Statistical decadal predictions for sea surface temperatures: a benchmark for dynamical GCM predictions

Article

Published Version

www.reading.ac.uk/centaur

Statistical decadal predictions for sea surface temperatures: a benchmark for dynamical GCM predictions

Chun Kit Ho · Ed Hawkins · Len Shaffrey ·
Fiona M. Underwood

Abstract Accurate decadal climate predictions could be used to inform adaptation actions to a changing climate. The skill of such predictions from initialised dynamical global climate models (GCMs) may be assessed by comparing with predictions from statistical models which are based solely on historical observations. This paper presents two benchmark statistical models for predicting both the radiatively forced trend and internal variability of annual mean sea surface temperatures (SSTs) on a decadal timescale based on the gridded observation data set HadISST. For both statistical models, the trend related to radiative forcing is modelled using a linear regression of SST time series at each grid box on the time series of equivalent global mean atmospheric CO_2 concentration. The residual internal variability is then modelled by (1) a first-order autoregressive model (AR1) and (2) a constructed analogue model (CA). From the verification of 46 retrospective forecasts with start years from 1960 to 2005, the correlation coefficient for anomaly forecasts using trend with AR1 is greater than 0.7 over parts of extra-tropical North Atlantic, the Indian Ocean and western Pacific. This is primarily related to the prediction of the forced trend. More importantly, both CA and AR1 give skillful predictions of the internal variability of SSTs in the subpolar gyre region over the far North Atlantic for lead time of 2–5 years, with correlation coefficients greater than 0.5. For the subpolar gyre and parts of the South Atlantic, CA is superior to AR1 for lead time of 6–9 years. These statistical forecasts are also compared with ensemble mean retrospective forecasts by DePreSys, an initialised GCM. DePreSys is found to out-perform the statistical models over large parts of North Atlantic for lead times of 2–5 years and 6–9 years, however trend with AR1 is generally superior to DePreSys in the North Atlantic Current region, while trend with CA is superior to DePreSys in parts of South Atlantic for lead time of 6–9 years. These findings encourage further development of benchmark statistical decadal prediction models, and methods to combine different predictions.

Keywords Decadal prediction · Statistical ·
Sea surface temperatures · Global climate model

1 Introduction

Climate predictions for the near-term (up to about 30 years), especially on regional scales, can be used to inform adaptation actions to a changing climate, for example infrastructure planning and hazard preparedness (Adger et al. 2005; Challinor 2009). In recent years, there has been major progress in the development of such predictions based on global climate models (GCMs) initialised with atmospheric and oceanic observations (e.g. Smith et al. 2007; Keenlyside et al. 2008; Fyfe et al. 2011). As part of the Fifth Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2011), experimental decadal predictions produced by initialised GCMs from different modelling groups will be compared (e.g. Smith et al. 2012). Proper evaluation of these dynamical decadal predictions can aid efforts in improving future GCM simulations, for example their initialisation schemes.

C. K. Ho (✉) · E. Hawkins · L. Shaffrey
NCAS-Climate, Department of Meteorology,
University of Reading, PO Box 243, Earley Gate,
Reading RG6 6BB, UK
e-mail: c.k.ho@reading.ac.uk

F. M. Underwood
Department of Mathematics and Statistics,
University of Reading, PO Box 220,
Whiteknights, Reading RG6 6AX, UK

On decadal timescales changes in climate are caused by both its response to radiative forcing and its internal variability (Keenlyside and Ba 2010; Solomon et al. 2011; Goddard et al. 2012b). With a non-stationary climate, an initialised GCM can show significant prediction skill if it predicts a long-term forced trend consistent with observations (e.g. Lee et al. 2006; van Oldenborgh et al. 2012), but often users are also interested in the magnitude of the internal variability. The skill of a particular initialised GCM in predicting such internal variability may be assessed by comparing its retrospective forecasts (also known as hindcasts in the decadal prediction literature) and those produced by the identical GCM without assimilation of observations (e.g. Smith et al. 2007). Alternatively, Laepple et al. (2008) proposed the use of bias-corrected ensemble mean projections from multiple uninitialised GCMs as benchmarks to evaluate the skill of decadal retrospective forecasts by initialised GCMs.

Being much less computationally expensive to run, predictions from statistical models can also serve as benchmarks when assessing the skill of dynamical GCM predictions. A good benchmark statistical model should be able to capture the basic characteristics of the climate system which we want to predict. Ideally it should also be trained solely by historical observations without information from physical climate models, but this is limited by the length and quality of available observational record. Simple benchmarks such as persistence and climatology have been extensively used in the evaluation of seasonal climate predictions (e.g. Barnston et al 1994; Colman and Davey 2003). As for decadal climate predictions, since both radiative forcing and internal variability can be important, more advanced benchmark statistical models which also incorporate these effects are more appropriate. For example, Lean and Rind (2009) projected global annual mean air temperatures using a multiple linear regression of temperature on anthropogenic influence, solar radiation, ENSO variability and volcanic aerosols. Krueger and von Storch (2011) produced decadal retrospective forecasts of global annual mean temperatures by separating the observed temperature time series into two components representing forced trend and internal variability respectively. The forced trend was modelled by a linear regression between multi-model ensemble mean temperatures and atmospheric CO_2 concentration. After removing this forced trend component, the residual internal variability was modelled as a first-order autoregressive process. Fildes and Kourentzes (2011) compared retrospective forecasts of global mean temperatures for up to 20 years ahead by the UK Met Office Decadal Prediction System DePreSys (Smith et al. 2007; which will also be considered in this study, see Sect. 2.2) with those by various time series models, ranging from a local linear trend model to a multivariate neural network.

DePreSys was found to be more skillful than the statistical models for lead time of 1–4 years, while some of the more complex statistical models, such as the neural network model, have better prediction skill for longer lead times.

Some other studies have produced benchmark statistical models for climate variables other than air temperatures. With a perfect model approach using control integrations of two GCMs, Hawkins et al. (2011) showed that two empirical methods, namely Linear Inverse Modelling (LIM) and Constructed Analogue (CA), have significant skill in predicting the internal variability of sea surface temperatures (SSTs) over parts of the Atlantic for up to a decade. Zanna (2012) applied LIM to gridded observed Atlantic SST anomalies, with the forced trend removed using cubic splines. The retrospective forecasts were found to be skillful relative to climatology for a lead time of up to around 5 years.

In this paper we present two benchmark statistical models for predicting annual mean SSTs with a lead time of up to a decade based on the gridded observed data set HadISST. Variations of SSTs have important implications for atmospheric conditions, such as precipitation patterns and tropical cyclone activity (e.g. Sutton and Hodson 2005; Zhang and Delworth 2006; Smith et al. 2010). We will assess the predictive skill of the two statistical models by verifying a set of retrospective forecasts with start times from mid-1960s to mid-2000s, with a focus on examining the regional skill for different lead times. In addition, a brief comparison will be made with the corresponding forecasts by DePreSys to assess the relative strengths of dynamical and statistical predictions for different regions. Predictions for the Atlantic Ocean are particularly interesting and our discussion will put greater emphasis on this sector.

This paper is structured as follows. Section 2 gives a summary of HadISST, DePreSys and the external forcing time series used in this study. This is followed by descriptions of our benchmark statistical models and verification measures in Sect. 3. In Sect. 4 we assess the prediction skill of the statistical models, while in Sect. 5 we compare the skill of these models with DePreSys. The statistical forecast of Atlantic SSTs for the years 2012 to 2021 is presented in Sect. 6. Further discussion on our findings and concluding remarks are given in Sect. 7.

2 Observational and GCM data

2.1 HadISST

We use the HadISST data set (Rayner et al. 2003) to train our statistical prediction models, correct biases in the DePreSys predictions and verify retrospective forecasts from both the statistical models and DePreSys. This data set contains global monthly interpolated fields of SSTs on a

$1° \times 1°$ grid from 1870 to 2011. For the rest of this paper, we consider the anomalies of annual means of SSTs. To allow direct comparison between the statistical and DePreSys retrospective forecasts, here the annual mean for a certain year is defined to be the 12-month average from the December of the previous year to the November of the year concerned. The anomalies for each grid box are calculated by removing the corresponding mean for the years 1986 to 2005. It should be noted that the amount of data (e.g. in-situ SST observations, satellite data) used to construct this gridded data set varies with space and time. Particularly, because of the sparseness of observations over the southern oceans and near the polar regions, we perform our analyses only from $30°S$ to $70°N$. Grid boxes covered with sea-ice are also omitted in our analyses.

2.2 DePreSys

The UK Met Office Decadal Prediction System, DePreSys (Smith et al. 2007), is based on the third Hadley Centre coupled GCM HadCM3 (Gordon et al. 2000). Its atmosphere component has a horizontal resolution of $2.5° \times 3.75°$ and 19 vertical levels, while its ocean component has a $1.25° \times 1.25°$ resolution with 20 vertical levels. We use a perturbed physics ensemble of DePreSys with nine variants designed to sample climate model uncertainty. One of these nine variants uses the standard HadCM3 settings of physical parameters, while for the other eight, simultaneous perturbations of 29 atmospheric parameters are employed. These variants are chosen from previous experiments with larger perturbed physics ensembles (Collins et al. 2006, 2011), with the aim of spanning a wide range of parameter values while simulating physically plausible climate variability. The equilibrium climate sensitivity of the ensemble ranges from 2.6 to 7.1 °C.

For each model variant, atmospheric and oceanic analyses were assimilated from December 1958 to November 2007 as anomalies with respect to the model climate to create the initial conditions. Time-varying radiative forcings derived from observed changes in greenhouse gases, aerosol and solar irradiance were used up to the year 2000, after which the forcing based on the SRES A1B scenario (Nakićenović and Swart 2000) was applied. A total of 46 retrospective forecasts of global SSTs, starting on 1 November of each year from 1960 to 2005 with each extending to 10 years ahead, are available for each ensemble variant. We consider only the ensemble mean of annual mean SST forecasts from December to November.

2.3 Global mean equivalent CO_2 concentration

In the modelling of the forced trend of observed SSTs, we use the global mean equivalent CO_2 concentration from the Representative Concentration Pathways (RCPs) greenhouse gases concentration historical and projection data set (Meinshausen et al. 2011). These data have been used to drive the CMIP5 climate simulations. The equivalent CO_2 concentration incorporates the global net effects of all anthropogenic forcing agents, including greenhouse gases and aerosols. Observed concentrations from 1760 to 2005 are available. For subsequent years, the RCP4.5 concentration scenario is used for the forecasts we present in this paper, but the sensitivity of our results to alternative scenarios have also been tested.

3 Statistical prediction models for SST hindcasts and forecasts

In order to compare with DePreSys, our statistical retrospective forecasts for annual mean SSTs cover the same time period, i.e. we produce 46 sets of retrospective forecasts starting from 1960 to 2005, each of 10 years in length. In addition, we produce a forecast starting from 2011 for the years 2012–2021.

Variations in annual mean SSTs are caused by both the changes in natural or anthropogenic radiative forcing and the internal dynamics of the ocean. Internal variability at a longer timescale may temporarily amplify or reduce the underlying long-term trend in SSTs due to radiative forcing (Ting et al. 2009; DelSole et al. 2011). For our predictions, we attempt to model the two effects separately. We decompose the time series of observed SST anomalies $X_{i,t}$ for each grid box i in year t (relative to the 1986–2005 mean) into two components,

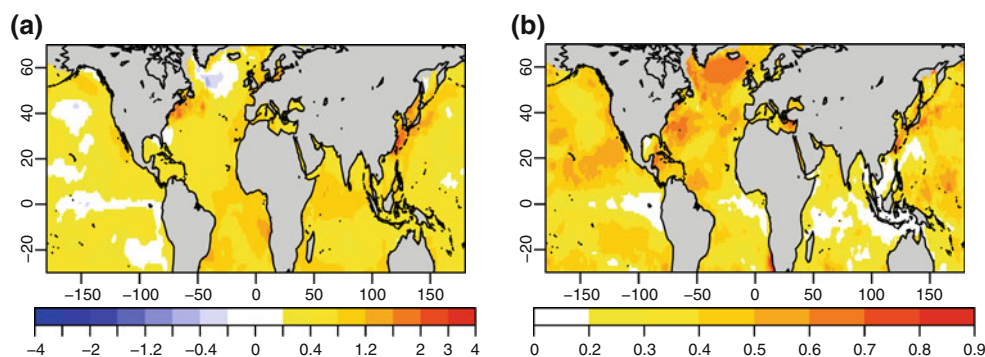$$X_{i,t} = X_{i,t}^f + X_{i,t}^v \tag{1}$$

where $X_{i,t}^f$ and $X_{i,t}^v$ represent the long-term trend due to radiative forcing and internal variability, respectively. The decomposition is done by first estimating $X_{i,t}^f$ with a simple regression of $X_{i,t}$ on the equivalent CO_2 concentration, as explained in Sect. 3.1 below. The residuals of the linear regression are taken as $X_{i,t}^v$, which are then modelled using either an autoregressive model or a constructed analogue, as described in Sects. 3.3 and 3.4.

3.1 Modelling the forced trend (Trend)

The time series of global mean equivalent CO_2 concentration is used to represent the variation of external radiative forcing. We then use the following linear regression model to estimate the response of SSTs in each grid box to this forcing:

$$X_{i,t} = \alpha_0 + \alpha_1 C_{t-1} + X_{i,t}^v, \tag{2}$$

Fig. 1 **a** Parameter estimate of α_1 in the forced trend regression model (Eq. 2) fitted to data from all available years, representing the dependence of SST on equivalent CO_2 concentration (in °C per 100 ppm); **b** Estimate of the autoregression coefficient γ_1 in the AR1 model (Eq. 4) fitted to data from all available years

where C_{t-1} is the equivalent CO_2 concentration for the previous year. The first two terms on the right hand side of (2) represent the forced component of observed SSTs, $X_{i,t}^f$. The residuals of this regression model, or the component of the SST time series not explained by the CO_2 concentration time series, are taken as the internal variability. These residuals are assumed to be normally distributed with zero mean and constant variance. For each forecast, this model is fitted to data from all available years excluding the 10 years to be predicted. This simple regression model has two main assumptions: SSTs at each location are only dependent on the equivalent CO_2 concentration with a lag time of 1 year and such dependence is linear. Although a logarithmic dependence might be more physically plausible, a linear approximation is considered adequate for the range of equivalent CO_2 concentration variation in the past 140 years. A similar regression approach was adopted by Krueger and von Storch (2011) to estimate the forced component of surface air temperatures, but with the multi-model ensemble mean temperatures from historical simulations as the response variable. In contrast, our model uses the observed SSTs and thus avoids the inclusion of climate model data in our statistical prediction scheme. Using estimates of the parameters, $\hat{\alpha}_0$ and $\hat{\alpha}_1$, the forced component of SST for a lead time of k years is predicted by

$$\hat{X}_{i,t+k}^f = \hat{\alpha}_0 + \hat{\alpha}_1 C_{i,t+k-1}. \qquad (3)$$

Figure 1a shows estimates of the slope parameter (α_1) for the model fitted to data from all available years (1871 to 2011 for SSTs; 1870 to 2010 for equivalent CO_2). For most places, SST increases with equivalent CO_2 concentration, with the notable exceptions of the far North Atlantic and parts of the tropical Pacific where the estimated α_1 is close to zero or even slightly negative. Note that among individual retrospective forecasts, there is a small variation in the magnitude (or even a change in the sign for the above two regions) of $\hat{\alpha}_1$; nevertheless Fig. 1a represents the general spatial pattern.
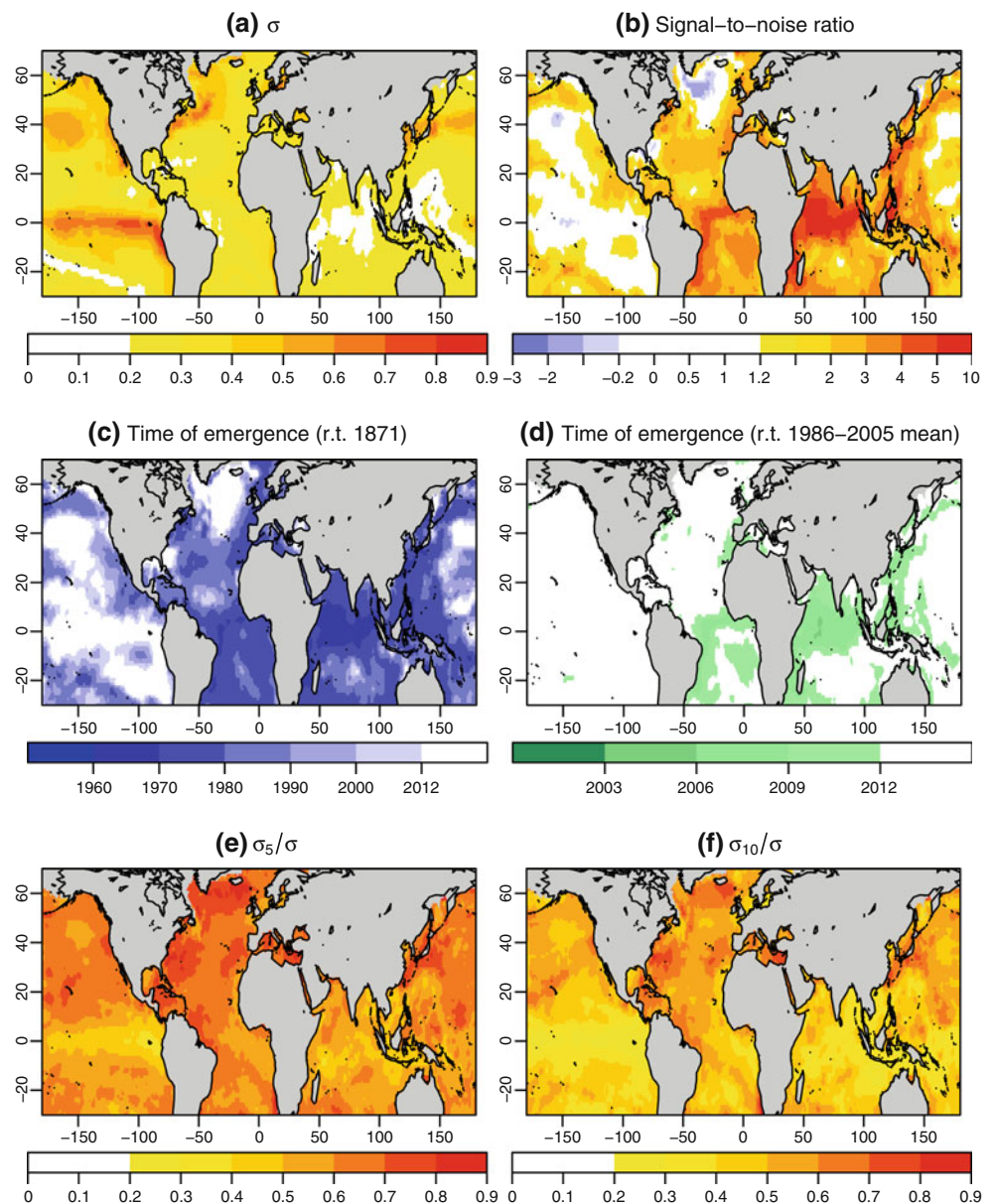
3.2 Comparing the forced trend and internal variability

Before describing how the internal variability component X^v is modelled, we first examine some of its features. Figure 2a shows the standard deviation of the residuals of model (2) (denoted by σ) fitted to data from all available years, and Fig. 2b shows the ratio of the estimated forced change over the period ($\hat{X}_{i,2011}^f - \hat{X}_{i,1871}^f$) to σ. The latter may be interpreted as the 'signal-to-noise' ratio of SSTs over the 141 years, with larger magnitudes indicating greater dominance of forced long-term change to internal variability. Over parts of the western Pacific, the Indian Ocean and the equatorial Atlantic, the forced change in SST is around five times larger than σ, therefore the effect of forced long-term change is dominant. The R-squared statistics of model (2) (not shown) indicate that the equivalent CO_2 concentration time series explain more than 50 % of the variance in SSTs in these regions. In contrast, over the tropical Pacific where the forced long-term change is small and σ is relatively large, the internal variability dominates. Less than 10 % of the variance in SSTs can be explained by model (2) (not shown).

A related variable, the time at which the climate change signal emerges from the interannual variability, is often of interest in climate change risk assessment. A number of studies (e.g. Giorgi and Bi 2009; Mahlstein et al. 2011; Diffenbaugh and Scherer 2011; Hawkins and Sutton 2012) have examined this 'time of emergence' for future climate change using climate model projections, but here we can assess when and where the signal has already emerged in the historical SST observations. Figure 2c shows the time of emergence defined as the year when the ratio of the magnitude in estimated forced change from the year 1871 (pre-industrial time) to σ last crossed and exceeded the threshold of one in the time series.[1] The climate signal

---

[1] The time of emergence may also be defined as the year when the ratio *first* crossed and exceeded the threshold of one. We use a more conservative measure here as the temporary reduction in equivalent CO_2 concentration in the 1950s led to a decrease in climate change signal in some places.

Fig. 2 **a** Standard deviation of X^v estimated as the residuals of the forced trend regression model (2) (in °C); **b** Ratio of estimated forced change in SST (signal) to the standard deviation of residuals (noise) of model (2); **c** and **d** Time of emergence of climate signal, defined as the year when the signal-to-noise ratio last crossed and exceeded one, relative to (r.t.) two reference baselines (1871 and 1986–2005 mean); **e** and **f** Potential predictability as a ratio of the standard deviation of 5 and 10 year means to the standard deviation of interannual variability in SSTs after the forced trend is removed

emerged over parts of the Indian Ocean in the 1960s, while most of the Atlantic and western Pacific saw emergence in the 1970s and 1980s. As the forced trend was small over the far North Atlantic and eastern Pacific, the climate signal remained smaller than the interannual variability throughout the 141 years. Figure 2d also shows the time of emergence but considers the forced change from the 1986–2005 mean. Even with this more recent baseline, the climate change signal has already emerged in tropical Indian Ocean and tropical Atlantic in the 2000s.

Another feature worth exploring is the importance of longer timescale variability relative to interannual variability. This is because the former is related to slow ocean processes and is considered to be, at least potentially, more predictable. We consider a simple measure of the potential predictability (Boer 2004, 2011; Boer and Lambert 2008) of X^v for a timescale of N years, σ_N/σ, where σ_N is the standard deviation of the running N-year mean of estimated X^v. Figure 2e, f show this measure for $N = 5$ and 10 years. Larger potential predictability can be found over the most of the North Atlantic. Previous observational and modelling studies have suggested that this is related to decadal variations in the North Atlantic thermohaline circulation and the North Atlantic Oscillation (e.g. Latif et al. 2006). The potential predictability is smaller over the tropical Pacific where ENSO-related interannual variability dominates.

For the rest of this paper we will put greater emphasis on the Atlantic sector for two reasons. First, the gridded historical observations in the Atlantic are more reliable

because there were more available in-situ observations in the region (Rayner et al. 2003). Second, over the Atlantic there is spatial variation in the relative importance of forced long-term trend and internal variability. In the far North Atlantic where the internal variability is more important, the longer timescale variability is also more potentially predictable. This constrasts with the Indian Ocean and western Pacific where the effect of long-term forced trend dominates and the forced trend prediction described in Sect. 3.1 is perhaps adequate for achieving reasonable predictive skill.

3.3 First-order autoregressive model (AR1)

In the first model for the internal variability component of SSTs, X^v for each grid box is modelled as a first-order autoregressive process (AR1):

$$X_{i,t}^v = \gamma_0 + \gamma_1 X_{i,t-1}^v + Z_{i,t}, \tag{4}$$

where $Z_{i,t}$ represents a purely random process. To ensure that the same length of time series is always used in the model training, the AR1 model is fitted to a total of 90 years of SST data (the start year of each forecast plus the previous 89 years). Autoregressive models of higher orders (e.g. AR2) were considered, but there was no significant improvement in predictive skill (not shown). With the estimated parameters $\hat{\gamma}_0$ and $\hat{\gamma}_1$, X^v at a lead time of k years is predicted iteratively:

$$\hat{X}_{i,t+k}^v = \begin{cases} \hat{\gamma}_0 + \hat{\gamma}_1 X_{i,t}^v & \text{for } k = 1 \\ \hat{\gamma}_0 + \hat{\gamma}_1 \hat{X}_{i,t+k-1}^v & \text{for } k = 2,\ldots,10. \end{cases} \tag{5}$$

Figure 1b shows estimates of γ_1 in the AR1 model fitted to data from all available years. The highest autoregression coefficients can be found in the North Atlantic subpolar gyre, indicating stronger persistence in the SST time series (or memory at a longer timescale; see also Zhu et al. 2010). The effect of persistence is generally weaker in the equatorial regions. Similar results can be seen for the AR1 models fitted for the individual retrospective forecasts (not shown).

3.4 Constructed analogue (CA)

The use of a second and more complex model for predicting X^v, constructed analogue (CA; van den Dool 2007, Chap. 7), is motivated by Hawkins et al. (2011) which employed this model for decadal predictions of SSTs using control integrations of two GCMs. It was found to be skillful over the far North Atlantic. The CA method has also been employed in operational seasonal SST forecasts for the Pacific (Barnston et al. 1994; Landsea and Knaff 2000) and seasonal predictions of soil moisture

(van den Dool et al. 2003). The rationale behind CA is to develop a weighted, linear combination of historical spatial patterns of observations which is closest to the initial spatial pattern. If the future evolution of such patterns resembles the historical evolution, the CA can make predictions by carrying forward the estimated weights. An analogue needs to be constructed because there is only a remote chance of finding a 'natural' analogue where the error between a state in the historical record and the desired initial state is within observational errors (van den Dool 1994).[2]

In this study we apply the CA method for the Atlantic sector (30°S–70°N, 100°W–20°E) only. For each forecast and for each lead time (k), the field of the internal variability component of SSTs X^v at start year t is constructed using a linear combination of X^v in the previous 89 years:

$$\mathbf{X}_t^v = \sum_{p=k}^{89} \beta_p \mathbf{X}_{t-p}^v, \tag{6}$$

where β_p are the weights. The weights are estimated by minimising

$$\sum_i^M \left(\tilde{\mathbf{X}}_{i,t}^v - \sum_{p=k}^{89} \beta_p \tilde{\mathbf{X}}_{i,t-p}^v \right), \tag{7}$$

where $\tilde{\mathbf{X}}^v$ are reconstructed fields of X^v using empirical orthogonal functions (EOFs), and M is the total number of grid boxes. This is analogous to estimating the coefficients in a multiple linear regression by minimising the sum of squared residuals on the previous 89 years of data, but here ridging (Draper and Smith 1998, Chap. 17) needs to be used to circumvent problems of an underdetermined system. The reason for estimating the weights in the EOF truncated space is to reduce the effects of noise. van den Dool (2007, Chap. 7) suggests the number of EOFs to be used should be half the number of training data length, which is around 45 in our case. On average about 99 % of the variance of $\mathbf{X}^v$ is explained using 45 EOFs. This gives minimal reduction in noise, and the CA weights are found to be unstable (not shown). After testing the effects of the number of EOFs on the stability of weights and predictive skill, we have chosen to use 9 EOFs, which explain on average about 80 % of the $\mathbf{X}^v$ variance. We note that making such a choice based on predictive skill may lead to some overestimation of skill in our verification.

There are two differences between our CA and AR1 models. First, a single CA predicts the whole Atlantic field of $\mathbf{X}^v$ for a particular year, while individual AR1 models are fitted to the time series of $\mathbf{X}^v$ for each grid box. Second, we use a direct prediction approach for the CA model which is fitted individually for each lead time, while for

---

[2] For example, Hawkins et al. (2011) estimated that for Atlantic SSTs, one would need 10^5 years of data to find a natural analogue.

AR1 predictions are made iteratively where the prediction for lead time $k + 1$ depends on the prediction for lead time k. Also, for CA predictions with lead time $k > 1$ year, $\mathbf{X}^v$ between years t and $t - k$ are not included in the construction. This avoids the use of 'future' data beyond year t in the prediction stage, where $\mathbf{X}^v$ fields are predicted by

$$\hat{\mathbf{X}}_{t+k}^v = \sum_{p=0}^{89-k} \hat{\beta}_{p+k} \mathbf{X}_{t-p}^v. \tag{8}$$

3.5 Verification for retrospective forecasts

We have seen how the forced trend (X^f) and the internal variability (X^v) components of SSTs are modelled and predicted. The predictions of SST anomalies ($\hat{X}$) are either the sum of predicted X^f and predicted X^v using AR1 [i.e. (3) + (5)] or the sum of predicted X^f and predicted X^v using CA [i.e. (3) + (8)]. We term these two anomaly predictions as 'Trend+AR1' and 'Trend+CA' respectively.

The three key questions in the evaluation of prediction skill of statistical SST retrospective forecasts are: (1) Are the retrospective statistical forecasts of the forced trend and internal variability skillful? (2) How does the skill vary spatially and with prediction lead times? (3) Where and for what lead times are initialised dynamical retrospective forecasts by DePreSys more skillful than the corresponding benchmark statistical forecasts? We now describe how these questions are to be addressed.

We mainly consider two skill measures, anomaly correlation coefficient (ACC) and mean squared skill score (MSSS), to verify up to 46 sets of statistical retrospective forecasts by Trend+AR1 and Trend+CA.[3] The ACC considers the prediction and observation of the anomaly ($\hat{X}$ and X). Since the ACC does not inform about whether the skill comes from the forced trend or internal variability (or both), we also consider the correlation coefficient between the prediction of the forced component ($\hat{X}^f$; 'Trend' prediction) and the observed anomaly (X) and the correlation coefficient between the predicted internal variability component ($\hat{X}^v$) and the estimated internal variability component of the verifying observations (X^v). Note that there are two major limitations of correlation coefficients in verification. They measure only the linear association between predictions and observations, and any conditional biases in the predictions are ignored (Murphy and Epstein 1989).

MSSS is a dimensionless measure of the improvement in mean squared error (MSE) of a prediction system relative to a reference prediction system. In the verification of our statistical forecasts, the MSSS of anomaly predictions

relative to Trend predictions can be used to evaluate the skill of our models in predicting the internal variability. This is analogous to comparing the initialised and uninitialised retrospective forecasts produced by a GCM. For a lead time of k years, we first calculate the MSE for $\hat{X}$ and $\hat{X}^f$ with start year t:

$$\text{MSE}_{\hat{X}} = \frac{1}{N} \sum_{j=0}^{N-1} \left(\hat{X}_{t+k+j} - X_{t+k+j} \right)^2 \tag{9}$$

$$\text{MSE}_{\hat{X}^f} = \frac{1}{N} \sum_{j=0}^{N-1} \left(\hat{X}_{t+k+j}^f - X_{t+k+j} \right)^2, \tag{10}$$

where N is the total number of available retrospective forecasts. Then MSSS of $\hat{X}$ relative to $\hat{X}^f$ is given by

$$\text{MSSS} = 1 - \frac{\text{MSE}_{\hat{X}}}{\text{MSE}_{\hat{X}^f}}. \tag{11}$$

A positive (negative) MSSS indicates that the anomaly predictions are more (less) skillful than the Trend predictions in terms of MSE.

The statistical significance of the correlation measures and MSSS is assessed using a non-parametric bootstrapping approach. In particular, a block bootstrapping technique is used to account for the temporal autocorrelation among successive observations (or forecasts). Details are given in the section Appendix.

For decadal climate predictions, verification is commonly performed on predictions averaged over a range of lead times, but there is no clear consensus on the choice of temporal averaging period (e.g. Smith et al. 2010; Hawkins et al. 2011; van Oldenborgh et al. 2012). We adopt the framework suggested by Goddard et al. (2012a) and consider three lead time periods, year 1, years 2–5 and years 6–9. Note that a skillful time-averaged (multi-annual) prediction does not imply that the predictions for individual years within the averaging period (annual predictions) are skillful, as errors in the individual predictions with different signs may cancel out by averaging.

The verification procedures for the ensemble mean DePreSys retrospective forecasts are similar to those described above but two additional steps are required. Since the spatial resolution of DePreSys and HadISST are different, the HadISST data are first interpolated onto the grid of DePreSys using bilinear interpolation prior to verification. In addition, the mean bias of the ensemble mean DePreSys retrospective forecasts (difference between the modelled and observed climate) is removed for each lead time. In order to avoid the introduction of artificial skill, this is performed in a cross-validation manner, i.e. for each lead time, we calculate and remove the mean bias for each forecast individually using forecast and observation data from all the other start times.

---

[3] Hindcasts with start year from 2002 onwards give predictions beyond year 2011, therefore less than 46 sets of retrospective forecasts can be verified for lead times longer than 6 years.
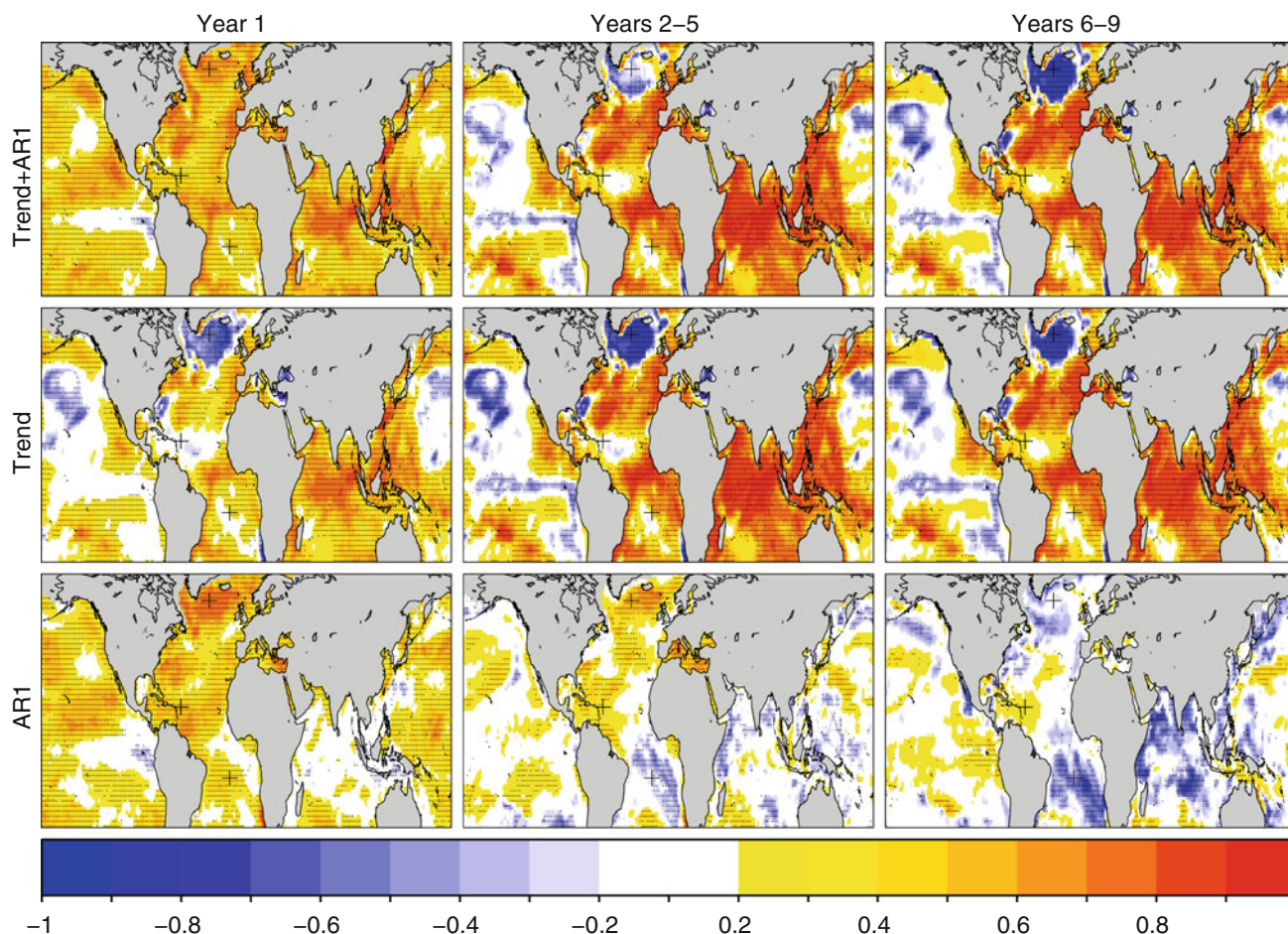
Fig. 3 Correlation measures over different lead time periods for Trend+AR1 and Trend retrospective forecasts with observed anomalies and for AR1 forecasts with the internal variability component of observed anomalies. The forecast time series of grid boxes marked with '+' are shown in Fig. 5. The stippled areas indicate where the correlation measure is significantly different from zero at the 10 % level

4 Skill of the statistical retrospective forecasts

We now assess the skill of retrospective forecasts by our statistical models, Trend+AR1 and Trend+CA. We will also compare our results with some other previous studies on benchmark statistical models reviewed in the Introduction.

4.1 Trend+AR1 forecasts

Figure 3 shows the global maps of correlation measures for forecasts using AR1. For year 1, the largest positive values of ACC for Trend+AR1 forecasts are found over parts of the equatorial and North Atlantic, the Indian Ocean and western Pacific. As we move to years 2–5 and 6–9, the ACC increases to over 0.7 in these areas. The source of skill may be explained by considering the other two correlation measures shown in Fig. 3. The spatial pattern of correlation between the predicted forced trend (Trend) and

observed anomalies matches well with that of Trend+AR1 for these two periods. Meanwhile, the correlation between the predicted and observed internal variability (AR1) are close to zero or even negative for many areas except the far North Atlantic. It appears that the large positive ACC for Trend+AR1 is primarily related to the prediction of the forced trend. The MSSS of Trend+AR1 forecasts relative to the Trend forecasts (Fig. 4) give consistent results. For years 2–5 and 6–9, the MSSS is close to zero for most areas, suggesting that AR1 predictions of the internal variability give little improvement in skill.

For the tropical Pacific, there is little predictive skill for all lead times. Despite large interannual variability (Fig. 2a), the effect of persistence is generally weak (Fig. 1b) as the timescale of ENSO is typically less than 1 year. Both the forced trend and AR1 forecasts for the internal variability do not show obvious skill.

The predictive skill for the far North Atlantic is more interesting. The ACC for Trend+AR1 drops from around
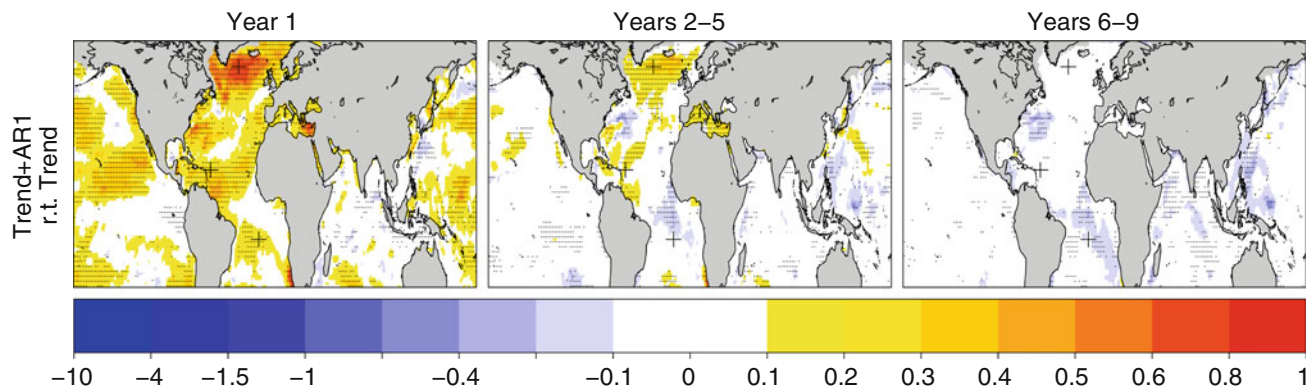
Fig. 4 Mean squared skill score (MSSS) of anomalies predicted by Trend+AR1 relative to the predicted forced trend component (Trend) for different lead time periods. Positive MSSS means that the mean squared error for Trend+AR1 retrospective forecasts is lower. The stippled areas indicate where the MSSS is significantly different from zero at the 10 % level

+0.7 to around −0.8 as we move from year 1 to years 6–9. The correlation for Trend is negative for all three periods, however the correlation for AR1 is greater than 0.5 for year 1 and years 2–5. The MSSS is also significantly greater than zero (at 10 % level; same for below unless otherwise stated) for these two periods. These suggest that the AR1 model gives skillful predictions of the internal variability over the far North Atlantic for shorter lead times, but the poor forced trend prediction offsets such skill. These results are similar to the global decadal air temperature retrospective forecasts by Krueger and von Storch (2011), where the ACC is larger than 0.7 for year 1 but is not significantly greater than zero (at 5 % level) for year 10.

Comparing the time series of observed and predicted SSTs for one of the grid boxes in the far North Atlantic (centred at 59.5°N 32.5°W; marked with a cross in Fig. 3) shown in Fig. 5b can offer further insights. Over the past 140 years, the observed SST (thick black line) has a weak negative long-term trend but large multi-decadal internal variability. Meanwhile, the equivalent CO_2 concentration increased gradually (Fig. 5a). The estimates of the parameter α_1 in the regression model (Eq. 2) for the 46 retrospective forecasts are slightly below zero (refer to Fig. 1a). Therefore with increasing equivalent CO_2 concentration, the predictions have a negative forced trend. This is not obvious in the time series of the year 1 Trend+AR1 forecasts (red line) in Fig. 5a because of strong persistence from the start year with a high autoregression coefficient (refer to Fig. 1b). The impact of the forced negative trend is much clearer in the time series of the year 8 forecasts (blue line). However, during the verification period (mid-1960s to mid-2000s), the SST increased by about 1 K. As a result, the correlation for Trend prediction is negative and the ACC decreases sharply with prediction lead time because the compensating skill from persistence by AR1 diminishes.
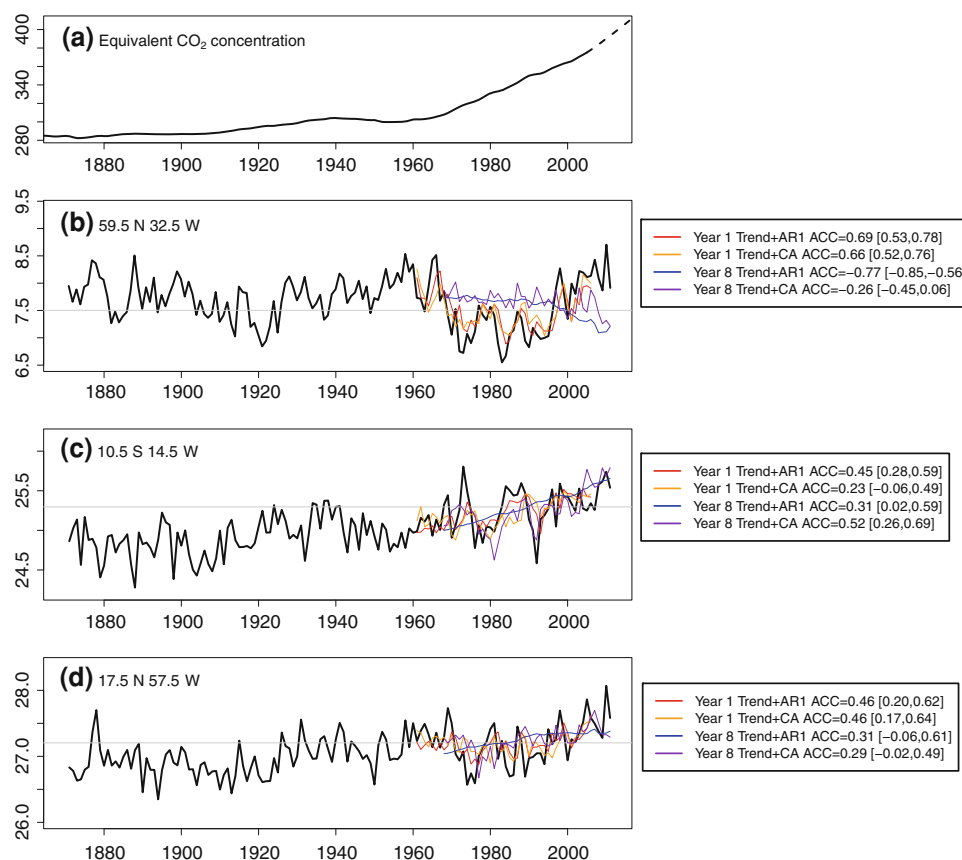
4.2 Trend+CA forecasts

We now consider the skill of forecasts using CA for the Atlantic using the correlation measures and MSSS shown in Figs. 6 and 7 respectively. The highest values of ACC for Trend+CA are found over parts of mid-latitude and equatorial Atlantic, especially for longer lead times. We have already seen that in mid-latitude and equatorial Atlantic the correlation coefficient for Trend prediction is also high. Since the correlation for internal variability by CA is generally weak, the predictive skill is therefore mainly related to the forced trend prediction like Trend+AR1. The MSSS of the Trend+CA forecasts relative to the Trend forecasts is significantly negative in mid-latitude and equatorial Atlantic, suggesting that the predictions of internal variability by CA actually increases the error.

There is quite strong evidence that CA is skillful in predicting the internal variability in the far North Atlantic, and performs considerably better than the simple AR1 model. Significant positive correlation coefficients for the internal variability (greater than 0.5) can be seen for all three lead time periods, while the MSSS also shows that the Trend+CA forecasts are significantly superior to the Trend forecasts even for years 6–9. These explain the changes in MSSS with lead time of Trend+CA relative to Trend+AR1. Both methods are similarly skillful for year 1, but Trend+CA has a significant advantage at extended lead times. Similar to Trend+AR1, the ACC for Trend+CA in the far North Atlantic changes from positive to negative as we move from year 1 to years 6–9 because the predicted forced trend is not consistent with observations within the verifying period. However, the drop in magnitude for Trend+CA is smaller as CA better predicts the internal variability.

Another region worth noting is the South Atlantic. There is some evidence that Trend+CA gives skillful predictions

Fig. 5 a Time series of global mean equivalent CO_2 concentration (in ppm) from 1870 to 2011, with the portion with dashed line indicating projections based on RCP 4.5; **b–d** Time series of observed SSTs (in °C) for three grid boxes (*thick black solid lines*) in the Atlantic. The *thin grey horizontal line* indicates the 1986–2005 mean. Retrospective forecasts by Trend+AR1 and Trend+CA for lead times of 1 and 8 years are overlaid with thin lines of different colours as indicated in the legend. The ACC for such forecasts and the corresponding 90 % bootstrap confidence interval (in parentheses) are also given in the legend

at longer lead times in parts of this region, even though we have seen that the South Atlantic is less potentially predictable than the far North Atlantic (Fig. 2e, f). From Fig. 1a the dependence of SSTs on radiative forcing appears to be stronger near the coast than in the central parts of the South Atlantic. This might explain the higher correlation coefficients for the Trend forecasts nearer to the coast. As for the CA predictions of internal variability, the correlation coefficients are rather low for year 1, but increase gradually as we move to longer lead time periods. This is consistent with the positive MSSS of Trend+CA relative to both Trend and Trend+AR1 forecasts, especially for years 6–9 with a large area of significantly positive MSSS (Fig. 7). Figure 5c shows the time series of observations and retrospective forecasts for a grid box in South Atlantic (centred at 10.5°S 14.5°W; also marked with a cross in Figs. 6, 7). The positive trend is well predicted. In addition, the Trend+CA forecasts for year 8 (purple line) appear to capture the interannual variability quite well, particularly from around 1980 to 2000.

Over parts of the tropical Atlantic, CA appears to have some predictive skill for the internal variability as well. The correlation coefficient for the internal variability exceeds 0.4 for year 1 and years 2–5. The MSSS of Trend+CA relative to both Trend and Trend+AR1 also shows that the CA is more skillful for these two
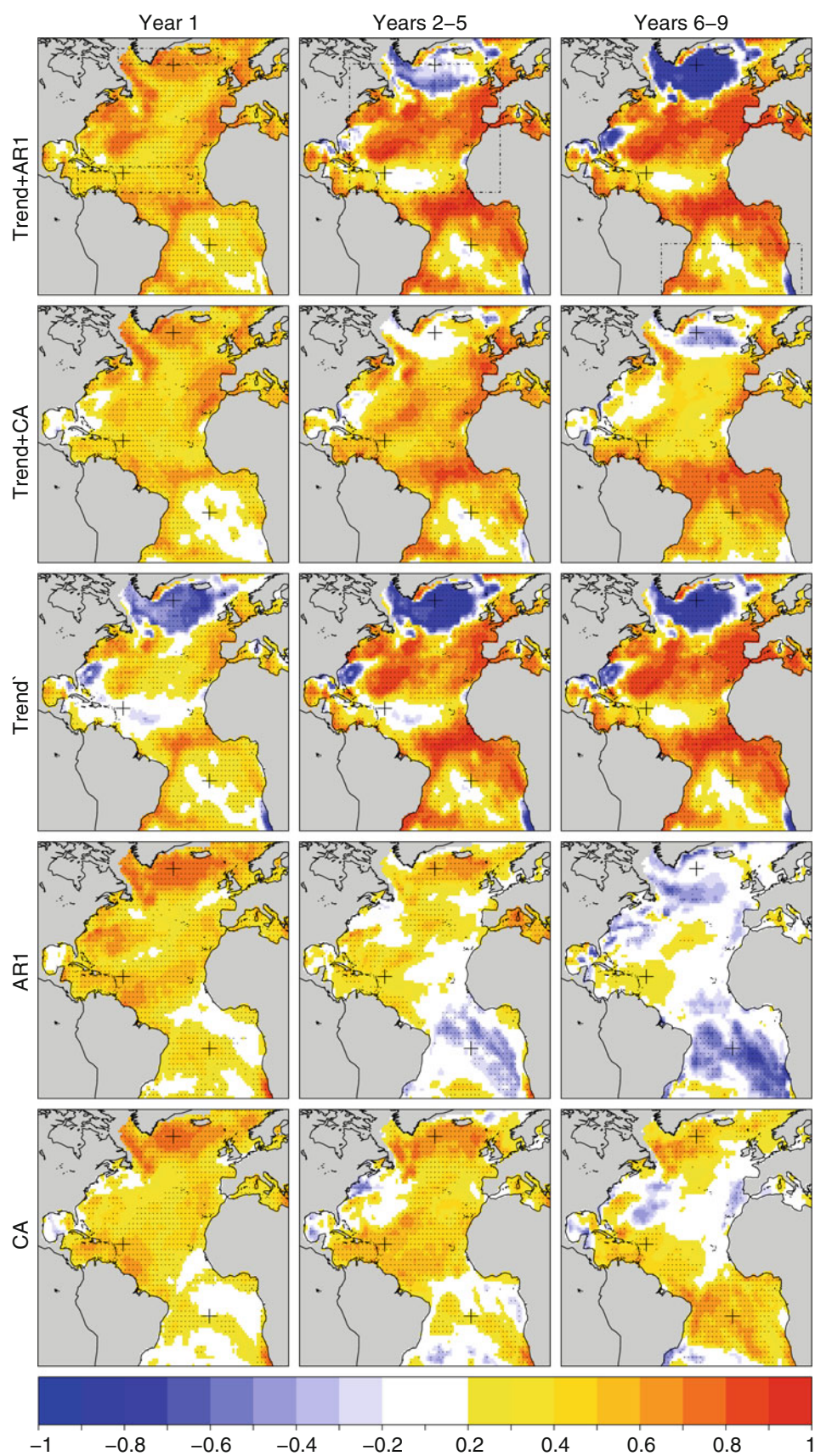
periods and perhaps even for years 6–9. However, comparing the time series of observations and year 8 forecasts for the grid box centred at 17.5°N 57.5°W in Fig. 5d, Trend+CA does not seem to have a clear advantage over Trend+AR1.

Our results are generally consistent with Hawkins et al. (2011). Their verification of retrospective forecasts by CA using the control integration of HadCM3 also gave ACC greater than 0.5 in the far North Atlantic for all lead time periods (years 1, 2, 3–5 and 6–10). However, while we have seen positive correlation coefficients for internal variability in the South Atlantic at extended lead times for our forecasts, the ACC for their CA forecasts in the region was close to zero.

4.3 Regional average retrospective forecasts

We have so far examined the skill of retrospective forecasts at a spatial scale identical to the resolution of HadISST data set (1°). It is often useful to consider also the skill of forecasts for larger regions. This is partly because certain regions are particularly important for the development of weather systems, for example tropical cyclone activity is strongly related to the atmospheric and oceanic conditions in the tropics (Goldenberg et al. 2001). Another reason is that it will be easier to assess the predictive skill of large

Fig. 6 As in Fig. 3 but for forecasts for the Atlantic for different lead time periods. The *boxes* on the first row of maps indicate the four regions mentioned in Tables 1, 2 and Fig. 8. The first, third and fourth rows are zoomed-in versions of the maps in Fig. 3
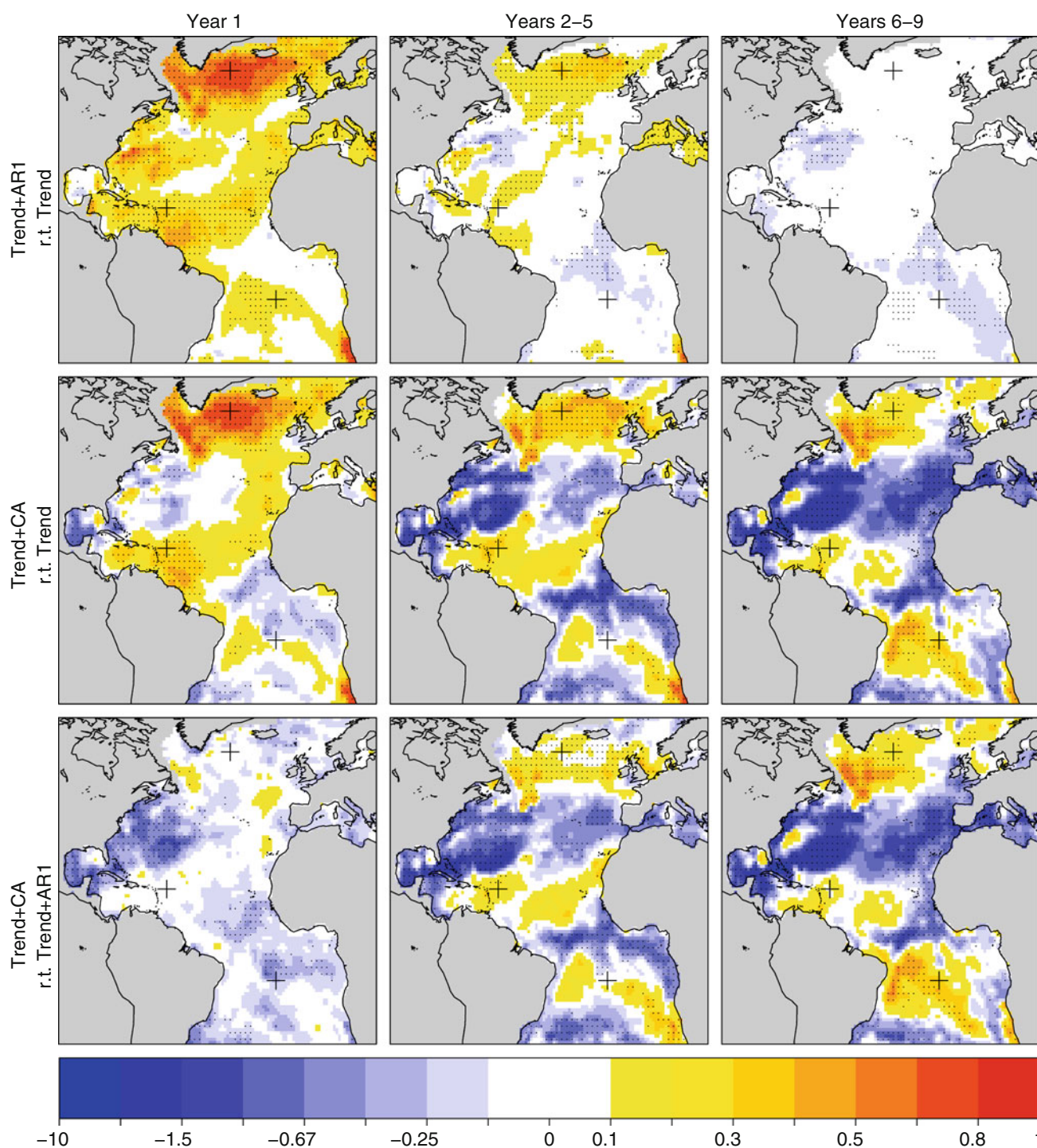
Fig. 7 As in Fig. 4 but comparing Trend+CA, Trend+AR1 and Trend forecasts for the Atlantic for different lead time periods

scale variability by considering regional average forecasts, because the local scale variability will be smoothed out (Goddard et al. 2012a). Here we consider four regions in the Atlantic: subpolar gyre (SPG), hurricane main development region (MDR), North Atlantic (NATL) and South Atlantic (SATL) (see Table 1 for definitions). The SSTs in MDR are directly relevant to the formation of hurricanes,

while a modelling study by Smith et al. (2010) found evidence of remote influence of temperatures in the SPG region on atmospheric conditions in MDR, which in turn affects hurricane activity. The results for NATL and SATL can give a general indication of skill for different statistical models over large areas. For each of the four regions, we average the observations and predictions for Trend,

Table 1 Correlation measures for predictions in four specified regions for different lead time periods

Model	Year 1	Years 2–5	Years 6–9
Subpolar gyre (SPG; 60–66°N, 10–60°W)			
Trend	−0.58 [−0.73,−0.22]	−0.64 [−0.83,−0.13]	−0.39 [−0.72,0.34]
Trend+AR1	0.65 [0.42,0.75]	−0.09 [−0.48,0.27]	−0.70 [−0.84,0.01]
Trend+CA	0.66 [0.40,0.77]	0.27 [0.01,0.43]	0.15 [−0.27,0.48]
DePreSys	**0.77** [0.55,0.85]	**0.84** [0.58,0.92]	**0.80** [0.38,0.89]
Main development region (MDR; 10–20°N, 20–80°W)			
Trend	0.06 [−0.26,0.39]	0.24 [−0.14,0.58]	0.40 [−0.05,0.73]
Trend+AR1	0.49 [0.20,0.67]	0.30 [−0.18,0.56]	0.42 [−0.03,0.73]
Trend+CA	0.44 [0.12,0.64]	0.53 [0.16,0.74]	0.47 [0.15,0.65]
DePreSys	**0.63** [0.37,0.79]	**0.61** [0.16,0.79]	**0.54** [0.09,0.76]
North Atlantic (NATL; 10–60°N, 0–75°W)			
Trend	0.43 [0.10,0.69]	0.59 [0.34,0.80]	0.68 [0.45,0.86]
Trend+AR1	0.87 [0.66,0.93]	0.78 [0.52,0.89]	0.73 [0.51,0.90]
Trend+CA	0.85 [0.65,0.92]	0.71 [0.33,0.85]	0.57 [0.15,0.76]
DePreSys	**0.89** [0.68,0.93]	**0.91** [0.75,0.96]	**0.89** [0.71,0.95]
South Atlantic (SATL; 10–30°S, 50°W–20°E)			
Trend	0.45 [0.23,0.65]	**0.65** [0.41,0.83]	0.60 [0.36,0.78]
Trend+AR1	0.47 [0.27,0.61]	0.64 [0.40,0.83]	0.60 [0.38,0.79]
Trend+CA	0.41 [0.18,0.57]	0.58 [0.32,0.80]	**0.65** [0.45,0.79]
DePreSys	**0.56** [0.35,0.72]	0.43 [0.08,0.67]	0.32 [0.07,0.51]

These four regions are marked by boxes with dashed lines in the top row of Fig. 6. The largest correlation coefficient for each lead time period in each region is shown in bold. The 90 % bootstrap confidence intervals for the correlation measures are shown in parentheses

Trend+AR1 and Trend+CA for all the grid boxes. The verification metrics are then calculated.

Tables 1 and 2 show the correlation measures and root-mean-squared error (RMSE),[4] respectively for the four regions. For SPG, Trend+CA is the best performing model with the highest ACC and the lowest RMSE for all three lead time periods. The improvement in these two measures from the Trend prediction, especially the significant improvement in correlation, confirms the skill of CA in predicting the internal variability at extended lead times. The ACC for Trend+CA is also greater than that for Trend+AR1 for years 6–9, but the improvement in RMSE is small. Figure 8a shows a selection of retrospective forecasts at different start years (indicated by thin lines of different colours) for SPG. The Trend+CA forecasts (solid lines) match the verifying observations slightly better than the Trend+AR1 forecasts (dashed lines), especially in the earlier part of the verification period.

For MDR, Trend+AR1 is the best performing model for year 1, while Trend+CA is the best for years 2–5 and 6–9, however the improvement in the verification metrics from the Trend prediction and Trend+AR1 to Trend+CA is smaller compared to SPG. Trend+AR1 appears to be

the best for all lead times for NATL. For SATL, Trend+CA performs better than the other models at longer lead times.

We note that another possible method to obtain regional average forecasts is to first average, year by year, the historical observations for all grid boxes in the region and then make forecasts from the spatially averaged annual mean SSTs as described in Sects. 3.1 and 3.3. This alternative method gives generally similar results.

5 Comparison with DePreSys retrospective forecasts

In order to illustrate the use of our statistical models as benchmarks to verify decadal predictions by initialised dynamical GCMs, we now compare the performance of mean bias corrected ensemble mean retrospective forecasts by DePreSys with forecasts by Trend+AR1 globally and Trend+CA for the Atlantic using MSSS (Fig. 9). The detailed evaluation of DePreSys, such as the investigation of sources of its prediction skill and the verification of probabilistic predictions by its ensemble, will be done in a future analysis.

For year 1, DePreSys significantly outperforms Trend+AR1 over large parts of the tropical Pacific, which is likely to be related to its skill in predicting ENSO at the

---

[4] We use the RMSE, the square root of MSE here as the RMSE is intuitively easier to interpret.
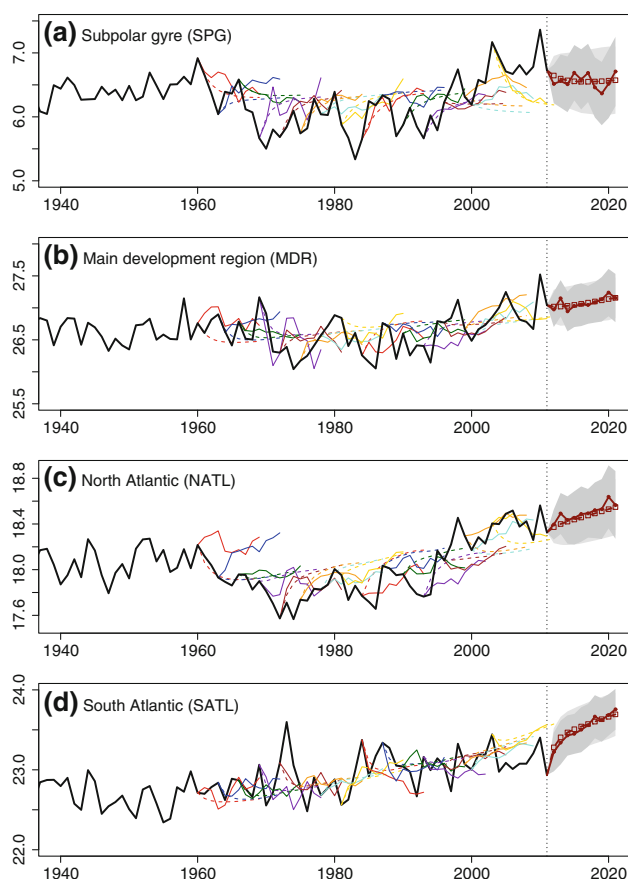
Fig. 8 Time series of observations (in °C; *thick black solid lines*) averaged over four specified regions (refer to Table 1 for definitions). Statistical retrospective forecasts for selected start times are overlaid using lines of different colours—*dashed lines* for Trend+AR1 and solid lines for Trend+CA. Each *line* starts at the observed SST at the start year of the forecast and ends at predicted value for year 10. The forecasts which start at year 2011 (for years 2012–2021) are shown by *thick dark red lines*, with *dots* and *squares* indicating the predicted values for each individual year. The *light grey* and *dark grey* shadings indicate the average root-mean-squared error (RMSE) for the retrospective forecasts of Trend+AR1 and Trend+CA respectively

Table 2 As in Table 1 but for root-mean-squared error (RMSE; in *K*)

Model	Year 1	Years 2–5	Years 6–9
Subpolar gyre (SPG; 60–66°N, 10–60°W)			
Trend	0.47 [0.39,0.53]	0.44 [0.36,0.51]	0.46 [0.36,0.50]
Trend+AR1	0.31 [0.26,0.33]	0.38 [0.31,0.44]	0.46 [0.36,0.50]
Trend+CA	0.30 [0.25,0.33]	0.37 [0.31,0.41]	0.42 [0.33,0.47]
DePreSys	**0.26** [0.23,0.29]	**0.21** [0.16,0.24]	**0.22** [0.19,0.26]
Main development region (MDR; 10–20°N, 20–80°W)			
Trend	0.32 [0.28,0.36]	0.23 [0.20,0.27]	**0.23** [0.19,0.27]
Trend+AR1	0.26 [0.23,0.30]	0.23 [0.19,0.26]	**0.23** [0.19,0.28]
Trend+CA	0.27 [0.24,0.30]	**0.21** [0.18,0.23]	**0.23** [0.19,0.27]
DePreSys	**0.25** [0.20,0.30]	**0.21** [0.17,0.25]	0.24 [0.19,0.29]
North Atlantic (NATL; 10–60°N, 0–75°W)			
Trend	0.23 [0.19,0.26]	0.21 [0.18,0.24]	0.22 [0.18,0.25]
Trend+AR1	0.14 [0.12,0.16]	0.19 [0.16,0.21]	0.22 [0.19,0.25]
Trend+CA	**0.13** [0.11,0.15]	0.19 [0.16,0.21]	0.23 [0.18,0.28]
DepreSys	**0.13** [0.11,0.14]	**0.12** [0.10,0.14]	**0.14** [0.11,0.16]
South Atlantic (SATL; 10–30°S, 50°W–20°E)			
Trend	0.26 [0.20,0.32]	0.18 [0.14,0.23]	0.18 [0.13,0.23]
Trend+AR1	0.24 [0.20,0.28]	0.18 [0.14,0.22]	0.19 [0.14,0.23]
Trend+CA	0.24 [0.20,0.29]	0.18 [0.12,0.22]	**0.16** [0.12,0.18]
DePreSys	**0.22** [0.18,0.26]	**0.16** [0.13,0.19]	**0.16** [0.13,0.18]

The lowest RMSE for each lead time period in each region is shown in bold

Tables 1 and 2 also show the ACC and RMSE of the regional average DePreSys retrospective forecasts for the four regions we considered in Sect. 4.3. DePreSys generally ourperforms the statistical models, especially for the SPG where for years 2–5 and 6–9, it has significantly higher ACC and lower RMSE than the best statistical model, Trend+CA. The results for SATL are different. Although the RMSE for DePreSys is the lowest for all three lead time periods, Trend and Trend+CA have higher ACC for years 2–5 and years 6–9 respectively.

6 Statistical forecast for 2012–2021

We finally give an overview of the forecasts for future Atlantic SST anomalies by Trend+AR1 and Trend+CA for three periods, 2012, the 2013–2016 average and the 2017–2020 average (Fig. 10). For the start year of this forecast, 2011, warm anomalies of less than 0.5 K (relative to the 1986–2005 mean) were observed in the SPG and the tropical North Atlantic. There is generally good agreement between the forecasts by Trend+AR1 and Trend+CA. Both predict the far North Atlantic SSTs to decrease gradually, but this should be viewed cautiously given the rather poor skill in predicting the past forced trend in this region. Meanwhile, the warm anomalies in the NAC region
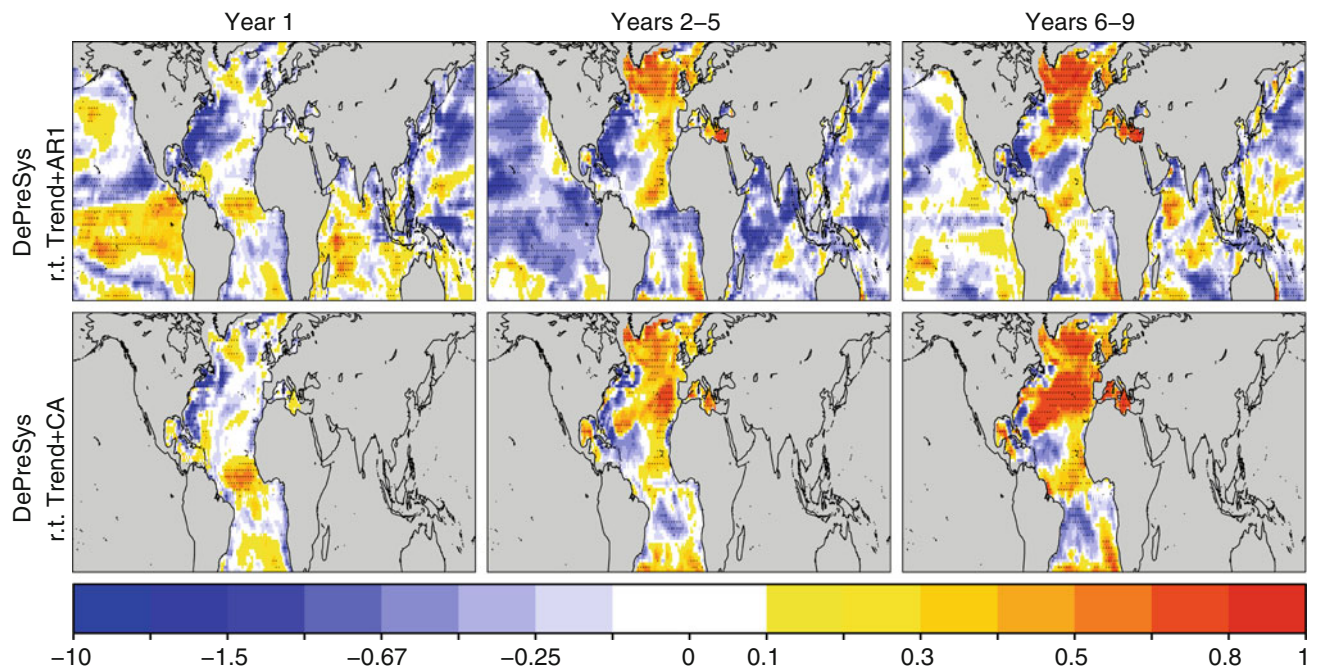
seasonal timescale. It also performs better than Trend+AR1 for parts of the Indian Ocean and better than both statistical models over parts of the tropical Atlantic. However, both Trend+AR1 and Trend+CA have clear advantages over DePreSys in the NAC region.

A different pattern evolves at longer lead times. The advantages of DePreSys in the tropical Pacific diminish, and DePreSys performs worse than Trend+AR1 for most parts of the Pacific and the Indian Ocean for years 2–5. On the other hand, DePreSys is clearly superior to Trend+AR1 over large parts of north Atlantic including the SPG for years 2–5 and 6–9. Similar results can be seen for the comparison between DePreSys and Trend+CA. For parts of the MDR and South Atlantic, however, Trend+CA performs significantly better than DePreSys.

Fig. 9 As in Fig. 4 but comparing DePreSys with Trend+AR1 and Trend+CA
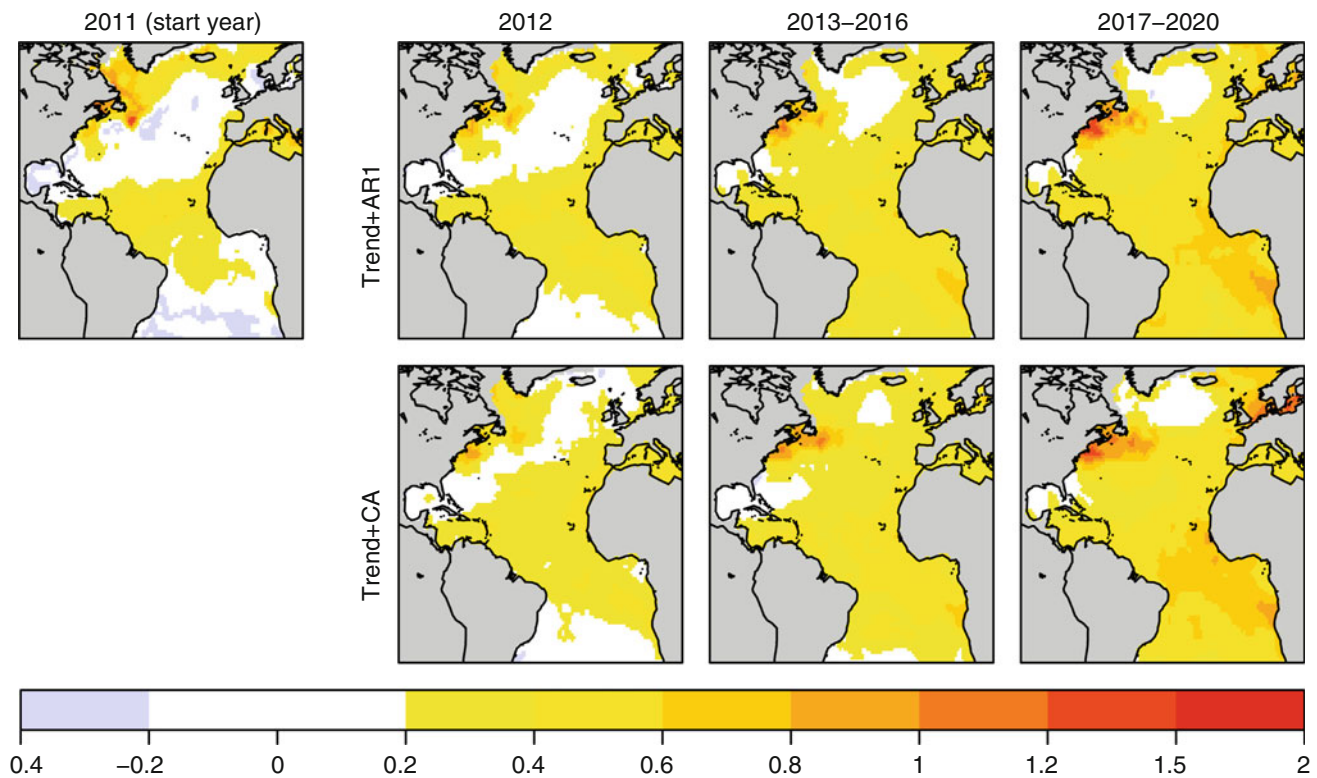

Fig. 10 Forecast anomalies (in °C, relative to the 1986–2005 mean) of annual mean Atlantic SSTs for 2012, the 2013–2016 average and the 2017–2020 average by Trend+AR1 and Trend+CA. The observed anomalies for 2011 are also shown as a comparison

are predicted to persist. As the rise in equivalent CO_2 concentration is projected to continue, both models predict the SSTs in other parts of the Atlantic to become warmer.

The regional average forecasts by Trend+AR1 and Trend+CA are shown by thick dark red dashed and solid lines respectively in Fig. 8. To give an indication of

possible errors in these forecasts, the corresponding average RMSE of the retrospective forecasts for different lead times are shown by the grey shadings. As in what we have seen in the forecast maps (Fig. 10), there is no major disagreement between forecasts by Trend+AR1 and Trend+CA. A slow cooling trend is predicted for SPG, while NATL and SATL are expected to have a warming trend. In particular, the magnitude of predicted rise in the average SST in SATL (about 0.5 K) is slightly larger than the average historical RMSE (about 0.3 K).

We note that the above forecasts are based on projected equivalent CO_2 concentration for the RCP4.5 concentration pathway. Forecasts using alternative pathways (e.g. RCP2.6 and RCP8.5) were considered. Since the difference in equivalent CO_2 concentration among various pathways are small (less than 10 ppm difference from RCP4.5), the forecasts are generally similar to those presented above (not shown).

7 Conclusions

With their potential benefits to climate change adaptation planning, accurate decadal climate predictions made by initialised dynamical GCMs are required. In this paper, we have presented two statistical models for the decadal prediction of SSTs which may be used as benchmarks in the evaluation of initialised GCM predictions. The two statistical models apply the same linear regression of SSTs on equivalent CO_2 concentration to model the effects of long-term trend due to radiative forcing, but use different methods to model the residual internal variability. The verification of retrospective forecasts by these two models give encouraging results. The main findings of this paper are:

- Both the simple AR1 model and the more complex CA model provide skillful predictions of the internal variability in the far North Atlantic for years 2–5, and the skill of CA extends to years 6–9.
- CA gives skillful predictions of the internal variability for parts of the South Atlantic for years 6–9.
- Although DePreSys, an initialised GCM, performs significantly better than the statistical models for most parts of the North Atlantic at extended lead times, Trend+AR1 or Trend+CA outperforms DePreSys significantly for certain regions, such as the NAC and South Atlantic.
- With a projected increase in equivalent CO_2 concentration, both statistical models forecast a small cooling trend in the subpolar gyre region for the years 2012–2021, while most other parts of the Atlantic are expected to warm, especially in the South Atlantic.

We have also attempted to understand the source of predictive skill in our statistical forecasts. We have identified certain regions where the prediction of the forced trend plays a dominant role in having skillful forecasts, such as parts of mid-latitude and tropical Atlantic, the Indian Ocean and the western Pacific. The prediction of internal variability is more important for some other regions, such as the far North Atlantic and perhaps parts of South Atlantic. These findings can help identify areas where decadal predictions may be improved by adding more observations.

We have seen that the relative skill of different prediction methods vary both spatially and with lead time. This is further illustrated in Fig. 11 which show summary maps of the best statistical prediction method for different lead time periods in terms of having the lowest MSE. Over the Atlantic, while Trend+AR1 is generally the best performing model for year 1, the more complex Trend+CA performs better for years 2–5 and 6–9. Even for these longer lead times, however, the forced trend model, which is the simplest benchmark in this study, performs best for certain regions. We can also see that the regions where
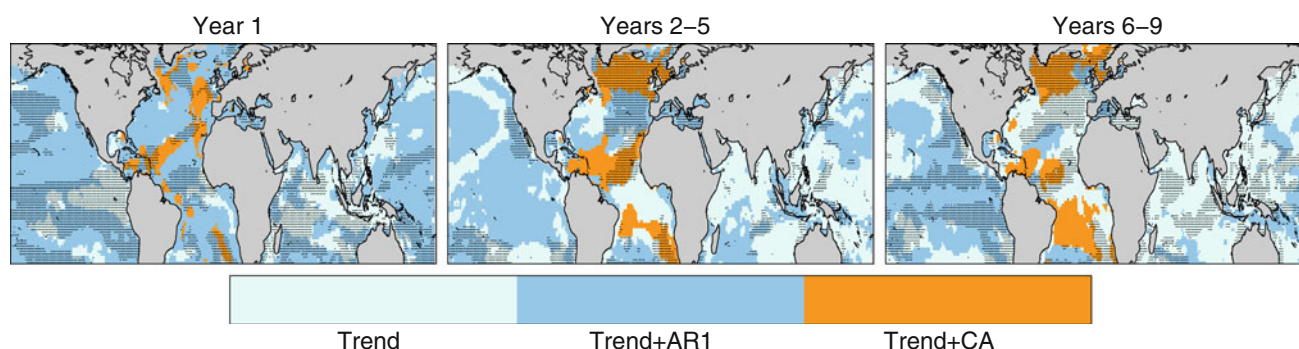

Fig. 11 The best prediction model with the lowest MSE as shown by *shades* of different colours. Note that Trend+CA forecasts are performed in the Atlantic sector only. The stippled areas are where DePreSys outperforms all the available statistical models

DePreSys outperforms all the statistical models (stippled areas in Fig. 11) vary with lead time. Combining predictions using different methods in order to improve accuracy has been attempted in seasonal forecasting (e.g. Colman and Davey 2003). The possibility of performing similar work on decadal predictions should be considered, for example combining different statistical forecasts to provide a better benchmark or combining statistical and dynamical predictions to give the 'best possible' forecast to the end user.

There is much scope for further work on developing benchmark statistical models for decadal predictions. While the retrospective forecasts by our two statistical models are skillful for certain regions on a decadal timescale, other modelling strategies may be considered. Specifically, the separation of the effects of radiative forcings and internal variability on SST variability is not trivial. The negative correlation for our forced trend prediction in the far North Atlantic suggests that our model is far from ideal. One aspect worth noting is the effect of tropospheric aerosols. The equivalent CO_2 concentration data used in this study represent the total net *global* effects of greenhouse gases and aerosols, however the spatial distribution of aerosols is not uniform which means that our model is unlikely to capture the *regional* climate effects of aerosols. Meanwhile, recent work by Booth et al. (2012) suggested that aerosols have played a key role in forcing North Atlantic SST changes. Some other known radiative forcing effects are also not included in this study, for example solar activity and volcanic eruptions. Further work is required to explore the best modelling option. In addition, for the modelling of the forced trend and AR1 processes, a separate model has been fitted to each grid box. A modelling approach in which all grid boxes are fitted together and parameters for each grid box assumed to be related would be worth exploring. Finally, this study has considered SSTs because of their influence on weather systems and climate patterns. It will be also interesting to explore if our statistical models can offer skillful regional predictions of other climate variables, for example land surface temperatures.

Appendix: Assessing the statistical significance of verification metrics

The statistical significance of correlation coefficients and MSSS is assessed using a non-parametric bootstrapping approach. Here we use the MSSS of the anomaly predictions (by Trend+AR1 or Trend+CA) relative to the forced trend predictions (Trend) as an example. The bootstrapping for the correlation coefficients is done in a similar manner. For lead time of k years, we have the anomaly predictions $\hat{X} = \left\{ \hat{X}_{t+k}, \hat{X}_{t+k+1}, \ldots, \hat{X}_{t+k+(N-1)} \right\}$ and the forced trend predictions $\hat{X}^f = \left\{ \hat{X}^f_{t+k}, \hat{X}^f_{t+k+1}, \ldots, \hat{X}^f_{t+k+(N-1)} \right\}$, where t is the start year and N is the total number of predictions available for verification. We also have the verifying observations $X = \left\{ X_{t+k}, X_{t+k+1}, \ldots, X_{t+k+(N-1)} \right\}$.

We produce $B = 999$ bootstrap samples of $\hat{X}$, $\hat{X}^f$ and X, each of length N. In the resampling process for each of the B bootstrap samples, we need to account for both the dependence among the three variables and the serial dependence within samples. We preserve the dependence among the three variables by resampling $\hat{X}_q$ and $\hat{X}^f_q$ whenever X_q is chosen, where $t + k \leq q \leq t + k + (N - 1)$. For the serial dependence, a block bootstrapping technique (Davison and Hinkley 1997, Chap. 8) is adopted. We first create overlapping blocks of data with length m within each original sample, then we resample the $N - m + 1$ blocks uniformly with replacement to produce each of the B bootstrap samples. Here we choose $m = 5$ based on studying the autocorrelation structure for SST time series at selected grid boxes.

After the resampling process, we calculate the MSSS for each bootstrap sample as described in Sect. 3.5. The $100(1 - \alpha)\%$ confidence interval (CI) of MSSS is constructed based on the empirical $\alpha/2$ and $(1 - \alpha)/2$ quantile of the MSSS distribution. If zero is not included in the CI, then MSSS is considered significant at the α level.

References

Adger W, Arnell N, Tompkins E (2005) Successful adaptation to climate change across scales. Glob Environ Change 15(2):77–86. doi:10.1016/j.gloenvcha.2004.12.005

Barnston A, van den Dool H, Zebiak S, Barnett T, Ji M, Rodenhuis D, Cane M, Leetmaa A, Graham N, Ropelewski C, Kousky V, O'Lenic E, Livezey R (1994) Long-lead seasonal forecasts—where do we stand? Bull Am Meteorol Soc 75(11):2097–2114

Boer G (2004) Long time-scale potential predictability in an ensemble of coupled climate models. Clim Dyn 23(1):29–44. doi:10.1007/s00382-004-0419-8

Boer GJ (2011) Decadal potential predictability of twenty-first century climate. Clim Dyn 36(5–6):1119–1133. doi:10.1007/s00382-010-0747-9

Boer GJ, Lambert SJ (2008) Multi-model decadal potential predictability of precipitation and temperature. Geophys Res Lett 35(5). doi:10.1029/2008GL033234

Booth B, Dunstone N, Halloran P, Andrews T, Bellouin N (2012) Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. Nature 484:228–232. doi:10.1038/nature10946

Challinor A (2009) Towards the development of adaptation options using climate and crop yield forecasting at seasonal to multi-decadal timescales. Environ Sci Policy 12(4):453–465. doi:10.1016/j.envsci.2008.09.008

Collins M, Booth BBB, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2006) Towards quantifying uncertainty in transient climate change. Clim Dyn 27(2–3):127–147. doi:10.1007/s00382-006-0121-0

Collins M, Booth BBB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2011) Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. Clim Dyn 36(9–10):1737–1766. doi:10.1007/s00382-010-0808-0

Colman A, Davey M (2003) Statistical prediction of global sea-surface temperature anomalies. Int J Climatol 23(14):1677–1697. doi:10.1002/joc.956

Davison AC, Hinkley DV (1997) Bootstrap methods and their applications. Cambridge University Press, Cambridge

DelSole T, Tippett MK, Shukla J (2011) A significant component of unforced multidecadal variability in the recent acceleration of global warming. J Clim 24(3):909–926. doi:10.1175/2010JCLI3659.1

Diffenbaugh NS, Scherer M (2011) Observational and model evidence of global emergence of permanent, unprecedented heat in the 20th and 21st centuries. Clim Change 107(3–4):615–624. doi:10.1007/s10584-011-0112-y

Draper NR, Smith H (1998) Applied regression analysis. 3rd edn. Wiley, New York, USA

Fildes R, Kourentzes N (2011) Validation and forecasting accuracy in models of climate change. Int J Forecast 27(4):968–995. doi:10.1016/j.ijforecast.2011.03.008

Fyfe JC, Merryfield WJ, Kharin V, Boer GJ, Lee WS, von Salzen K (2011) Skillful predictions of decadal trends in global mean surface temperature. Geophys Res Lett 38:L22801. doi:10.1029/2011GL049508

Giorgi F, Bi X (2009) Time of emergence (TOE) of GHG-forced precipitation change hot-spots. Geophys Res Lett 36. doi:10.1029/2009GL037593

Goddard L, Kumar A, Solomon A, Smith D, Boer G, Gonzalez P, Kharin V, Merryfield W, Deser C, Mason S, Kirtman B, Msadek R, Sutton R, Hawkins E, Fricker T, Hegerl G, Ferro C, Stephenson D, Meehl G, Stockdale T, Burgman R, Greene A, Kushnir Y, Newman M, Carton J, Fukumori I, Delworth T (2012a) A verification framework for interannual-to-decadal predictions experiments. Clim Dyn. doi:10.1007/s00382-012-1481-2 (in press)

Goddard L, Hurrell L, Kirtman B, Murphy J, Stockdale T, Vera C (2012b) Two timescales for the price of one. Bull Am Meteorol Soc 93(5):621–629. doi:10.1175/BAMS-D-11-00220.1

Goldenberg S, Landsea C, Mestas-Nunez A, Gray W (2001) The recent increase in Atlantic hurricane activity: causes and implications. Science 293(5529):474–479. doi:10.1126/science.1060040

Gordon C, Cooper C, Senior C, Banks H, Gregory J, Johns T, Mitchell J, Wood R (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16(2–3):147–168. doi:10.1007/s003820050010

Hawkins E, Sutton R (2012) Time of emergence of climate signals. Geophys Res Lett 39. doi:10.1029/2011GL050087

Hawkins E, Robson J, Sutton R, Smith D, Keenlyside N (2011) Evaluating the potential for statistical decadal predictions of sea surface temperatures with a perfect model approach. Clim Dyn 37(11–12):2495–2509. doi:10.1007/s00382-011-1023-3

Keenlyside NS, Ba J (2010) Prospects for decadal climate prediction. Wiley Interdisc Rev Clim Change 1(5):627–635. doi:10.1002/wcc.69

Keenlyside NS, Latif M, Jungclaus J, Kornblueh L, Roeckner E (2008) Advancing decadal-scale climate prediction in the North Atlantic sector. Nature 453(7191):84–88. doi:10.1038/nature06921

Krueger O, von Storch JS (2011) A simple empirical model for decadal climate prediction. J Clim 24(4):1276–1283. doi:10.1175/2010JCLI3726.1

Laepple T, Jewson S, Coughlin K (2008) Interannual temperature predictions using the CMIP3 multi-model ensemble mean. Geophys Res Lett 35(10). doi:10.1029/2008GL033576

Landsea C, Knaff J (2000) How much skill was there in forecasting the very strong 1997–98 El Nino? Bull Am Meteorol Soci 81(9):2107–2119

Latif M, Collins M, Pohlmann H, Keenlyside N (2006) A review of predictability studies of Atlantic sector climate on decadal time scales. J Clim 19(23):5971–5987. doi:10.1175/JCLI3945.1

Lean JL, Rind DH (2009) How will Earth's surface temperature change in future decades? Geophys Res Lett 36. 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(20):5305–5318. doi:10.1175/JCLI3912.1

Mahlstein I, Knutti R, Solomon S, Portmann RW (2011) Early onset of significant local warming in low latitude countries. Environ Res Lett 6(3). doi:10.1088/1748-9326/6/3/034009

Meinshausen M, Smith SJ, Calvin K, Daniel JS, Kainuma MLT, Lamarque JF, Matsumoto K, Montzka SA, Raper SCB, Riahi K, Thomson A, Velders GJM, van Vuuren DPP (2011) The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. Clim Change 109(1–2):213–241. doi:10.1007/s10584-011-0156-z

Murphy A, Epstein E (1989) Skill scores and correlation coefficients in model verification. Mon Wea Rev 117(3):572–581

Nakićenović N, Swart R (2000) Emission scenarios. A special report of working group III of the intergovernmental panel on climate change. Cambridge University Press, New York

Rayner N, Parker D, Horton E, Folland C, Alexander L, Rowell D, Kent E, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res Atmos 108(D14). doi:10.1029/2002JD002670

Smith D, Scaife A, Boer G, Caian M, Doblas-Reyes F, Guemas V, Hawkins E, Hazeleger W, Hermanson L, Ho C, Ishii M, Kharin V, Kimoto M, Kirtman B, Lean J, Matei D, Merryfield W, Müller W, Pohlmann H, Rosati A, Wouters B, Wyser K (2012) Real-time multi-model decadal climate predictions. Clim Dyn (Submitted)

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(5839):796–799. doi:10.1126/science.1139540

Smith DM, Eade R, Dunstone NJ, Fereday D, Murphy JM, Pohlmann H, Scaife AA (2010) Skilful multi-year predictions of atlantic hurricane frequency. Nature Geosci 3(12):846–849. doi:10.1038/NGEO1004

Solomon A, Goddard L, Kumar A, Carton J, Deser C, Fukumori I, Greene AM, Hegerl G, Kirtman B, Kushnir Y, Newman M, Smith D, Vimont D, Delworth T, Meehl GA, Stockdale T (2011) Distinguishing the roles of natural and anthropogenically forced decadal climate variability and implications for prediction.

Bull Am Meteorol Soc 92(2):141–156. doi:10.1175/2010BA MS2962.1

Sutton R, Hodson D (2005) Atlantic Ocean forcing of North American and European summer climate. Science 309(5731): 115–118. doi:10.1126/science.1109496

Taylor K, Stouffer R, Meehl G (2011) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93(4):485–498. doi: 10.1175/BAMS-D-11-00094.1

Ting M, Kushnir Y, Seager R, Li C (2009) Forced and internal twentieth-century sst trends in the North Atlantic. J Clim 22(6):1469–1481. doi:10.1175/2008JCLI2561.1

van den Dool H (1994) Searching for analogues, how long must we wait? Tellus Ser A Dyn Meteorol Oceanogr 46(3):314–324

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

van den Dool H, Huang J, Fan Y (2003) Performance and analysis of the constructed analogue method applied to US soil moisture over 1981–2001. J Geophys Res Atmosp 108(D16). doi:10.1029/2002JD003114

van Oldenborgh G, Doblas-Reyes F, Wouters B, Hazeleger W (2012) Decadal prediction skill in a multi-model ensemble. Clim Dyn 38(7–8):1263–1280. doi:10.1007/s00382-012-1313-4

Zanna L (2012) Forecast skill and predictability of observed Atlantic sea surface temperatures. J Clim 25(14):5047–5056. doi:10.1175/JCLI-D-11-00539.1

Zhang R, Delworth TL (2006) Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. Geophys Res Lett 33(17). doi:10.1029/2006GL026267

Zhu X, Fraedrich K, Liu Z, Blender R (2010) A demonstration of long-term memory and climate predictability. J Clim 23(18):5021–5029. doi:10.1175/2010JCLI3370.1