considered the trends due to the radiative forcing component and the spread of natural variability (Meehl et al. 2007). However, the *phase* of the internal variability component is particularly important for near-term predictions on regional spatial scales (e.g. Hawkins and Sutton 2009b). The potential to improve near-term climate forecasts by predicting the internal variability, as well as the radiatively forced component, has led to the development of global climate model (GCM) based 'decadal prediction systems' (e.g. Smith et al. 2007, Keenlyside et al. 2008). Analysing the skill of these types of predictions will be a major part of the next IPCC assessment, although many challenges remain (Meehl et al. 2009).

The development of GCM based decadal predictions is complex, challenging, and computationally expensive, so it seems prudent to assess the skill of the GCM predictions by using benchmarks of simpler, statistically based alternatives. This type of assessment has been a useful feature of seasonal predictions for many years (e.g. Penland and Magorian 1993; Barnston et al. 1994; Colman and Davey 2003; Saha et al. 2006), and has begun for decadal timescales. Lee et al. (2006) considered a Bayesian method for making decadal predictions using ensemble simulations of GCMs and demonstrated some skill in predicting climate trends. Laepple et al. (2008) improved near-term predictions of global mean temperature using the CMIP3 projections by simply adjusting for the internal variability at the start of the forecast, but this still requires the CMIP3 projections to predict the future trend. Ideally we would like to develop a benchmark prediction system which is independent of any GCM-based information. Lean and Rind (2009) recently developed a statistical decadal prediction of the global temperature response to radiative forcings to allow a comparison with the GCM estimates. However, none of these papers attempted to predict the internal variability component of temperatures for the coming decade.

The presence of considerable decadal and multi-decadal variability in historical observations of Atlantic sea surface temperatures (SSTs) (e.g. Delworth and Mann 2000) gives rise to the potential for more skillful predictions and motivates analysing this region in detail. This type of variability is also thought to be potentially predictable in GCMs (e.g. Boer 2004; Collins et al. 2006; Boer and Lambert 2008). Additionally, there are thought to be regional temperature and precipitation impacts over land of these long-term Atlantic SST changes (e.g. Folland et al. 1986; Sutton and Hodson 2005), which may then also be potentially predictable. The number and intensity of Atlantic hurricanes are also sensitive to the SSTs in the tropical regions (e.g. Goldenberg et al. 2001; Emanuel 2005; Vecchi et al. 2008). Recently, Smith et al. (2010) demonstrated skill in predicting hurricane numbers on seasonal to multi-year timescales using a GCM based decadal prediction system.

In this paper we consider the first stage of building an operational statistical decadal prediction system by analysing GCM output in a ‘perfect model’ framework, i.e. using model data as substitute observations, and trying to predict the subsequent model data. This type of analysis could be considered as a best case scenario, with perfect availability of observations and a perfect model, although this assumes that the GCM has a realistic representation of the internal variability of the real ocean. The robustness of the results can be considered by analysing more than one GCM and by comparing this idealised skill with operational predictions using the same GCM. We further simplify the problem by analysing control simulations where there are no complicating external forcing factors. This type of approach has been regularly used in GCM based predictability studies (e.g. Griffies and Bryan 1997; Collins and Sinha 2003; Collins et al. 2006; Dunstone and Smith 2010) to explore the capability of GCMs to make decadal predictions, and we consider it valuable to assess the potential skill of various statistical methods before starting to analyse the real observations which are complicated by the forced trends and lack of complete observational coverage. Once the potential predictive skill is established then the methods can be subsequently applied to the observational records.

This paper is structured as follows. In Section 2 we discuss the data used and consider the potential predictability in two GCMs and the historical observations, and Section 3 describes the statistical methods we use. Section 4 discusses the skill of the predictions and considers the benefit of utilising sub-surface data as well as SSTs to improve the predictions. We conclude and discuss the findings in Section 5, including a brief comparison with existing operational decadal predictions.

2 SST data and potential predictability

We consider historical observations and data from two GCMs to explore the predictability of SSTs, with a focus on the Atlantic Ocean.

2.1 HadCM3, HadGEM1 and HadISST

Firstly, we use 1000 years of a long, stable control integration of the third version of the Hadley Centre climate model (HadCM3, Gordon et al. 2000) with constant pre-industrial radiative forcings. HadCM3 is a global coupled ocean-atmosphere model with an atmospheric resolution of $2.5^\circ \times 3.75^\circ$ and 19 vertical levels. The ocean component has a resolution of $1.25^\circ \times 1.25^\circ$ with 20 vertical levels.

We also use an 846 year control integration of the first version of the Hadley Centre Global Environmental Model (HadGEM1, Johns et al. 2006), but we remove the first 300 years to avoid spin-up issues. The atmospheric component has a resolution of $1.25^\circ \times 1.875^\circ$ with 38 layers in the vertical. The ocean component has 40 levels in the vertical with a zonal resolution of 1° everywhere and a meridional resolution of 1° between the poles and 30° latitude, increasing to $\frac{1}{3}^\circ$ at the equator.

The reason for choosing these models is that HadCM3 is also the base model used in the UK Met Office's decadal prediction system (DePreSys; Smith et al. 2007) for which a direct comparison of skill can be made. HadGEM1 has also been recently updated to run at a higher horizontal resolution (HiGEM; Shaffrey et al. 2009), and HiGEM is being used to make decadal predictions as part of the forthcoming CMIP5, allowing a future comparison of a similar model. Unfortunately, the HiGEM control integration is not long enough to allow the analysis presented here to be performed.

The HadISST dataset (Rayner et al. 2003) includes estimates of observed monthly mean interpolated SSTs, available on a $1^\circ \times 1^\circ$ grid for 140 years from 1870-2009. In all that follows we consider annual means of SSTs from both GCMs and observations.

2.2 Comparing potential predictability in the GCMs and observations

Before attempting to make decadal predictions of SSTs in these models, it is enlightening to consider the potential predictability (e.g., Boer 2004; Boer and Lambert 2008) of these two GCMs, for different length time means, defined as,

$$\text{potential predictability} = \frac{\sigma_N}{\sigma_1}, \tag{1}$$

where σ_N represents the standard deviation of N-year means of SST. This measure of potential predictability essentially assumes that much of the interannual variability is chaotic and unpredictable, whereas the longer timescale variability relies on slower ocean dynamics and is therefore more predictable. A threshold for 'useful' potential predictability is hard to define, as it is likely to be purpose and situation dependent.

The left hand column of Fig. 1 shows the interannual standard deviation (σ_1) of SSTs in the GCMs and observations. The remaining columns illustrate the potential predictability for $N = 3, 5$ and 10 years.

Both GCMs (top two rows) identify the far North Atlantic in general, and the North Atlantic Current (NAC) region in particular, as regions with high potential predictability. There are also encouraging signals in the south tropical Atlantic. A similar analysis for the HadISST observations is complicated by the trends due to historical radiative forcings. The bottom row in Fig. 1 shows the variability and potential predictability for the annual mean observations, detrended at each grid point using a cubic spline with two breakpoints. This qualitative effort at removing the long-term trend will give approximate estimates of the variability and potential predictability.

Comparing the observed and model estimates suggests that the GCMs overestimate the interannual variability in SSTs, especially in the far North Atlantic region. However, Minobe and Maeda (2005) conclude that HadISST under-represents the magnitude of the variability, especially in the NAC region, so this appearance should not be regarded as a robust finding. Additionally, the processing of point observations into a gridded data set (Rayner et al. 2003) will mean that the comparison with GCM data is not really comparing like with like. We also note that HadGEM1 has a far weaker ENSO than found in the observations (Johns et al. 2006; also see Supplementary Fig. S1 which shows the global version of Fig. 1). Nevertheless, the potential predictability estimates for the GCMs have comparable magnitudes to HadISST.

2.3 Interpretation of potential predictability

It is often suggested that the existence of this type of potential predictability in a particular region is a necessary prerequisite for actual predictability in that region. Later, we will compare the potential predictability with the actual skill of the prediction methods developed and demonstrate that this is not always true.

We also note here that if the SSTs in a certain region evolved randomly (i.e. white noise) then the potential predictability from Eqn. 1 would be $1/\sqrt{N}$. In many regions, and for different values of N, the potential predictability in both the observations and

GCMs is *less* than this value. This finding implies that the SSTs in these regions have negative auto-correlations on timescales less than N years, due to the impact of El Nino for instance. However, the existence of these anti-correlations might also be considered as potential predictability which is not captured by the simple definition of Eqn. 1.

3 Empirical prediction methods

3.1 Rationale

The end goal of this research is to create decadal predictions of SSTs, based solely on observational data, to act as a 'benchmark' with which to compare and contrast results from GCM predictions. The aim of the present paper is to take the first step towards this goal by testing different statistical methods using data from a GCM. The experimental design is chosen to match the availability of historical data, but in the 'perfect' GCM framework. However, note especially that any regions identified as having significant skill in the GCMs may not be the same as those found when analysing observations. For checking the robustness of the findings we consider data from two different GCMs. Analysis of the observed SSTs in some high latitude regions is impossible because there are no SST observations in regions covered by sea ice. For most of the GCM analysis that follows we do not consider latitudes polewards of 66°N to roughly simulate the observational coverage.

We will compare different statistical methods for making predictions of SSTs. For each individual prediction we consider 140 years of SST anomalies from the control integration as training data, i.e. the same length as the historical observations. We then make a statistical prediction of the next 10 years and compare to the actual SSTs in the control integration, i.e. the 'perfect model' approach. We then repeat the analysis by shifting the training data to start 10 years later in the control integration. Thus the different training sets are pseudo-independent, allowing us to assess the cross-validated skill of the predictions as the forecast data is not used in the training of the statistical model. For the HadCM3 (HadGEM1) data this allows 85 (40) separate forecasts, allowing the average skill statistics to be calculated.

Although we make a prediction for each lead time of 1-10 years, we consider the prediction skill using four different lead times and averaging periods, namely, year 1, year 2, the mean of years 3-5 and the mean of years 6-10.

3.2 Climatology, persistence and lagged correlations

The simplest reference statistical methods that we consider are,

$$\text{Climatology}: \quad x(t_0 + \tau) = 0, \tag{2}$$

$$\text{Persistence}: \quad x(t_0 + \tau) = x(t_0), \tag{3}$$

$$\text{Lagged correlation}: \quad x(t_0 + \tau) = \beta(\tau)x(t_0), \tag{4}$$

where x is the SST anomaly at a particular grid point, t_0 is the start year for the forecast, τ is the forecast lead time from 1-10 years, and $\beta(\tau)$ is the auto-correlation of x at a lag of τ, estimated from the training data. In these methods, each grid point is treated independently.

The lagged correlation forecast is a version of 'damped persistence' (DP; Lorenz 1973) which damps the anomalies (x) at each grid point towards zero, although here β can also be negative. Lorenz (1973) showed that DP is as good as, or better, than both a climatological forecast ($\beta = 0$) and persistence ($\beta = 1$), as defined above, assuming the underlying auto-correlation structure can be reliably estimated. If only short periods of observations exist then 'persistence' may be the only option for a prediction.

3.3 Linear Inverse Modelling

The above methods utilise information from only one grid point to make a forecast, whereas it may be possible to add skill by including covariance information from other grid points. Linear Inverse Modelling (LIM) is one way of doing this and is currently being used in experimental forecasts of tropical SSTs on seasonal[1] (Penland and Magorian 1993) and annual (Newman et al. 2010) timescales. Here we extend this approach to decadal timescales.

The LIM method models the evolution of SSTs as a linear dynamical system, forced by (atmospheric) white noise. The number of degrees of freedom are reduced by considering only the leading modes of variability; we use empirical orthogonal functions (EOFs) as the spatial fields and their respective principal components (PCs) as the time variation of these spatial fields[2]. The technical steps required are given in Penland and Magorian (1993) and here we give brief details:

1. We estimate the leading correlation EOFs and corresponding PCs of the SST data, weighted to ensure each unit area of the sea surface is treated equally. In the analysis

[1] These forecasts are available at: http://www.esrl.noaa.gov/psd/forecasts/sstlim/

[2] The leading EOFs of the two GCMs are compared to the observations in the Supplementary Information.

presented here we use this methodology on the entire global domain, and separately using just the Atlantic region. The EOFs are estimated for each training period individually, to ensure that we are cross-validating our predictions appropriately.

2. The evolution of the leading PCs is modelled as a linear dynamical system with (atmospheric) stochastic forcing (ξ),
$$\frac{d\mathbf{x}}{dt} = \mathbf{B}\mathbf{x} + \xi, \tag{5}$$
where $\mathbf{x}$ is the PC state vector and the matrix $\mathbf{B}$ defines the temporal evolution of the state vector. This model can then be used to make forecasts of $\mathbf{x}$ (denoted by $\widehat{\mathbf{x}}$), with lead time τ,
$$\widehat{\mathbf{x}}(t+\tau) = \mathbf{P}(\tau)\mathbf{x}(t), \tag{6}$$
where the forecast propagator ($\mathbf{P}$) is derived from,
$$\mathbf{P}(\tau) = \exp(\tau\mathbf{B}) = \exp\left(\frac{\tau}{\tau_0}\ln\left[\frac{\mathbf{C}(\tau_0)}{\mathbf{C}(0)}\right]\right), \tag{7}$$
where
$$\mathbf{C}(\tau_0) = \langle \mathbf{x}(t+\tau_0)\,\mathbf{x}^{\mathrm{T}}(t)\rangle, \tag{8}$$
$$\mathbf{C}(0) = \langle \mathbf{x}(t)\,\mathbf{x}^{\mathrm{T}}(t)\rangle, \tag{9}$$
where $\langle\cdot\rangle$ denotes an average over all t, and τ_0 is the lead time over which the propagator is estimated. We choose $\tau_0 = 1$ year, and select the largest number of EOFs which ensure that $\mathbf{B}$ remains real. Typically this is around 7 (12) EOFs, explaining $\sim 40\%$ ($\sim 62\%$) of the variance for the global (Atlantic) domain in HadCM3, and similarly, 9 (12) EOFs, explaining $\sim 47\%$ ($\sim 68\%$) of the variance for HadGEM1.
3. The spatial pattern of the forecast can be recovered by multiplying $\widehat{\mathbf{x}}$ by the respective EOF patterns.

3.4 Constructed Analogue

The use of historical matches (or ‘analogues’) for forecasting has a long history (see van den Dool 2007 for a recent review). The natural analogue (NA) method looks for particular times in the historical record which are ‘closest’ to the desired forecast start state. The subsequent evolution in the historical records would then be the analogue forecast. van den Dool (1994) described how unlikely finding a good natural analogue might be[3], so suggested

[3] van den Dool (1994) showed that, for the atmospheric flow at 500mb to match over an entire hemisphere, one would need a library of 10^{30} years to find an analogue within the observational errors. Making a similar estimate, with optimistic assumptions, for Atlantic SST would require more than 10^5 years of data to provide a natural analogue. Given 140 years of data, the chance of finding a natural analogue is remote.

the constructed analogue (CA) method, which has successfully been applied to seasonal SST forecasts in the Pacific (Barnston et al. 1994) and soil moisture forecasts in the US (van den Dool et al. 2003). The essential ingredient of this method is constructing an analogue for the recent past from a weighted, linear combination of historical data. The assumptions made are of linearity, and that the future evolution will be similar to the historical evolution for similar situations.

Here we use a simple CA methodology to forecast SSTs at lead times (τ) of 1-10 years by reconstructing the spatial pattern of the last year of our training data, denoted as $\mathbf{u}_t$. Thus, if $\mathbf{u}_i$ represents SST anomalies from year i in the control integration, then an analogue can be constructed by minimising,

$$J = \sum_{M} \left(\mathbf{u}_t - \sum_{i=t-L+1}^{t-\tau} w_i \mathbf{u}_i \right)^2 , \tag{10}$$

over the M gridpoints by choosing the weights, w_i, and where L is the length of the training data (here $L = 140$ years). When minimising J it is also necessary to ensure equal weight for each unit area of the domain. The forecast of SST, $\widehat{\mathbf{u}}$, at a lead time of τ years is then,

$$\widehat{\mathbf{u}}_{t+\tau} = \sum_{i=t-L+1}^{t-\tau} w_i \mathbf{u}_{i+\tau}, \tag{11}$$

where the procedure is repeated for different values of τ. Like the LIM method we perform the analysis for a global domain and the Atlantic basin separately. For HadCM3 (HadGEM1) the SST for the forecast start time is reconstructed with a mean correlation across the forecasts of 0.79 (0.77) for the global domain and 0.83 (0.84) for the Atlantic domain.

4 Predictive skill of the statistical methods

We now consider the skill of our various statistical methods using anomaly correlation and root mean square (RMS) error, and compare the results to potential predictability estimates. The reason for choosing these two skill measures is because they are simple, commonly used for assessing skill in predictions, and measure different aspects of the prediction which may be useful in different situations.

4.1 Anomaly correlation

Figs. 2, 3 show the anomaly correlation skill for some chosen prediction types for the HadCM3 and HadGEM1 GCMs respectively, focussing on the Atlantic domain. The global maps of skill for all prediction types are shown in the Supplementary Information. Assuming independent predictions, the 95% confidence level is around $r \approx 0.2$ (0.3) for HadCM3 (HadGEM1), so most of the coloured regions are significant. Generally, the statistical methods employed are superior to persistence forecasts, especially for lead times greater than 5 years. We note that, in some regions, the skill appears to be increasing with lead time, but this is mostly due to the averaging period also increasing, and hence increasing the signal-to-noise.

There is considerable skill for many regions for 1 year ahead (left columns) in both GCMs, especially in higher latitudes. By year 2, the skill has dropped markedly, with little skill outside the far north Atlantic, although the Atlantic-only LIM and CA methods are performing well in HadCM3 in this region. For years 3-5, the skill has dropped further, though it remains close to zero rather than significantly negative. The largest grid point skill remains in the Atlantic-only LIM and CA methods for HadCM3, especially in the north-east Atlantic. There is very little skill anywhere in the South Atlantic beyond 1 year.

For years 6-10 the skill has declined further in the HadCM3 regions where significant skill was found at shorter lead times. However, the largest skill ($r \approx 0.4 - 0.5$) is in the northern tropical Atlantic for HadGEM1 using the CA method, although the LIM method also has some skill in this region. This is particularly interesting given the lack of skill in this region at shorter lead times. This region is important for the development of Atlantic hurricanes, and may suggest some long term predictability in this region when averaged over longer time periods, i.e. 5 years rather than 3, even though there is little potential predictability for this region in HadGEM1 (Fig. 1). However, this finding is not consistent between the GCMs. In HadCM3, the CA method only produces significant skill in the North Atlantic Current region, in agreement with potential predictability estimates.

Table 1 shows anomaly correlations for specific regional averages inside the Atlantic domain, with the best performing methods highlighted in bold for each GCM and lead time. Lagged correlations tend to perform best for shorter lead times, and CA for longer lead times. LIM sometimes has the highest skill in year 2 or years 3-5. Particularly encouraging is the skill for predicting mean North Atlantic SSTs, with $r \sim 0.5$ at a lead time of 6-10 years using CA in both GCMs. The Atlantic-only LIM tends to outperform the global LIM estimates for both GCMs, presumably because more SST variance is being retained inside the Atlantic domain for the Atlantic-only case.

4.2 RMS error

We now consider an alternative measure of skill, namely RMS error. We note here that the LIM method ‘damps’ anomalies towards zero, and is therefore unlikely to give a worse forecast than climatology. The CA method has no such natural damping, and is likely to perform poorly on this measure. However, methods exist to apply artificial damping to forecasts to minimise RMS error (e.g. Jewson and Hawkins 2009), although various assumptions are needed to derive the optimal damping factors which would vary for each forecast lead time and region.

Here we consider a very simple approach to reduce the RMS error in the CA forecasts. In the RMS errors shown below we use a CA forecast that has been damped by dividing each forecast (after lead time averaging) by a factor of $\sqrt{2}$, which improves the RMS skill virtually everywhere (not shown) but leaves the correlation unchanged[4].

Figs. 4, 5 show the RMS error for the HadCM3 and HadGEM1 GCMs respectively, relative to the RMS error of a climatological forecast for N year means,

$$\mathrm{RMS}_{\mathrm{relative}} = \frac{\mathrm{RMS}_{\mathrm{pred}}}{\mathrm{RMS}_{\mathrm{clim}}} = \frac{\mathrm{RMS}_{\mathrm{pred}}}{\sigma_N}. \tag{12}$$

Thus, the blue regions indicate where the prediction is performing better than climatology, and the red areas indicate where the performance is worse. The global maps of skill for all prediction types are shown in the Supplementary Information.

For year 1, the best method in both models is the lagged correlations, with less skill seen for the LIM and CA methods. For year 2 and years 3-5 there is little skill in any method in either model, although lagged correlations and Atlantic LIM do not produce worse forecasts than climatology. The CA method produces skill in certain regions, matching the skill found using anomaly correlations, but is also worse than climatology in many regions. However, in years 6-10, the enhanced skill in the tropical north Atlantic in HadGEM1 reappears as a significantly lower RMS error than a climatological forecast, again suggesting some long term predictability for this important region. However, outside this region the RMS error is worse than climatology.

It is clearly important to consider more than one measure of prediction skill when analysing prediction systems as these two metrics have produced different estimates of which method works best, and the choice of which to use will be situation dependent.

[4] This choice of artificial damping is motivated by noting that a forecast time series with zero correlation skill, but with the same variance as the true time series will produce an RMS error that is a factor of $\sqrt{2}$ larger than a climatological forecast, assuming Gaussian distributions. It would be possible to derive optimal damping factors if the mean RMS error was known before the forecasts are made, but in this case, it is not known.

Although the lagged correlations predictions tend to perform best for shorter lead times, the methods which use non-local information tend to perform better for longer lead times, suggesting that they are predicting some of the dynamical evolution of the SSTs.

The tropical north Atlantic skill in HadGEM1 at long lead times is an intriguing region of skill to explain, as this region has low potential predictability (Fig. 1). Interestingly, the skill is larger for the mean of years 6-10 than the mean of years 1-5 (not shown) suggesting that the skill comes from a non-local source. To more convincingly demonstrate the non-local mechanisms, a series of data withholding experiments were performed. Fig. 6 shows how the correlation skill for HadGEM1 for a lead time of 6-10 years changes as different regions are masked out of the construction of the statistical model[5] - a far North Atlantic region (northwards of 47°N) and a Gulf Stream region (GSR). These regions are chosen as they are significantly correlated with the tropical north Atlantic at 6-10 year lead times (not shown).

For the Atlantic LIM method (left column), removing the far North Atlantic region completely removes the skill from the tropical north Atlantic. For the Atlantic CA method (right column), both regions seem important, and removing each in turn reduces the skill, which again disappears completely when both regions are removed from the domain. A final test, removing the tropical region itself (bottom row), shows that even when the local tropical data is not used in the construction of the statistical model, there is still significant skill in the tropical North Atlantic region predictions, but the far north Atlantic is unaffected.

The skill that derives from within the Atlantic therefore appears to be non-local to the tropical north Atlantic, and instead relies on reconstructing the far north Atlantic, especially the convection regions. The mechanism for propagation could be related to changes in ocean deep convection leading to wave propagation along the boundaries from the north Atlantic to the tropics (e.g. Johnson and Marshall 2002), but Hodson and Sutton (in prep.) suggest that these wave anomalies remain sub-surface in HadGEM1 and that, instead, there is an atmospheric teleconnection to the tropical Atlantic.

It is also interesting to note that the global CA method performs better than the Atlantic CA method at long lead times in HadGEM1 for the tropical north Atlantic region (Table 1, Fig. S4), suggesting that some skill comes from outside the Atlantic

[5] For the LIM method, the estimated EOFs can be extended into the masked regions through regression onto the SSTs, allowing a prediction to subsequently be made for all regions. For the CA method, the estimation of the weights in Eqn. 10 does not include the masked regions, but Eqn. 11 can use all the data to make a prediction.

basin. An analysis of lagged correlations between SSTs in the tropical north Atlantic and global SSTs (not shown) indicates that this additional skill emanates from reconstructing the Indian Ocean SSTs, which leads the tropical Atlantic SSTs, though this relationship does not appear stationary and should be treated with caution. Possible mechanisms for this communication might rely on an indirect influence of the Agulhas current on the tropical north Atlantic (e.g. Biastoch et al. 2008; Haarsma et al. 2010) or an atmospheric teleconnection. However, it is essential to reiterate that this finding of enhanced skill in the tropical North Atlantic is only found in HadGEM1 and may not apply to observations.

An important conclusion of this analysis is that the regions identified as potentially predictable in Fig. 1 are not necessarily the same regions which will show enhanced predictive skill in forecasts. As described above, this seems to be because the skill comes from knowledge non-local to where the predictions are validated and the statistical methods are capturing some of the dynamics of low-frequency SST evolution. Thus, potential predictability estimates may under- or over-estimate the ability to actually predict, and prediction studies are required to reliably estimate actual predictability. Finally, these findings indicate that the far North Atlantic regions may be optimal for targeted observations to improve predictions, as found in previous studies (e.g. Tziperman et al. 2008; Hawkins and Sutton 2009a).

4.4 Effects of additional sub-surface observations

The historical ocean observational record also consists of sub-surface observations, though significant amounts of reliable measurements only exist for roughly the last 50 years. Dunstone and Smith (2010) recently performed a set of idealised GCM-based prediction experiments and found that using sub-surface data to initialise the HadCM3 GCM led to improved predictions compared to predictions which only used SSTs to initialise the ocean. We now consider whether the inclusion of such sub-surface information and/or different quantities of SST data would improve the statistical forecasts described above.

Fig. 7 compares the correlation skill of the LIM approach using the Atlantic domain in HadCM3 using varying amounts of data (Table 2). The top row (Short SST) shows the skill when using just 50 years of SSTs for training, and the second row repeats the previous analysis for 140 years of SST training data (Long SST). Unsurprisingly, there is a clear increase in skill when more data is used in the construction of the statistical model. The skill grows as longer amounts of training SST data are used (not shown), and the extreme case of using all the SST data is shown in the third row (All SST). Note however that the 'All SST' case is not strictly cross-validated as the forecast data is used in the construction of the EOFs.

When adding sub-surface temperature data down to 100m the situation becomes more complicated. In this case the temperature EOFs used in LIM are three-dimensional (3D) (e.g. Hawkins and Sutton 2007). The fourth row (Short 3D) demonstrates that using 50 years of the full 3D temperature data reduces the skill in the predictions from the 'Long SST' case, especially for short lead times. This is presumably because 50 years is not long enough to fully sample and represent the sub-surface temperature variability. The fifth row (Mixed 3D) shows that using the full length of the SST record with a shorter sub-surface record produces more skill for longer lead time predictions than the SSTs alone. This case is the closest to the actual historical observational data amounts. The sixth row (Long 3D) shows a more idealised case, and a further increase in skill for long lead times if sub-surface data was available for as long as the SST data. However, for predictions up to 5 year lead times, using the SST data alone tends to produce the highest skill (not shown). Finally, the extreme case of using all 3D data is shown in the bottom row (All 3D), and the skill has again improved from the 'Long 3D' case. Interestingly, some skill is now found in the tropical north Atlantic in HadCM3 with the addition of sub-surface data. However, elsewhere the skill has decreased with the addition of sub-surface data which may be because we are representing different modes of variability.

Fig. 8 repeats the same analysis for HadGEM1, and finds similar increases in skill when using more SST data, but that including the sub-surface temperature data generally makes the predictions worse, even in the 'Long 3D' case, in all regions. This difference between HadGEM1 and HadCM3 could be because the mechanisms of decadal variability are different in the two GCMs. As discussed above, Hodson and Sutton (in prep.) suggest that there is an atmospheric teleconnection between variability in the far north Atlantic and tropical Atlantic in HadGEM1, indicating that adding sub-surface data may not help the predictions, as seen in Fig. 8. In HadCM3, Dong and Sutton (2005) demonstrated the important role of the sub-surface ocean in decadal Atlantic variability, consistent with the improved predictive skill when sub-surface data is added (Fig. 7), especially in the tropical Atlantic.

Lastly, we note that using temperature data on even deeper levels tends to reduce the correlation skill of the LIM predictions for both GCMs (not shown), but better ways of including this information could perhaps be found.

5 Conclusions and discussion

We have analysed the potential to make statistical decadal forecasts of Atlantic SSTs, using control integrations of two different GCMs in a perfect model approach. The main findings are as follows:

- Statistical decadal predictions of Atlantic SSTs can be made with significant skill for up to a decade, and should provide a suitable benchmark for GCM based decadal predictions in the future.
- The specific regions of significant predictive skill differ between the two GCMs, and are likely to be different again in observations.
- Regions with low potential predictability can have high actual predictability as information can propagate from non-local regions. Potential predictability is therefore not necessarily a robust measure of potential predictive skill.
- Prediction skill should be measured using more than one metric, e.g. correlation *and* RMS error.

It is important to again reiterate that we have considered an idealised case of predictive skill, and actual skill in predicting observed variability may be lower than found here because we do not have a perfect model or perfect observations. However, we believe that this type of perfect model approach is valuable to assess methods before moving to a more complex situation. We have considered two different GCMs to try and explore the sensitivity of the results to the GCM chosen. For example, HadGEM1 has a relatively weak ENSO (Johns et al. 2006) which may enhance the appearance of tropical North Atlantic prediction skill. A more comprehensive study using a wider range of GCMs and statistical techniques would be desirable.

When considering predictions in the real world, there will be additional skill from predicting the trend component. In Table 3 we show the anomaly correlation skill for predicting SSTs from an operational GCM-based prediction system for 1981-2001 based on the HadCM3 GCM (Smith et al. 2007), both initialised (DePreSys) and uninitialised (NoAssim). The skill for NoAssim is an estimate of the skill due to the trend component only, and this exceeds $r \sim 0.6$ for most regions at 6-10 year lead times. The DePreSys predictions generally perform better on short lead times as they predict some of the variability component as well as the trend. However, DePreSys does not produce significantly better predictions for longer lead times, apart from in the far North Atlantic (also see Robson 2010, Smith et al. 2010). A quantitative comparison with the skill from the statistical predictions is not appropriate, but we explore the spatial pattern of this skill in more detail in the Supplementary Information for interested readers.

The next step in this perfect model approach is to analyse transient forced integrations of the same GCMs. This should provide a suitable approach to subsequently analyse the historical observations and compare with the GCM-based decadal predictions being prepared for the forthcoming CMIP5 assessment. We will also need to consider the uncertainties in the SST observations, perhaps generating an ensemble of predictions.

We also note that it may be possible to further increase skill by using longer lead time averaging periods, depending on the application that the forecast may be required for. Future work will also focus on predictions in other ocean basins and on examining other measures of prediction skill. There are also undoubtedly more complex statistical methods that may be appropriate for making such predictions and further improve skill; this will be explored in future studies. Better use of deeper temperature data, and the addition of other climate variables (such as ocean salinity or sea level pressure) into the statistical model could also improve skill (e.g. Newman et al. 2010). The statistical methods also provide a testbed for exploring where additional observations would be most beneficial for improving predictions (Fig. 6, also see e.g. Hawkins and Sutton 2009a).

Furthermore, the skill in predicting SSTs could produce some additional skill in predicting other climate variables such as surface air temperature or precipitation, either by using the predicted SSTs to drive atmosphere-only GCMs (e.g. Sutton and Hodson 2005; Scaife et al. 2009) or through a further statistical step (e.g. Colman and Davey 1999). Statistical predictions of SSTs are also being used to make operational predictions of regions likely to experience coral bleaching on seasonal timescales (Liu et al. 2009) and similar applications could be found on decadal timescales.

Acknowledgements We thank Chun-Kit Ho, Fiona Underwood, Len Shaffrey, Dan Hodson and Manoj Joshi for useful suggestions. The two anonymous reviewers provided valuable comments which helped improve the paper. The authors are supported by NCAS-Climate (EH, JR, RS) and the NERC VALOR project (JR). The research leading to these results has received funding (EH, NK) from the European Community's 7th framework programme (FP7/2007-2013) under grant agreement No. GA212643 (THOR: 'Thermohaline Overturning - at Risk', 2008-2012).

References

Arnell NW, Delaney EK (2006) Adapting to climate change: Public water supply in England and Wales. Clim Change 78: 227–255 doi:10.1007/s10584-006-9067-9

Barnston AG, van den Dool HM, Rodenhuis DR, Ropelewski CR, Kousky VE, O'Lenic EA, Livezey RE, Zebiak SE, Cane MA, Barnett TP, Graham NE, Ji M, Leetmaa A (1994) Long-Lead Seasonal Forecasts? Where Do We Stand? Bull Amer Meteor Soc 75: 2097–2114

Biastoch A, Boning CW, Lutjeharms JRE (2008) Agulhas leakage dynamics affects decadal variability in Atlantic overturning circulation. Nature 456: 489–492 doi: 10.1038/nature07426

Boer GJ (2004) Long time-scale potential predictability in an ensemble of coupled climate models. Clim Dyn 23: 29–44 doi:10.1007/s00382-004-0419-8

Boer GJ, Lambert SJ (2008) Multi-model decadal potential predictability of precipitation and temperature. Geophys Res Lett 35: L05 706 doi:10.1029/2008GL033234

Collins M, Botzet M, Carril AF, Drange H, Jouzeau A, Latif M, Masina S, Otteraa OH, Pohlmann H, Sorteberg A, Sutton R, Terray L (2006) Interannual to decadal climate predictability in the North Atlantic: a multimodel-ensemble study. J Climate 19: 1195–1202 doi:10.1175/JCLI3654.1

Collins M, Sinha B (2003) Predictability of decadal variations in the thermohaline circulation and climate. Geophys Res Lett 30: 1306 doi:10.1029/2002GL016504

Colman A, Davey M (1999) Prediction of summer temperature, rainfall and pressure in Europe from preceding winter North Atlantic Ocean temperature. Int J Clim 19: 513 – 536

Colman A, Davey M (2003) Statistical prediction of global sea-surface temperature anomalies. Int J Clim 23: 1677 – 1697 doi:10.1002/joc.956

Delworth TL, Mann ME (2000) Observed and simulated multi-decadal variability in the Northern Hemisphere. Climate Dyn 16: 661–676 doi:10.1007/s003820000075

Dong B, Sutton RT (2005) Mechanism of interdecadal thermohaline circulation variability in a coupled ocean-atmosphere GCM. J Climate 18: 1117 – 1135

Dunstone NJ, Smith DM (2010) Impact of atmosphere and sub-surface ocean data on decadal climate prediction. Geophys Res Lett 37: L02 709 doi:10.1029/2009GL041609

Emanuel K (2005) Increasing destructiveness of tropical cyclones over the past 30 years. Nature 436: 686–688 doi:10.1038/nature03906

Folland CK, Palmer TN, Parker DE (1986) Sahel rainfall and worldwide sea temperatures, 1901-85. Nature 320: 602–607 doi:10.1038/320602a0

Goldenberg SB, Landsea CW, Mestas-Nuñez AM, Gray WM (2001) The Recent Increase in Atlantic Hurricane Activity: Causes and Implications. Science 293: 474 – 479 doi: 10.1126/science.1060040

Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, Johns TC, Mitchell JFB, Wood RA (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Climate Dyn 16: 147–168

Griffies SM, Bryan K (1997) A predictability study of simulated North Atlantic multidecadal variability. Climate Dyn 13: 459–487

Haarsma RJ, Campos EJD, Drijfhout S, Hazeleger W, Severijns C (2010) Impacts of interruption of the Agulhas leakage on the tropical Atlantic in coupled ocean-atmosphere simulations. Clim Dyn in press doi:10.1007/s00382-009-0692-7

Hawkins E, Sutton R (2007) Variability of the Atlantic thermohaline circulation described by three-dimensional empirical orthogonal functions. Climate Dyn 29: 745–762 doi: 10.1007/s00382-007-0263-8

Hawkins E, Sutton R (2009a) Decadal predictability of the Atlantic Ocean in a coupled GCM: estimation of optimal perturbations using Linear Inverse Modelling. J Climate 22: 3960–3978 doi:10.1175/2009JCLI2720.1

Hawkins E, Sutton R (2009b) The potential to narrow uncertainty in regional climate predictions. Bull Amer Met Soc 90: 1095–1107 doi:10.1175/2009BAMS2607.1

Hodson DLR, Sutton RT (in prep.) The impact of model resolution on MOC adjustment in a coupled climate model

Jewson S, Hawkins E (2009) Improving the expected accuracy of forecasts of future climate using a simple bias-variance tradeoff available from http://arxiv.org/abs/0911.1904

Johns TC, Durman CF, Banks HT, Roberts MJ, McLaren AJ, Ridley JK, Senior CA, Williams KD, Jones A, Rickard GJ, Cusack S, Ingram WJ, Crucifix M, Sexton DMH, Joshi MM, Dong BW, Spencer H, Hill RSR, Gregory JM, Keen AB, Pardaens AK, Lowe JA, Bodas-Salcedo A, Stark S, Searl Y (2006) The New Hadley Centre Climate Model (HadGEM1): Evaluation of Coupled Simulations. J Climate 19: 1327–1353 doi: 10.1175/JCLI3712.1

Johnson HL, Marshall DP (2002) A theory for the surface Atlantic response to thermohaline variability. J Phys Ocean 32: 1121–1132

Keenlyside NS, Latif M, Jungclaus J, Kornblueh L, Roeckner E (2008) Advancing decadal-scale climate prediction in the North Atlantic sector. Nature 453: 84–88 doi: 10.1038/nature06921

Laepple T, Jewson S, Coughlin K (2008) Interannual temperature predictions using the CMIP3 multi-model ensemble mean. Geophys Res Lett 35: L10701 doi: 10.1029/2008GL033576

Lean JL, Rind DH (2009) How will Earth's surface temperature change in future decades? Geophys Res Lett 36: L15 708 doi:10.1029/2009GL038932

Lee TCK, Zwiers FW, Zhang X, Tsao M (2006) Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing. J Clim 19: 5305–5318 doi: 10.1175/JCLI3912.1

Liu G, Matrosova LE, Penland C, Gledhill DK, Eakin C, Webb RS, Christensen TRL, Heron SF, Morgan JA, Skirving WJ, Strong AE NOAA Coral Reef Watch Coral Bleaching Outlook System in Proceedings of the 11th International Coral Reef Symposium, Ft. Lauderdale, Florida 951–955 2009

Lorenz EN (1973) On the existence of extended range predictability. J Atmos Sci 12: 543–546

Meehl GA, Goddard L, Murphy J, Stouffer RJ, Boer G, Danabasoglu G, Dixon K, Giorgetta MA, Greene AM, Hawkins E, Hegerl G, Karoly D, Keenlyside N, Kimoto M, Kirtman B, Navarra A, Pulwarty R, Smith D, Stammer D, Stockdale T (2009) Decadal prediction: can it be skillful? Bull Amer Met Soc 90: 1467–1485 doi: 10.1175/2009BAMS2607.1

Meehl GA, Stocker TF, Collins W, Friedlingstein P, Gaye AT, Gregory JM, Kitoh A, Knutti R, Murphy JM, Noda A, Raper SCB, Watterson IG, Weaver AJ, Zhao ZC Global climate projections. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change Cambridge University Press, Cambridge, UK. 2007

Michaels A, Close A, Malmquist D, Knap A (1997) Climate science and insurance risk. Nature 389: 225–227 doi:10.1038/38378

Minobe S, Maeda A (2005) A 1° monthly gridded sea-surface temperature dataset compiled from ICOADS from 1850 to 2002 and Northern Hemisphere frontal variability. Int J Climatol 25: 881–894 doi:10.1002/joc.1170

Newman M, Alexander MA, Scott JD (2010) An empirical model of tropical ocean dynamics. Clim Dyn submitted

Penland C, Magorian T (1993) Prediction of Niño 3 Sea Surface Temperatures using linear inverse modelling. J Climate 6: 1067 – 1076

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108: 4407 doi: 10.1029/2002JD002670

Robson JI (2010) Understanding the performance of a decadal prediction system Ph.D. thesis University of Reading, UK

Saha S, Nadiga S, Thiaw C, Wang J, Wang W, Zhang Q, Van den Dool HM, Pan HL, Moorthi S, Behringer D, Stokes D, Peña M, Lord S, White G, Ebisuzaki W, Peng P, Xie P (2006) The NCEP Climate Forecast System. Journal of Climate 19 (15): 3483–3517 doi:10.1175/JCLI3812.1

Scaife A, Kucharski F, Folland C, Kinter J, Bronnimann S, Fereday D, Fischer A, Grainger S, Jin E, Kang I, Knight J, Kusunoki S, Lau N, Nath M, Nakaegawa T, Pegion P, Schubert S, Sporyshev P, Syktus J, Yoon J, Zeng N, Zhou T (2009) The CLIVAR

C20C project: selected twentieth century climate events. Clim Dyn 33: 603–614 doi: 10.1007/s00382-008-0451-1

Shaffrey LC, Stevens I, Norton WA, Roberts MJ, Vidale PL, Harle JD, Jrrar A, Stevens DP, Woodage MJ, Demory ME, Donners J, Clark DB, Clayton A, Cole JW, Wilson SS, Connolley WM, Davies TM, Iwi AM, Johns TC, King JC, New AL, Slingo JM, Slingo A, Steenman-Clark L, Martin GM (2009) U.K. HiGEM: The New U.K. High-Resolution Global Environment Model: Model Description and Basic Evaluation. J Climate 22: 1861–1896 doi:10.1175/2008JCLI2508.1

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. Science 317: 796–799 doi:10.1126/science.1139540

Smith DM, Eade R, Dunstone NJ, Fereday D, Murphy JM, Pohlmann H, Scaife A (2010) Skilful multi-year predictions of Atlantic hurricane frequency. Nature Geosci in press doi:10.1038/ngeo1004

Sutton RT, Hodson DLR (2005) Atlantic Ocean forcing of North American and European summer climate. Science 309: 115–118 doi:10.1126/science.1109496

Tziperman E, Zanna L, Penland C (2008) Non-normal thermohaline circulation dynamics in a coupled ocean-atmosphere GCM. J Phys Ocean 38: 588 – 604 doi: 10.1175/2007JPO3769.1

van den Dool H (1994) Searching for analogues, how long must we wait? Tellus 46A: 314–324

van den Dool H Empirical methods in short-term climate prediction Oxford University Press, Oxford, UK. 2007

van den Dool H, Huang J, Fan Y (2003) Performance and analysis of the constructed analogue method applied to U.S. soil moisture over 1981-2001. J Geophys Res 108: D16, 8617 doi:10.109/2002JD003114

Vecchi GA, Swanson KL, Soden BJ (2008) Whither Hurricane Activity? Science 322: 687–689 doi:10.1126/science.1164396

Table 1 Anomaly correlation skill measures for region averages for HadCM3 (HadGEM1). Values in bold indicate the highest correlation for each GCM for particular lead times and regions.

Whole Atlantic (30°S - 66°N)				
Method	Year 1	Year 2	Years 3-5	Years 6-10
Lag. Corr.	**0.67** (**0.66**)	**0.50** (0.39)	0.42 (0.30)	0.12 (0.21)
Glo. LIM	0.28 (0.32)	0.20 (0.27)	0.08 (0.12)	0.07 (0.05)
Atl. LIM	0.59 (0.60)	0.47 (0.41)	0.34 (0.30)	0.09 (0.16)
Glo. CA	0.61 (0.47)	0.49 (**0.43**)	0.40 (0.35)	**0.41** (**0.42**)
Atl. CA	0.53 (0.49)	0.41 (0.30)	**0.46** (**0.42**)	0.31 (0.38)
North Atlantic (0°N - 66°N)				
Method	Year 1	Year 2	Years 3-5	Years 6-10
Lag. Corr.	**0.79** (**0.74**)	**0.65** (**0.58**)	0.48 (0.40)	0.22 (0.27)
Glo. LIM	0.43 (0.40)	0.24 (0.31)	0.13 (0.23)	0.07 (0.06)
Atl. LIM	0.70 (0.66)	0.57 (0.55)	0.47 (0.45)	0.16 (0.19)
Glo. CA	0.75 (0.59)	0.62 (0.56)	0.47 (0.47)	**0.50** (**0.51**)
Atl. CA	0.71 (0.66)	0.57 (0.47)	**0.59** (**0.56**)	0.43 (0.47)
Atlantic Main Development Region (10°N - 20°N)				
Method	Year 1	Year 2	Years 3-5	Years 6-10
Lag. Corr.	0.39 (**0.32**[*])	**0.20** (0.15)	-0.33 (-0.09)	-0.06 (-0.06)
Glo. LIM	0.35 (0.26)	0.03 (0.05)	-0.09 (0.12)	-0.01 (0.00)
Atl. LIM	0.33 (0.32)	0.16 (**0.22**)	**0.06** (**0.17**)	0.09 (0.16)
Glo. CA	**0.45** (0.23)	0.10 (0.16)	0.03 (0.05)	**0.26** (**0.36**)
Atl. CA	0.34 (0.26)	0.16 (0.19)	0.06 (0.13)	0.15 (0.22)
Far North Atlantic (50°N - 66°N)				
Method	Year 1	Year 2	Years 3-5	Years 6-10
Lag. Corr.	**0.77** (**0.77**)	0.61 (0.59)	0.36 (0.35)	0.07 (0.20)
Glo. LIM	0.33 (0.39)	0.22 (0.32)	0.21 (0.17)	0.01 (-0.01)
Atl. LIM	0.73 (0.68)	**0.64** (**0.63**)	0.44 (0.39)	-0.07 (-0.08)
Glo. CA	0.75 (0.66)	0.60 (0.55)	0.33 (0.23)	0.17 (0.18)
Atl. CA	0.77 (0.72)	0.64 (0.58)	**0.54** (**0.42**)	**0.30** (**0.26**)

[*] Persistence forecast has an anomaly correlation of 0.33 here.

Table 2 Amounts of data considered in the different observing system experiments. Sub-surface temperature data is included on all model depth levels down to 100m. Practically, the 'Mixed 3D' case is achieved by taking the full 140 years of 3D data, but assuming all sub-surface temperature data for the first 90 years to have zero anomalies.

Experiment	SST data	Sub-surface data
	[years]	[years]
Short SST	50	0
Long SST	140	0
All SST		
- HadCM3	1000	0
- HadGEM1	546	0
Short 3D	50	50
Mixed 3D	140	50
Long 3D	140	140
All 3D		
- HadCM3	1000	1000
- HadGEM1	546	546

Table 3 Anomaly correlation skill measures for region averages for DePreSys and NoAssim from 1981-2001 (Smith et al. 2007). Bold values indicate where one system is performing substantially better.

<table>
<tr><th colspan="5">Whole Atlantic (30°S - 66°N)</th></tr>
<tr><th>Method</th><th>Year 1</th><th>Year 2</th><th>Years 3-5</th><th>Years 6-10</th></tr>
<tr><td>DePreSys</td><td><b>0.82</b></td><td>0.52</td><td><b>0.70</b></td><td>0.68</td></tr>
<tr><td>NoAssim</td><td>0.62</td><td>0.51</td><td>0.52</td><td><b>0.74</b></td></tr>
<tr><th colspan="5">North Atlantic (0°N - 66°N)</th></tr>
<tr><th>Method</th><th>Year 1</th><th>Year 2</th><th>Years 3-5</th><th>Years 6-10</th></tr>
<tr><td>DePreSys</td><td><b>0.86</b></td><td>0.70</td><td><b>0.74</b></td><td>0.76</td></tr>
<tr><td>NoAssim</td><td>0.77</td><td>0.70</td><td>0.62</td><td>0.76</td></tr>
<tr><th colspan="5">Atlantic Main Development Region (10°N - 20°N)</th></tr>
<tr><th>Method</th><th>Year 1</th><th>Year 2</th><th>Years 3-5</th><th>Years 6-10</th></tr>
<tr><td>DePreSys</td><td><b>0.70</b></td><td>0.33</td><td>0.46</td><td>0.62</td></tr>
<tr><td>NoAssim</td><td>0.43</td><td><b>0.45</b></td><td>0.47</td><td>0.60</td></tr>
<tr><th colspan="5">Far North Atlantic (50°N - 66°N)</th></tr>
<tr><th>Method</th><th>Year 1</th><th>Year 2</th><th>Years 3-5</th><th>Years 6-10</th></tr>
<tr><td>DePreSys</td><td><b>0.86</b></td><td><b>0.77</b></td><td><b>0.82</b></td><td><b>0.68</b></td></tr>
<tr><td>NoAssim</td><td>0.69</td><td>0.60</td><td>0.35</td><td>0.61</td></tr>
</table>

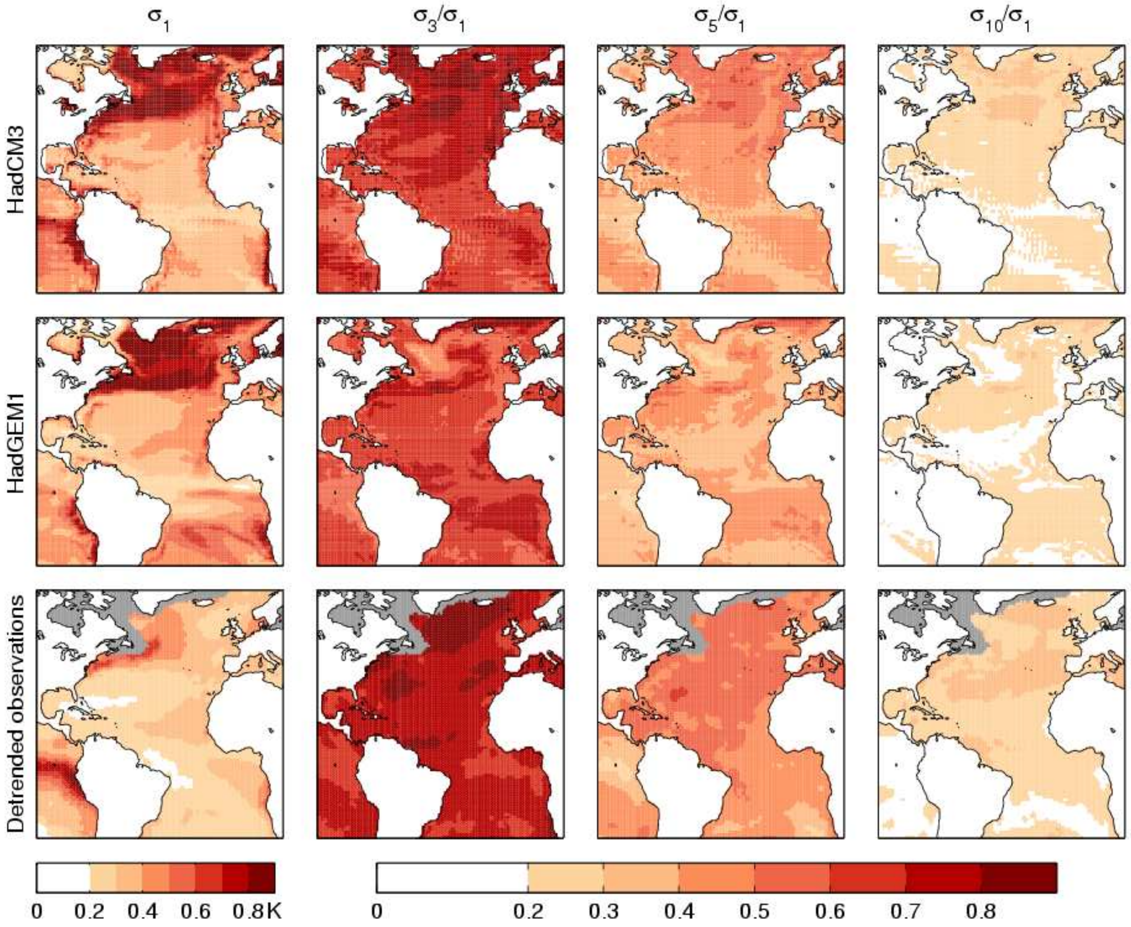


Fig. 1 Potential predictability of Atlantic SSTs. Left column: interannual standard deviation of SST. Other columns: ratio of standard deviation of 3, 5 and 10 year means to the interannual standard deviation. Rows: HadCM3 (top), HadGEM1 (middle) and HadISST observations from 1870-2009, detrended using cubic spline (bottom). Regions where sea ice precludes analysis are shown in grey. The global version of this figure is shown as Fig. S1.

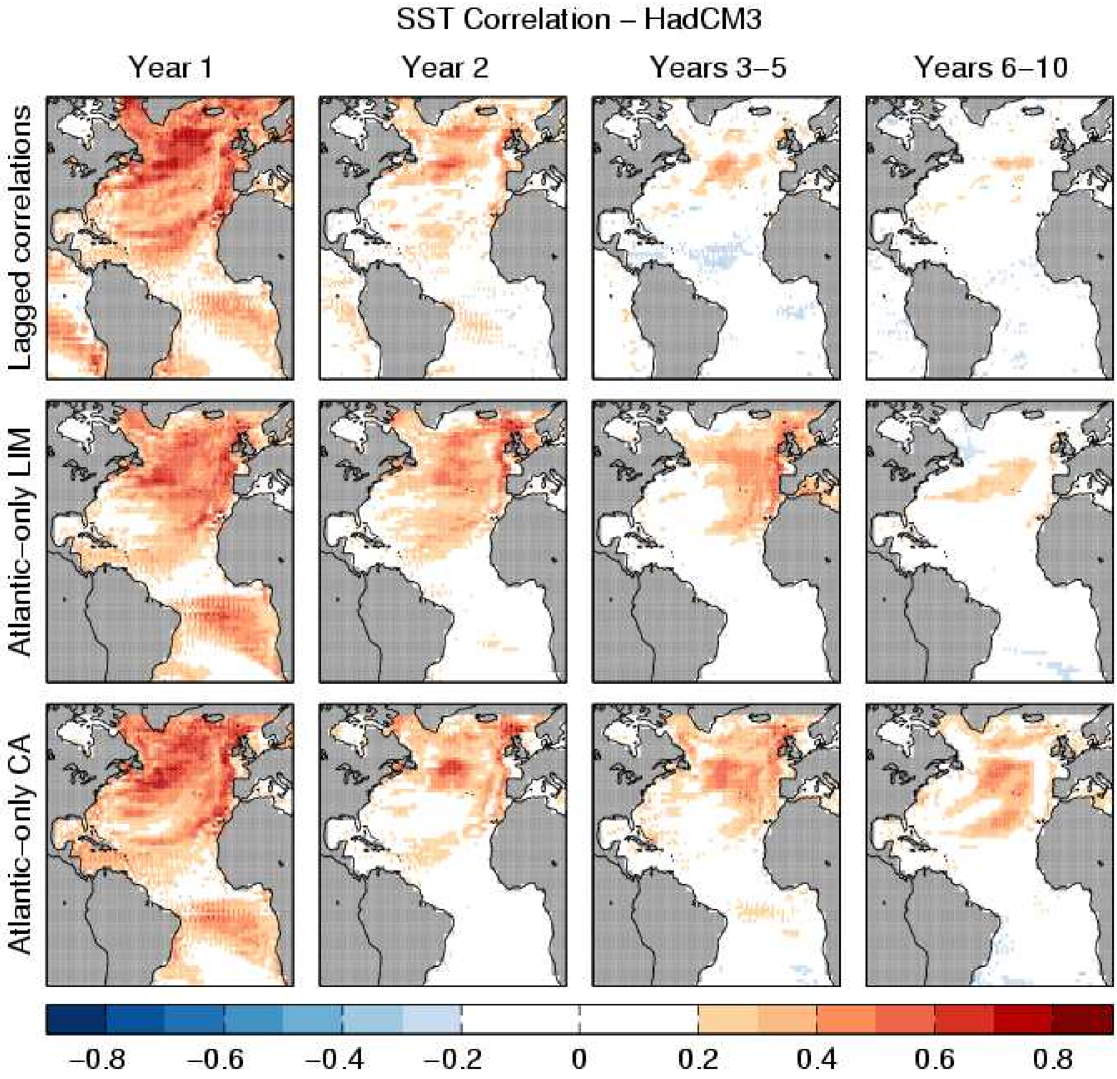


Fig. 2 SST anomaly correlation skill for HadCM3, for lagged correlations, Linear Inverse Modelling (LIM) and Constructed Analogue (CA) predictions using only the Atlantic domain. The skill for the full set of predictions is shown in Fig. S3.

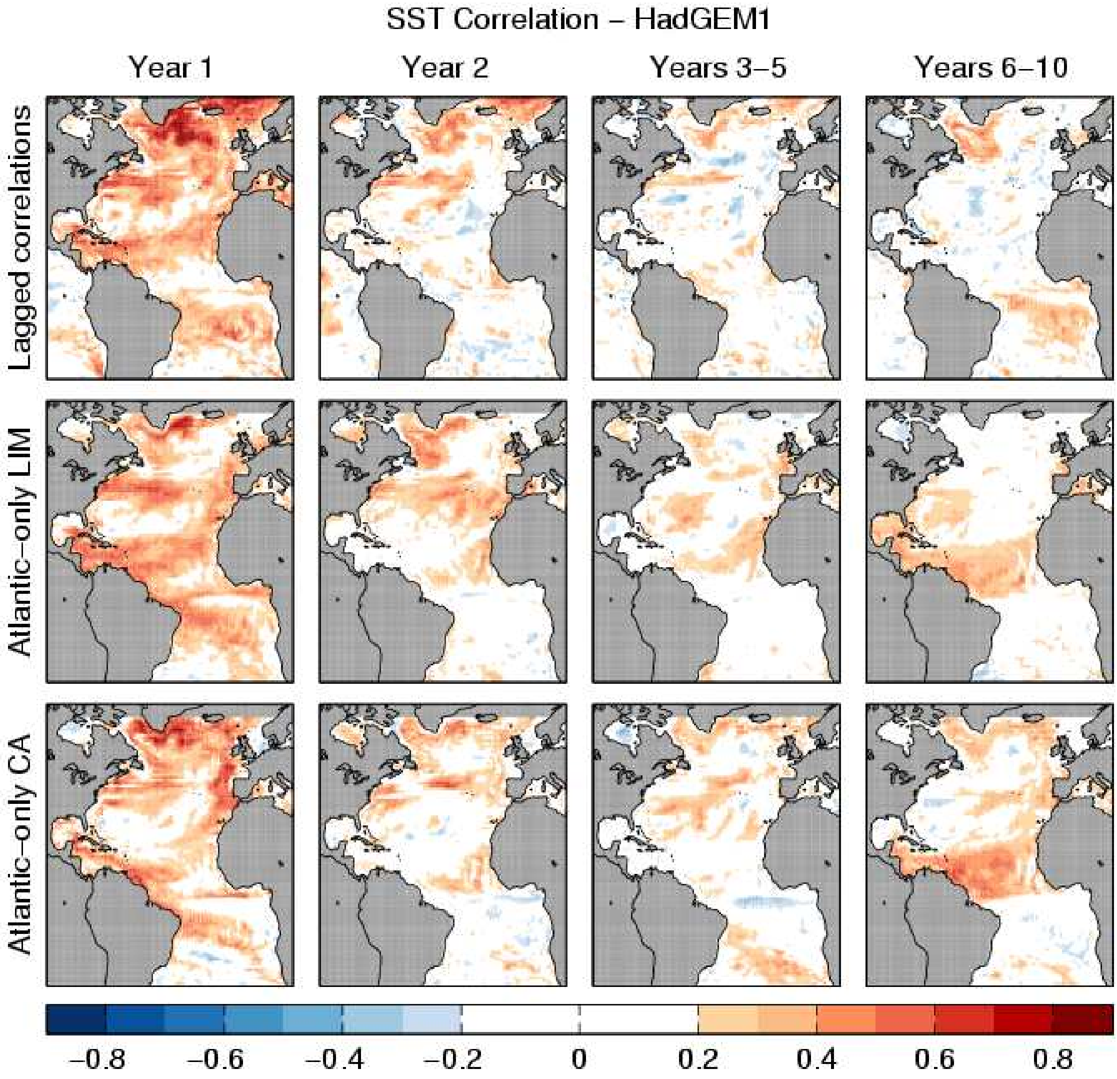


Fig. 3 Same as Fig. 2 for HadGEM1. The skill for the full set of predictions is shown in Fig. S4.

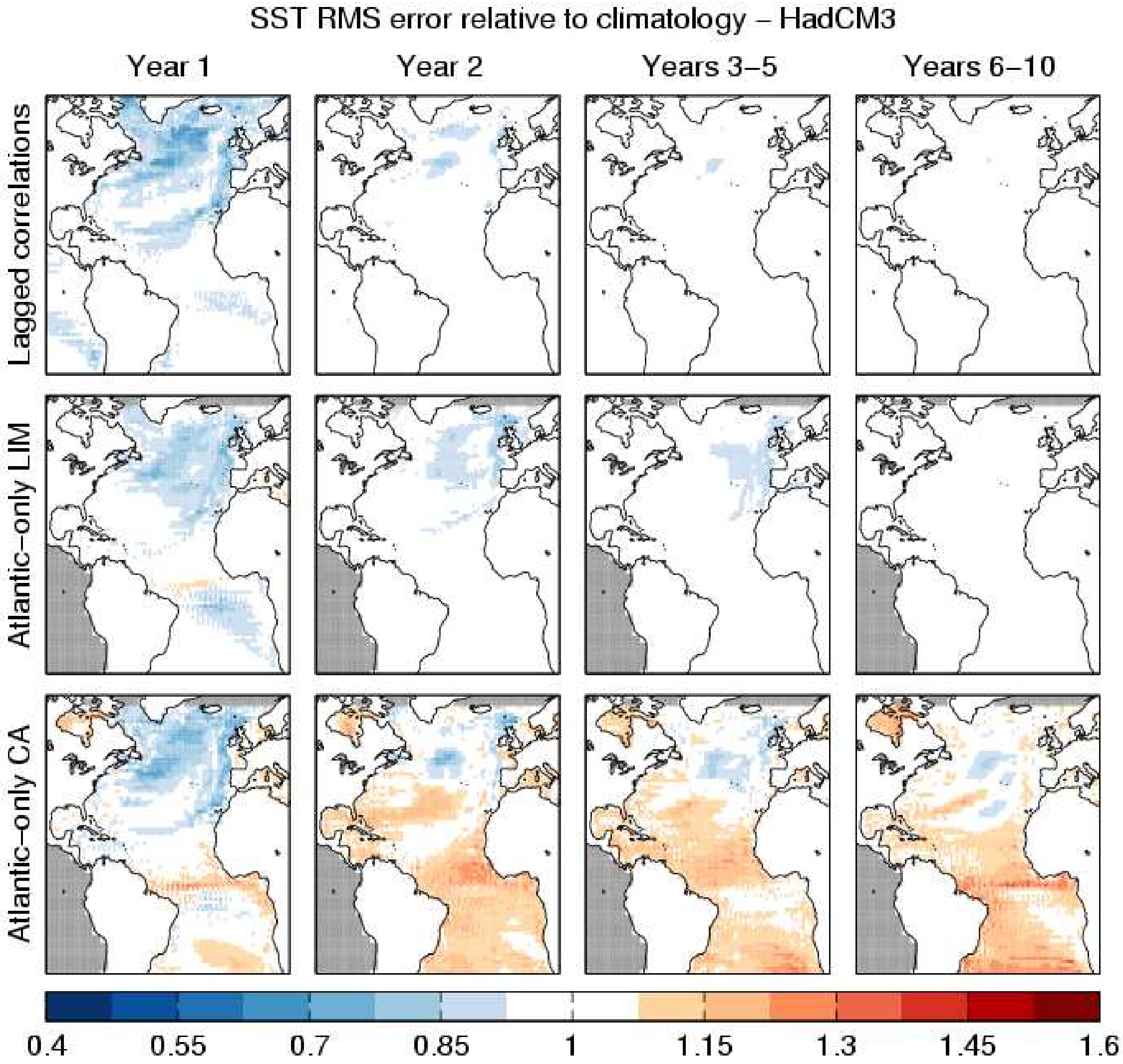


Fig. 4 Root mean square error relative to the climatological RMS error for HadCM3, for lagged correlations, LIM and CA predictions using the Atlantic domain. The skill for the full set of predictions is shown in Fig. S5.

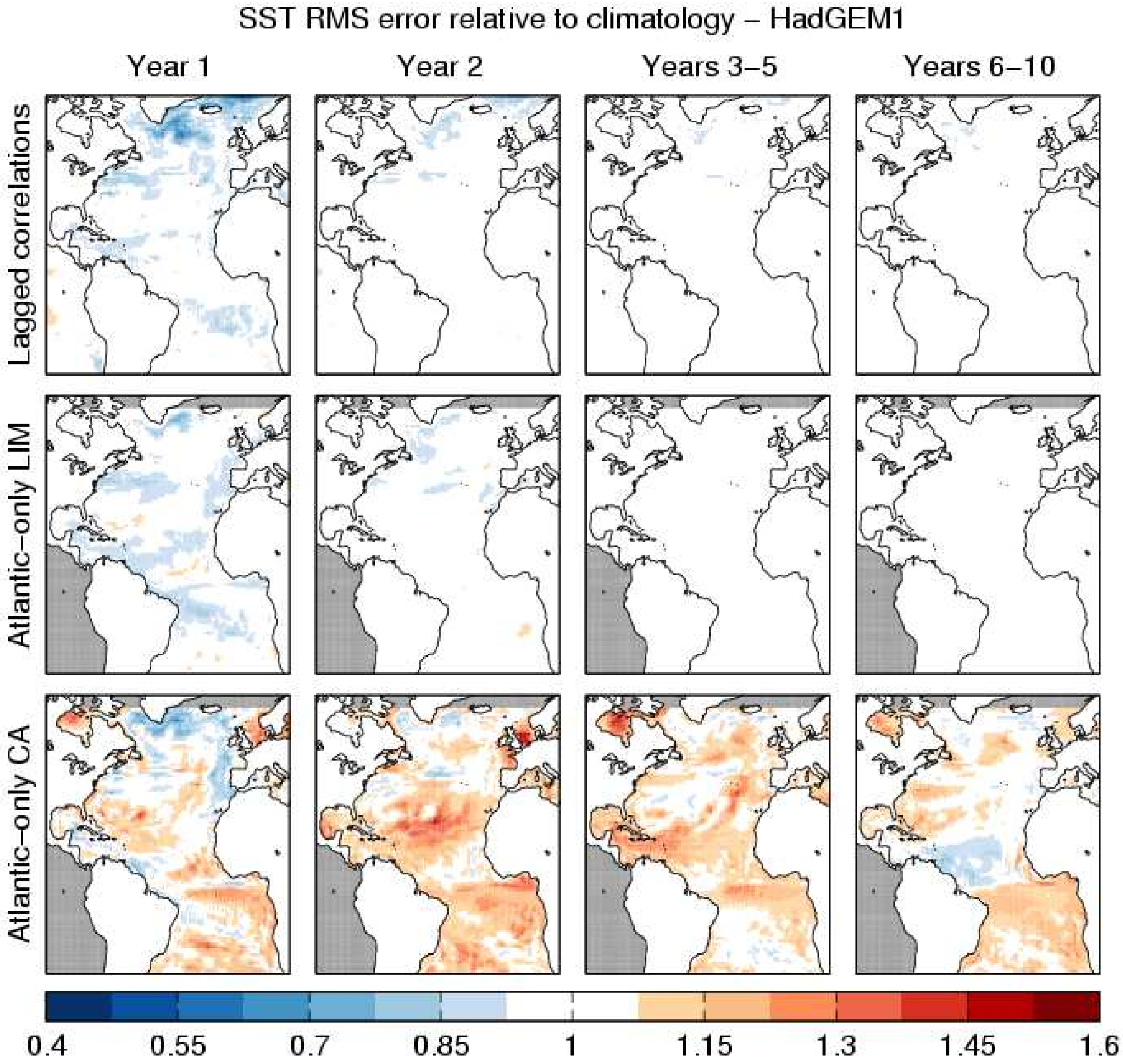


Fig. 5 Same as Fig. 4 for HadGEM1. The skill for the full set of predictions is shown in Fig. S6.

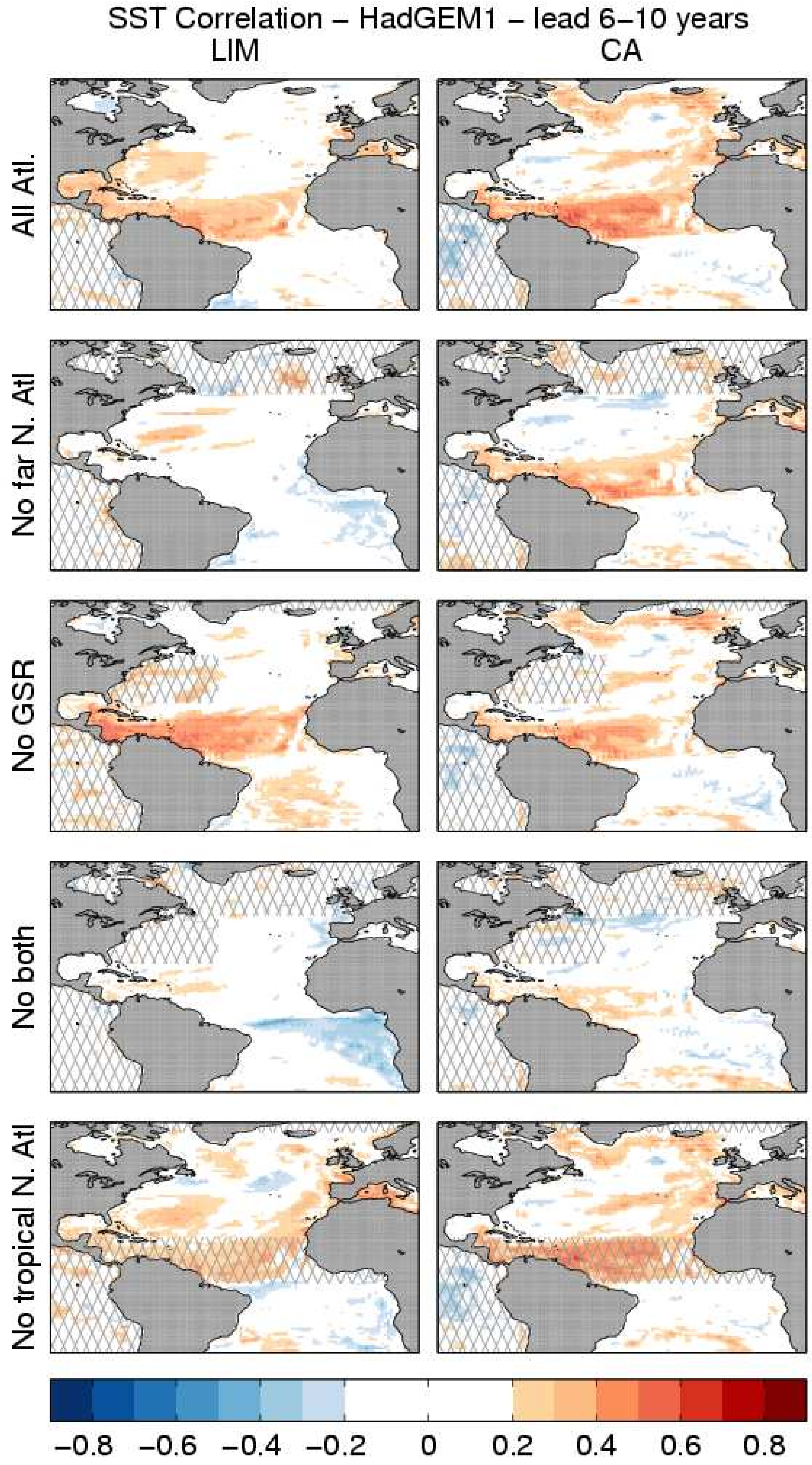


Fig. 6 Withholding data experiments: SST anomaly correlation skill for HadGEM1 for years 6-10, using different domains and statistical methods. The hatched regions are not included in constructing the statistical models. GSR = Gulf Stream Region.

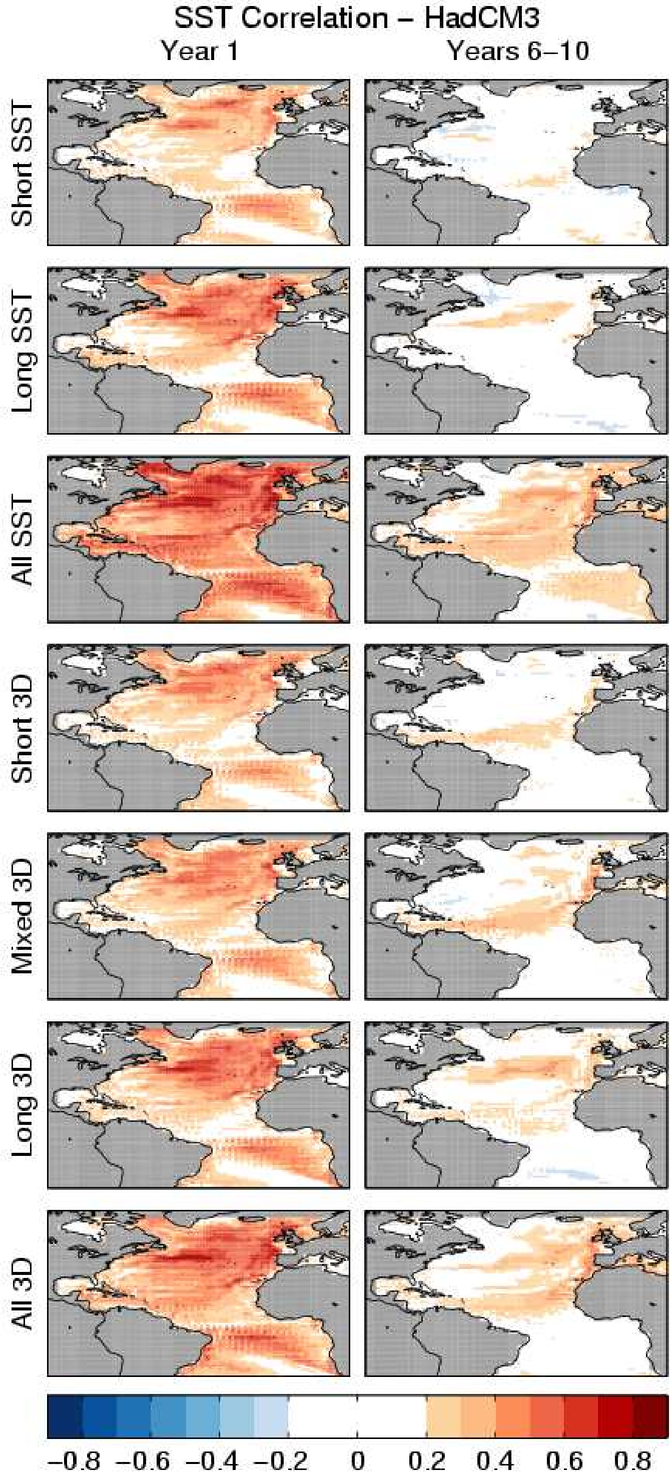


Fig. 7 Comparing the correlation skill of the Atlantic LIM predictions in HadCM3 using different amounts of surface and sub-surface data (see Table 2).

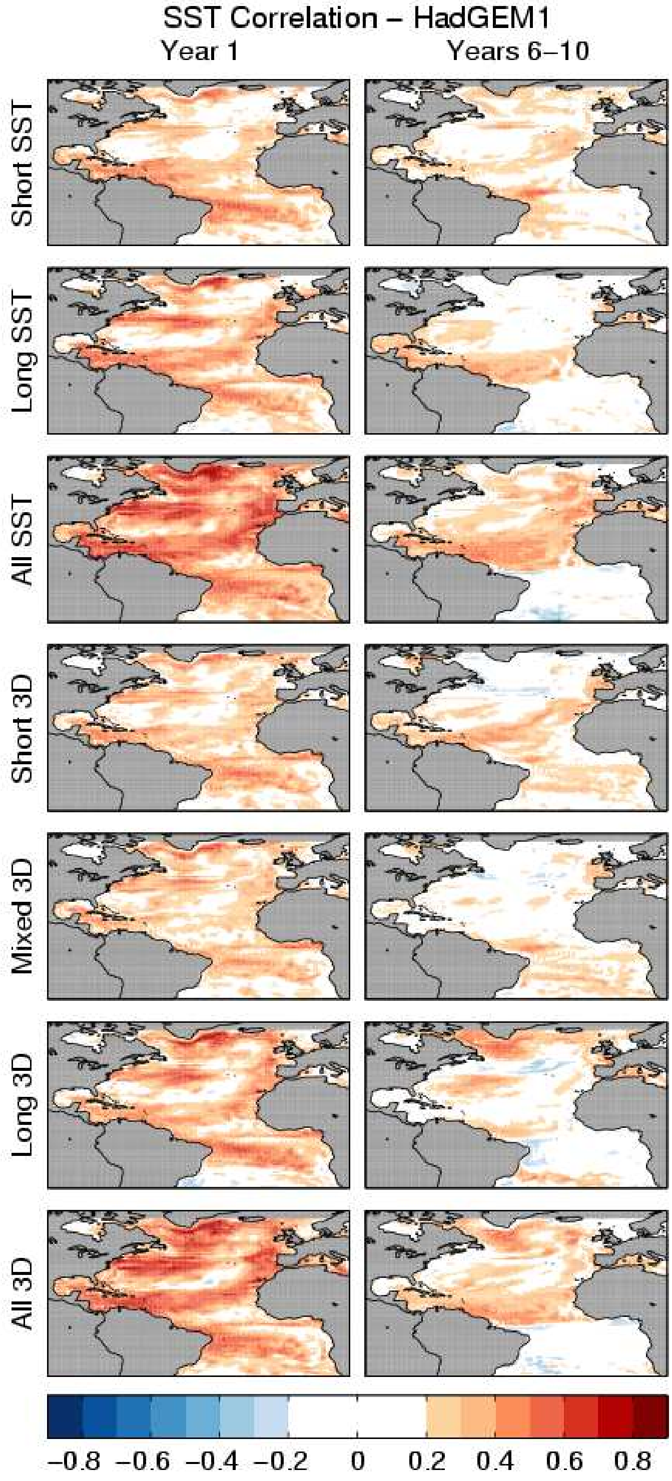


Fig. 8 Same as Fig. 7 for HadGEM1.

Evaluating the potential for statistical decadal predictions of sea surface temperatures with a perfect model approach:

Supplementary Information

Ed Hawkins[*], Jon Robson, Rowan Sutton, Doug Smith, Noel Keenlyside

Resubmitted to Climate Dynamics

This supplementary information augments the findings of the main paper by showing the global versions of many of the figures, and by including the skill of the prediction methods that were not shown in the main paper. Additionally, we include a brief discussion comparing the leading EOFs of the observations and GCMs. Finally, we discuss the skill of SST predictions in operational decadal prediction systems.

Potential predictability

Fig. S1 is the global version of Fig. 1 and shows the potential predictability (PP) estimates for the two GCMs and observations outside the Atlantic. Note particularly that Fig. S1b shows PP using the raw observations (top) and the observations detrended with a spline as in the main text (bottom).

It can be seen that much of the Atlantic has the highest levels of potential predictability, although the Pacific is also highlighted. Although the presence of potential predictability is not *necessarily* an indicator of actual predictability, this motivates considering predictions focussed on the Pacific in more detail in future work to establish whether this potential predictability can be realised.

Leading EOFs of models and observations

In the LIM analysis in the main paper, EOFs are used to build the statistical model. Fig. S2 shows the leading EOFs of the observations (HadISST; Rayner et al. 2003) and

[*] E-mail: e.hawkins@reading.ac.uk

the control runs of the two GCMs considered in this study. Note that in the LIM analysis, the EOFs are estimated for each 140 year training period separately, but here we use the whole control run.

The leading EOF of HadISST shows an overall warming, likely associated with the radiatively forced signal. The leading EOF of both GCMs shows a pattern with opposite signs across the equator, which also resembles EOF2 of the observations. The second EOF of both GCMs shows a similarly signed signal either side of the equator, and this pattern is not clearly seen in EOF3 of the observations. This is not particularly surprising as EOFs will not separate the natural and anthropogenic components cleanly.

Prediction skill

Figs. S3-S6 are the global versions of Figs. 2-5, showing the skill measures over the entire global domain for all prediction types, and they highlight the other regions where some skill may be found. In particular, both the global LIM and CA methods show high levels of skill in the Indian (HadCM3) and Pacific (HadGEM1) basins, although only for lead times up to 2 years. From a comparison of the global and Atlantic-only predictions it can be seen that using the global SSTs is often superior in year 1 in the tropical Atlantic, likely related to predicting aspects of ENSO in the Pacific.

Comparison of operational decadal prediction systems

We now qualitatively compare the potential skill of the statistical approaches with the actual skill found in two operational GCM-based decadal prediction systems, DePreSys (Smith et al. 2007) (hereafter S07), which is based on HadCM3, and Keenlyside et al. (2008) (hereafter K08), which uses the ECHAM5/MPI-OM climate model. In these decadal predictions, skill comes from both predicting the trend and the variability around the trend. We also use the companion simulations which only predict the trend (denoted as NoAssim for S07 and '20th C' for K08). Note that we do not consider any forecast bias correction[1].

For DePreSys (S07), we use the original retrospective forecasts (or hindcasts) which

[1] It is possible that the skill measures used are sensitive to the bias correction and this will be explored in future work.

are started every season from 1981-2001, and we consider the skill of the ensemble mean of all 4 members for each prediction. Both the predictions and the NoAssim simulations do not include knowledge of volcanic eruptions that occur later in each hindcast. The predictions from K08 are started every 5 years from 1955-2005, and we consider the skill of the ensemble mean of all 3 members for each prediction. Again, the initialised predictions do not include future volcanic eruptions, but the ‘20th C’ simulations have the volcanic forcing prescribed. The other major difference between the two systems is that K08 only assimilate historical SSTs into the model, whereas S07 nudge to an analysis of temperature and salinity at all depths (Smith and Murphy 2007). Table 1 summarises the experimental configurations. When estimating the skill in these predictions below, the SST anomalies in the hindcasts are compared to observed anomalies relative to a chosen climatological period (1955-2004 for K08 and 1951-2006 for DePreSys [following Robson 2010]).

Anomaly correlation

There is no simple decomposition of the correlations of such an initialised and uninitialised decadal prediction system. However, Fig. S7a shows the SST correlation skill of the predictions from S07 for NoAssim (top row), DePreSys (middle row) and the difference (bottom row). The initialised system, DePreSys, generally outperforms NoAssim for year 1, especially in the tropics and in the far North Atlantic. At longer lead times there is additional skill only in the far North Atlantic.

Although a direct quantitative comparison between DePreSys and Fig. S3 is not appropriate, it is interesting to note that the initialised system does outperform the uninitialised system in many regions for year 1, and some regions for longer lead times. As demonstrated, the statistical methods are also providing additional skill over a climatological forecast in some regions, suggesting that they could match the skill of the GCM-based predictions, and this motivates their further development.

Fig. S8a shows the equivalent SST correlation skill from K08, again demonstrating increased predictive ability for the first year in most regions, and the north-east Atlantic at longer lead times. The correlation skill in years 6-10 is still above 0.8 over a large region, however many regions see a reduction in skill with initialisation. Note that the

95% confidence level for these correlations, assuming 9 independent predictions, is $r \approx 0.7$.

Both systems seem to show the potential to predict ENSO up to 2 years ahead. Beyond 2 years, the main skill outside the Atlantic is found in the east Pacific for K08.

Relative RMS error

As noted in the main paper, it is important to consider more than one measure of skill. For a comparison of the RMS error we consider that,

$$\mathrm{RMS}_{\mathrm{relative}} = \frac{\mathrm{RMS}_{\mathrm{init}}}{\mathrm{RMS}_{\mathrm{uninit}}}, \tag{1}$$

where ‘init’ and ‘uninit’ refer to the initialised and uninitialised predictions respectively. Note that this measure is relatively insensitive to the presence of the same trend in both predictions.

Fig. S7b shows this relative RMS error for DePreSys (S07) and identifies similar regions to the correlations where there is enhanced skill for DePreSys over NoAssim (shown in blue colours), particularly for the far North Atlantic.

Fig. S8b shows the relative RMS error for K08, again demonstrating improved skill for year 1. However, for longer lead times the RMS error is significantly worse than the ‘20th C’ projections for many regions. It may also be noted that regions of high correlation often also have large RMS errors, and this is probably due to the model’s internal variability being too large in this region (see K08).

Overall, it is seen that the initialised GCM based predictions provide some additional skill in predicting the internal variability than the uninitialised simulations, especially on shorter lead times. The trends provide most of the apparent skill on longer lead times.

References

Keenlyside NS, Latif M, Jungclaus J, Kornblueh L, Roeckner E (2008) Advancing decadal-scale climate prediction in the North Atlantic sector. Nature 453: 84–88 doi: 10.1038/nature06921

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108: 4407 doi: 10.1029/2002JD002670

Robson JI (2010) Understanding the performance of a decadal prediction system Ph.D. thesis University of Reading, UK

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. Science 317: 796–799 doi:10.1126/science.1139540

Smith DM, Murphy JM (2007) An objective ocean temperature and salinity analysis using covariances from a global climate model. J Geophys Res 112: C02 022 doi: 10.1029/2005JC003172

Table 1: Description of the GCM based decadal prediction systems considered, Smith et al. (2007) and Keenlyside et al. (2008). The two systems are based on different GCMs, consider different historical time periods, have different frequency of start dates, and initialisation strategies for both temperature (T) and salinity (S).

Study	GCM used	Years considered	Frequency	Initialisation strategy
S07	HadCM3	1981-2001	every season	anomalies, full depth T, S
K08	ECHAM5/MPI-OM	1955-2005	every 5 years	anomalies, SST only

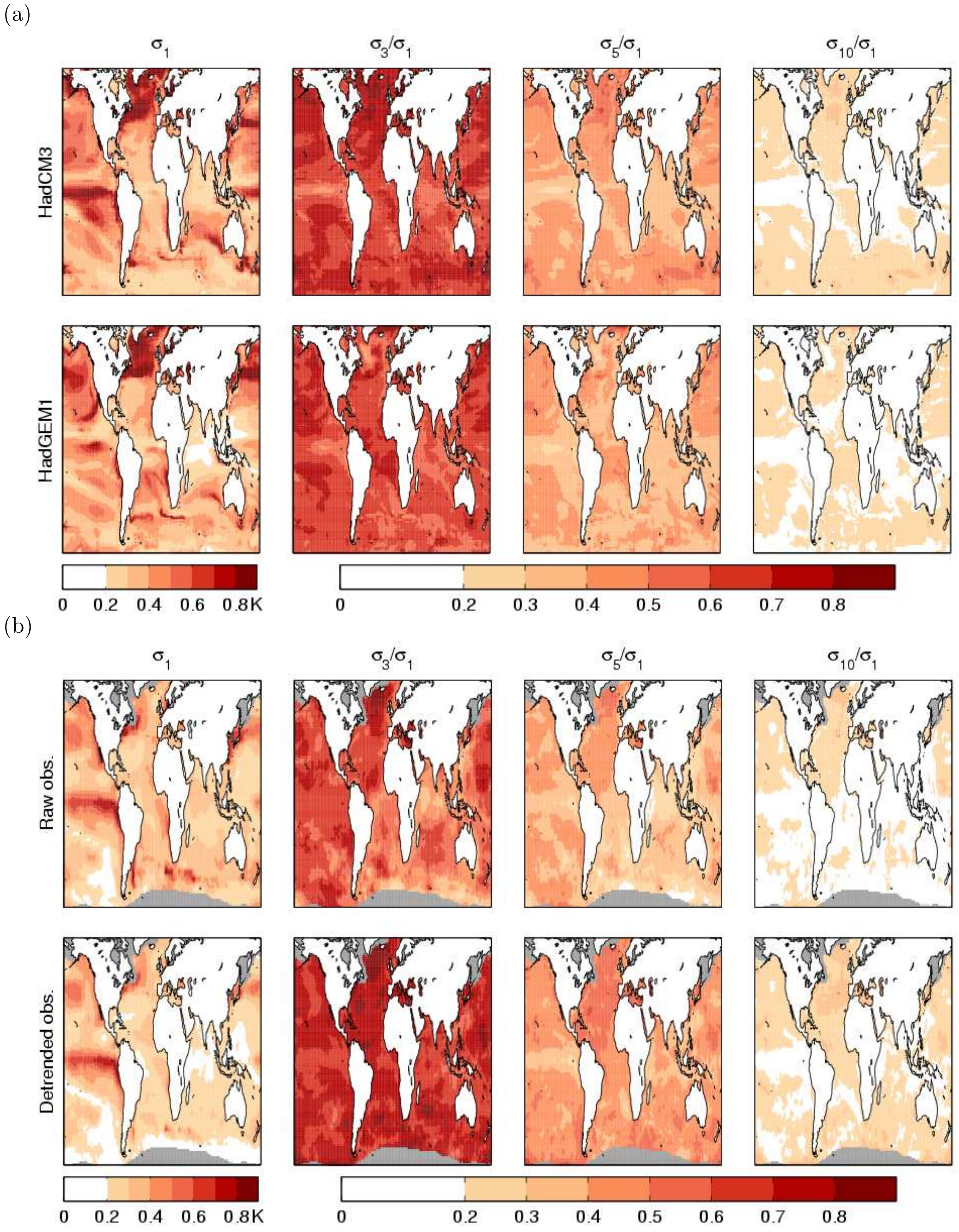


Figure S1: Potential predictability of Atlantic SSTs. Left column: interannual variability of SST. Other columns: ratio of standard deviation of 3, 5 and 10 year means to the interannual standard deviation. (a) Top row: HadCM3. Bottom row: HadGEM1. (b) HadISST observations from 1870-2009. Top row: raw observations. Bottom row: detrended observations, using cubic spline. Regions where sea ice precludes analysis are shown in grey.

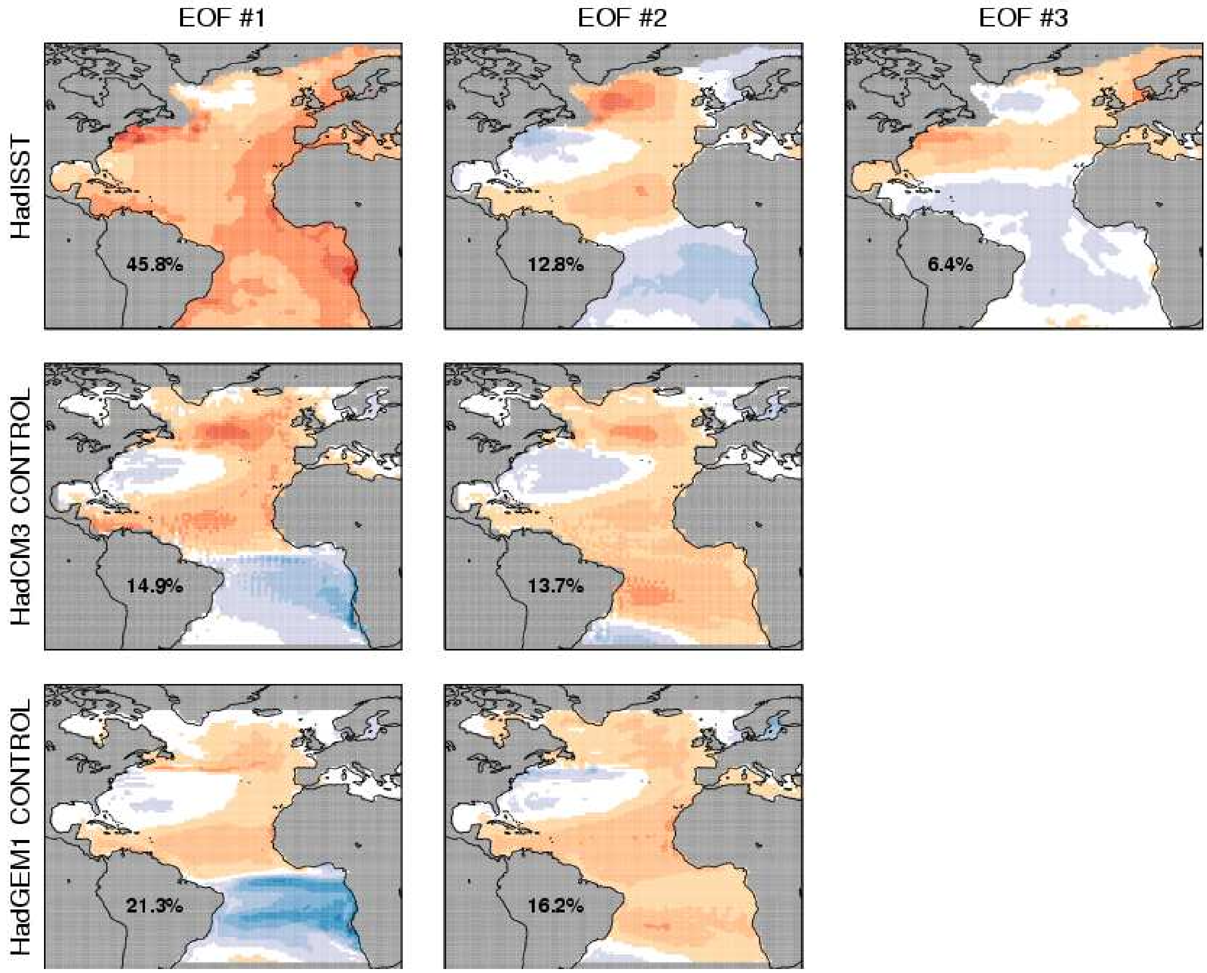


Figure S2: The leading correlation EOFs of the observations (HadISST, top) and the control runs of HadCM3 and HadGEM1.

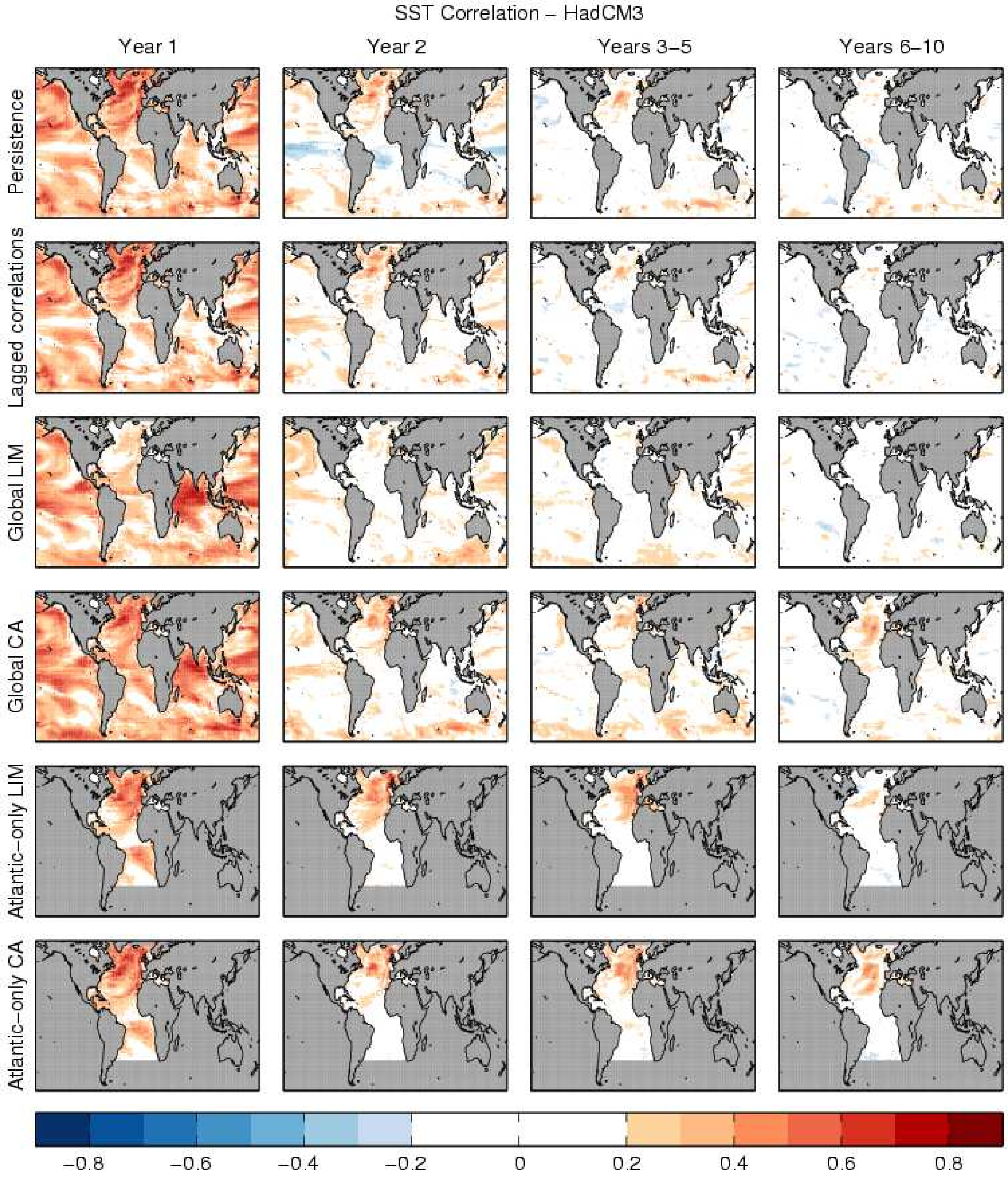


Figure S3: SST anomaly correlation skill for HadCM3, for persistence, lagged correlations, Linear Inverse Modelling (LIM) and Constructed Analogue (CA) predictions. The LIM and CA methods are repeated for both Global and Atlantic domains. The grey regions are masked out of the analysis.

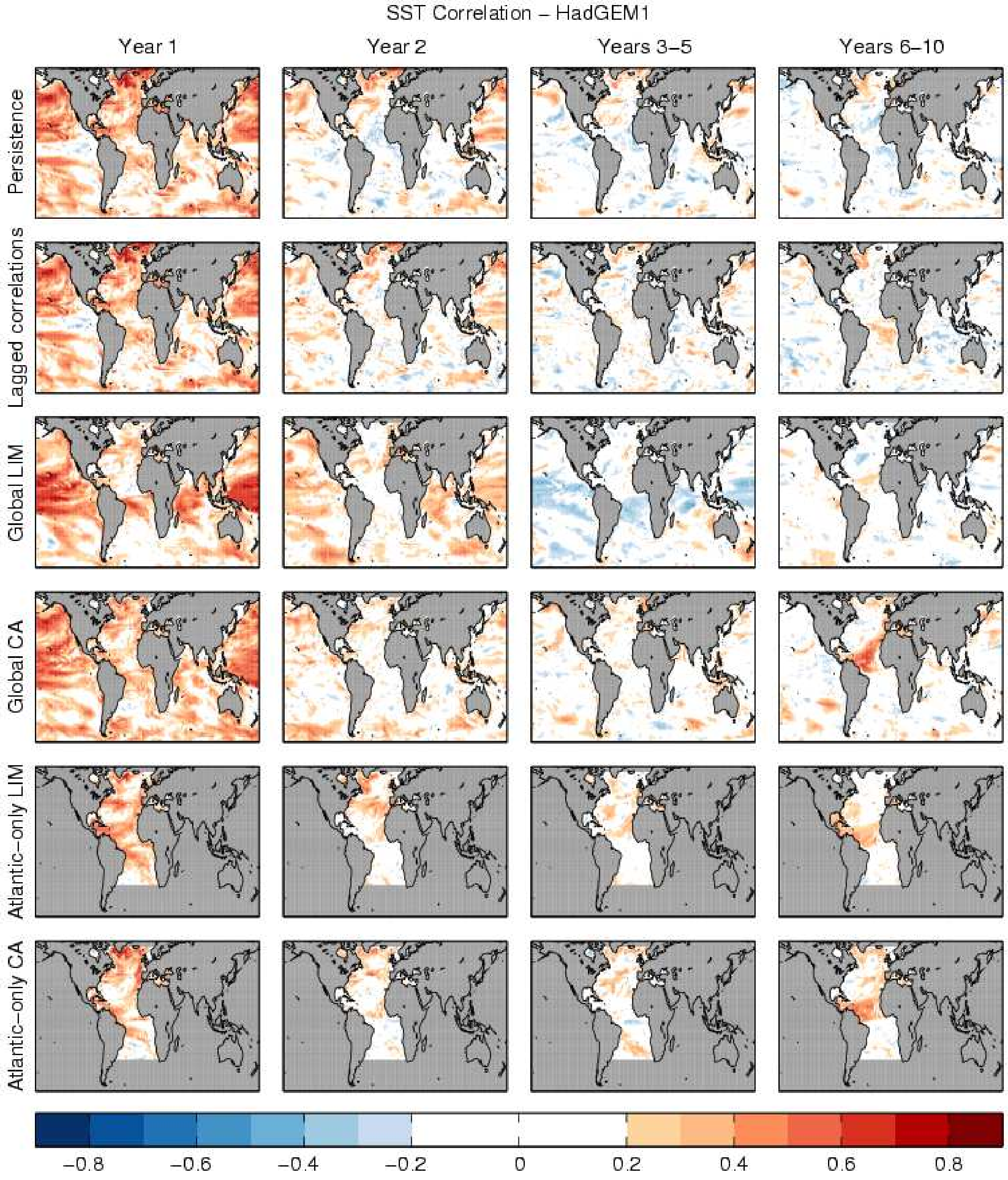


Figure S4: Same as Fig. S3 for HadGEM1.

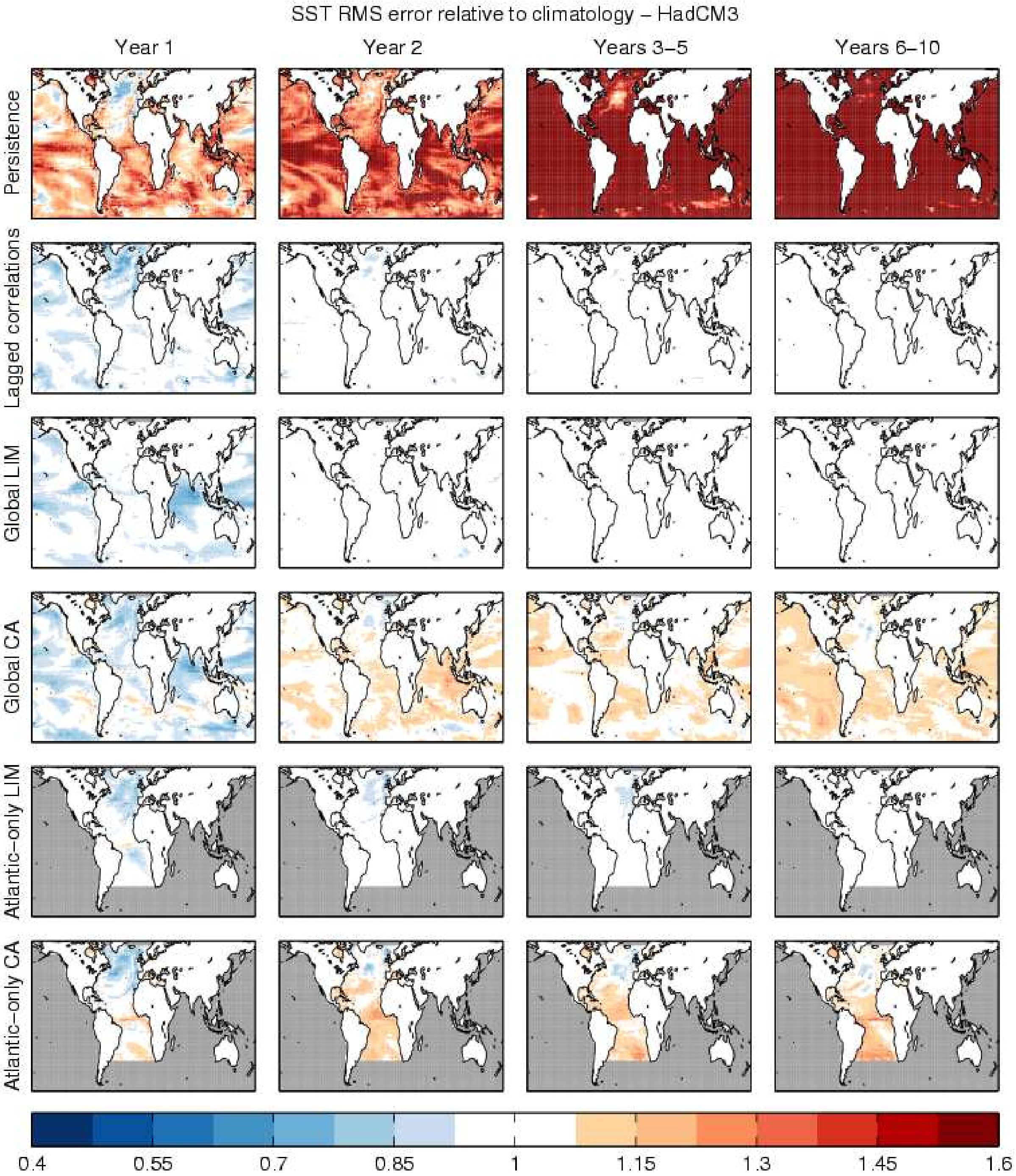


Figure S5: Root mean square error relative to the climatological RMS error for HadCM3, for persistence, lagged correlations, LIM and CA predictions. The LIM and CA methods are repeated for both Global and Atlantic domains. The grey regions are masked out of the analysis.

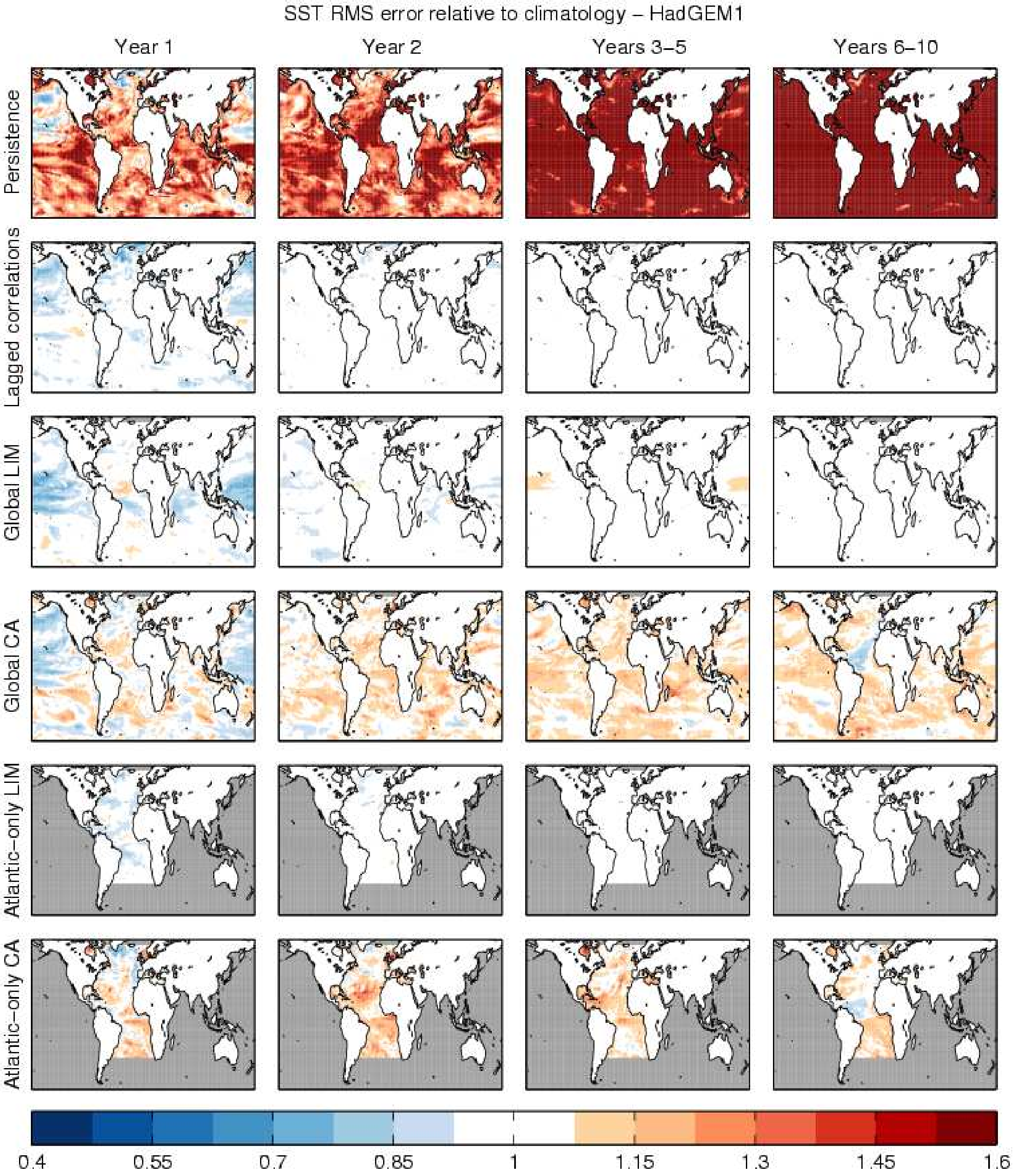
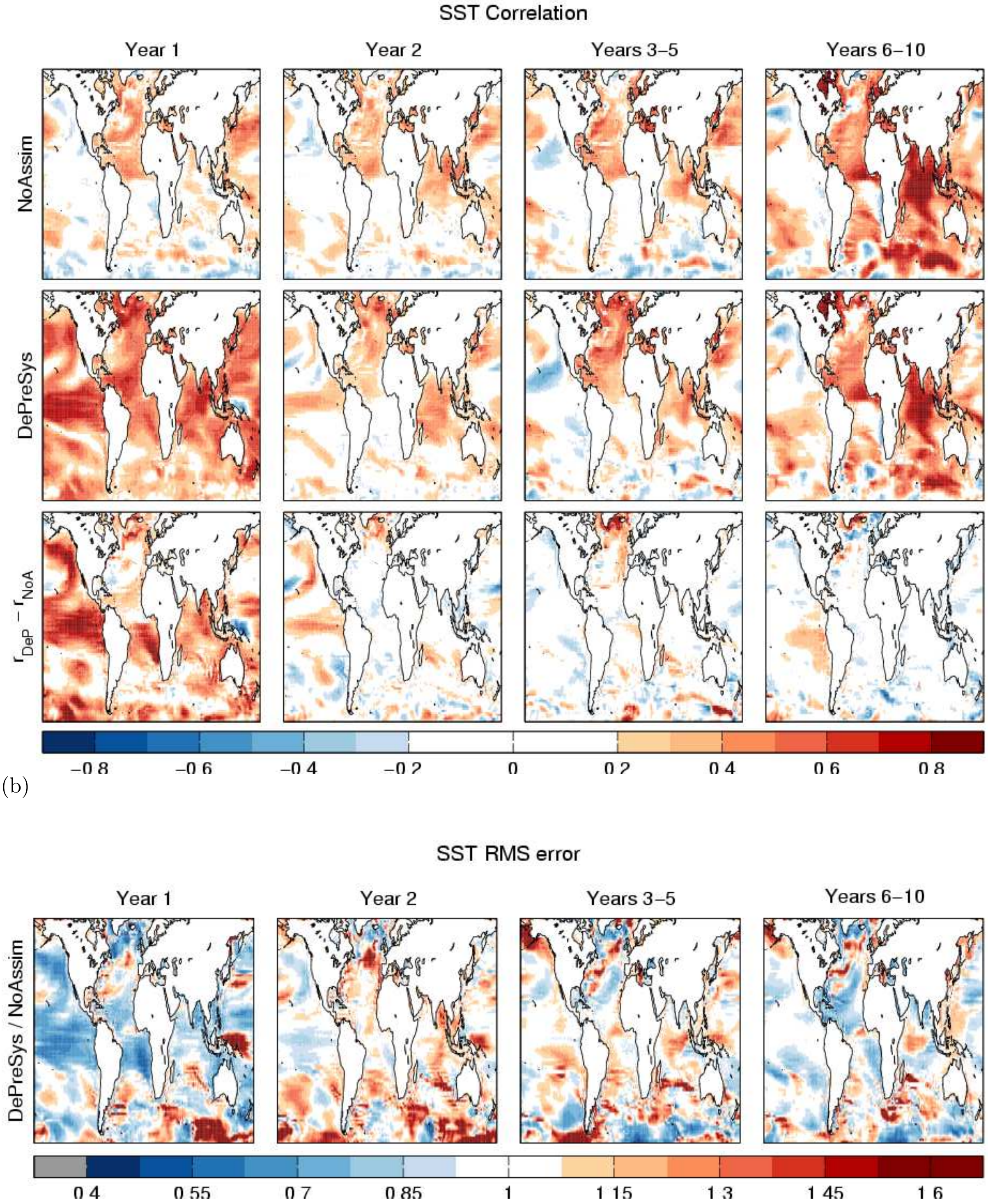


Figure S6: Same as Fig. S5 for HadGEM1.

(a)



(b)

Figure S7: Prediction skill of an operational GCM-based decadal prediction system (DePreSys; Smith et al. 2007) for years 1981-2001. (a) Correlation skill for the uninitialised ensemble (NoAssim), initialised ensemble (DePreSys) and the difference. (b) RMS error of DePreSys relative to NoAssim.

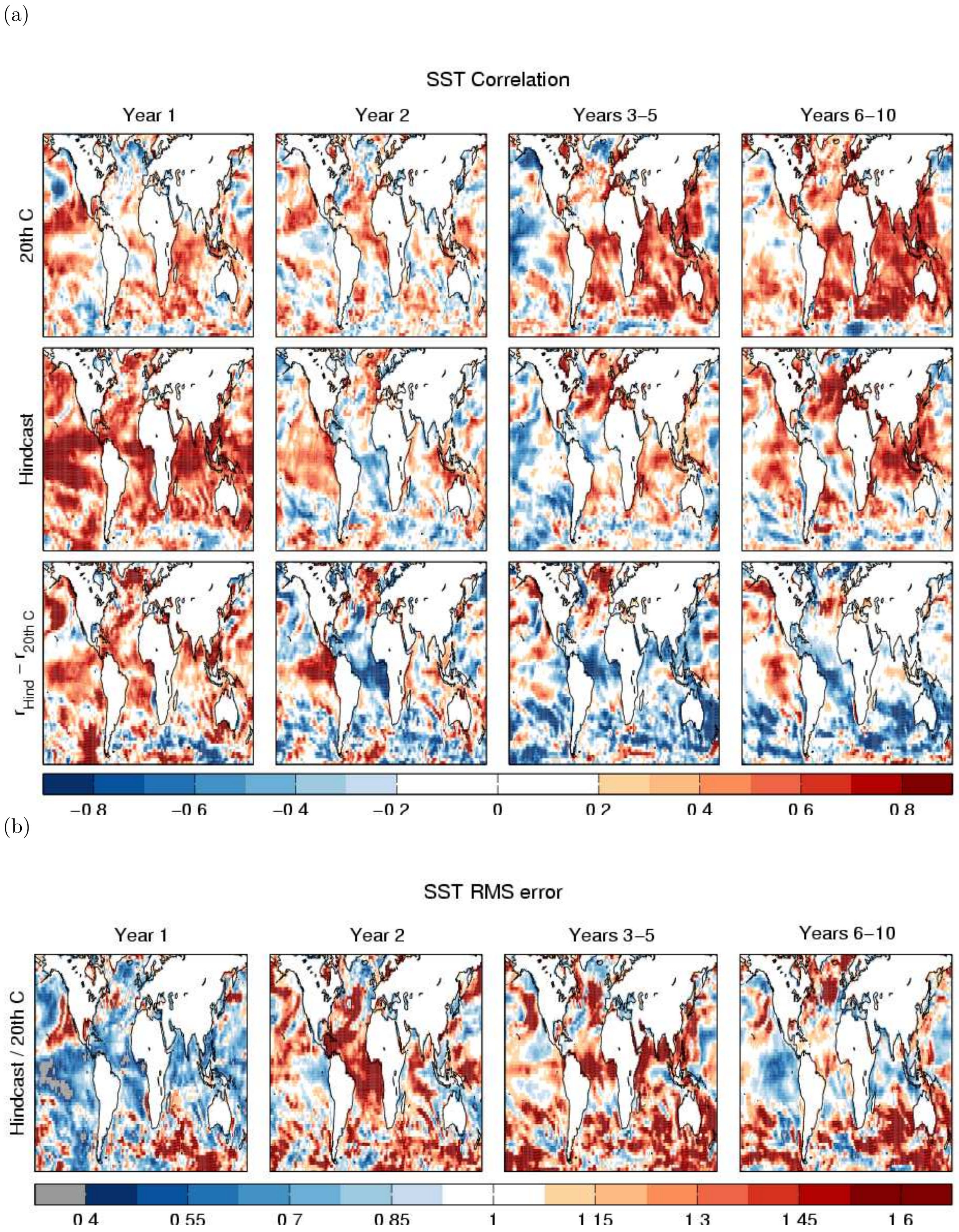


Figure S8: Prediction skill of an operational GCM-based decadal prediction system (Keenlyside et al. 2008) for years 1955-2005. (a) Correlation skill for the uninitialised ensemble (20th C), initialised ensemble (Hindcast) and the difference. (b) RMS error of Hindcast relative to 20th C.