ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Skilful decadal predictions of subpolar North Atlantic SSTs using CMIP model-analogues

Matthew B Menary[1,2], **Juliette Mignot**[2] **and Jon Robson**[3]

[1] LMD-IPSL, École Normale Supérieure, Paris 75005, France
[2] LOCEAN Laboratory, Sorbonne Université-CNRS-IRD-MNHN, Paris 75005, France
[3] NCAS, Department of Meteorology, University of Reading, Reading, United Kingdom

E-mail: matthew.menary@locean.ipsl.fr

Keywords: AMOC, decadal prediction, analogue, CMIP5, CMIP6, North Atlantic, sea surface temperatures

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
15 December 2020

REVISED
21 May 2021

ACCEPTED FOR PUBLICATION
28 May 2021

PUBLISHED
16 June 2021

Abstract

Predicting regional climate variability is a key goal of initialised decadal predictions and the North Atlantic has been a major focus due to its high level of predictability and potential impact on European climate. These predictions often focus on decadal variability in sea surface temperatures (SSTs) in the North Atlantic subpolar gyre (NA SPG). In order to understand the value of initialisation, and justify the high costs of such systems, predictions are routinely measured against technologically simpler benchmarks. Here, we present a new model-analogue benchmark that aims to leverage the latent information in uninitialised climate model simulations to make decadal predictions of NA SPG SSTs. This system searches through more than one hundred thousand simulated years in Coupled Model Intercomparison Project archives and yields skilful predictions in its target region comparable to initialised systems. Analysis of the underlying behaviour of the system suggests the origins of this skill are physically plausible. Such a system can provide a useful benchmark for initialised systems within the NA SPG and also suggests that the limits in initialised decadal prediction skill in this region have not yet been reached.

1. Introduction

As the global climate continues to change in response to anthropogenic influences (Bindoff *et al* 2013), estimates of how climate might evolve in the near term and on a regional level are becoming increasingly important (Kushnir *et al* 2019). These estimates can take the form of projections of the next century using different shared socioeconomic pathways from the 6th Coupled Model Intercomparison Project (CMIP6) (Eyring *et al* 2016). They can also take the form of initialised predictions of the next decade, conducted as part of the Decadal Climate Prediction Project (DCPP) (Boer *et al* 2016). Both of these methods utilise complex coupled climate models in which external forcings such as greenhouse gases (GHGs) and aerosols are prescribed.

Previous work has shown that sea surface temperatures (SSTs) in the North Atlantic (NA) including the subpolar gyre (SPG) can impact climate both locally and remotely (Sutton and Hodson 2005, Monerie *et al* 2018). In addition, they are potentially predictable on decadal timescales (Collins *et al* 2006) and successfully initialising them can provide skill elsewhere (Dunstone *et al* 2011). Recent analysis of the CMIP6 archive has shown an improvement since CMIP5 in multiannual skill in SSTs in the NA SPG, in both uninitialised and initialised simulations (Borchert *et al* 2021). In addition, technologically simpler benchmarks are a useful tool to help quantify the uncertainty in climate model projections/predictions (Brunner *et al* 2020).

Various types of analogue methods have been proposed as skill benchmarks, with skill greater than comparable dynamical forecast systems at particular lead times or in particular regions (Hawkins *et al* 2011, Ho *et al* 2013). These methods are based on the underlying assumption that a pair of climate states that are similar (analogous) will remain so for a certain amount of time, with the timescale dependent on both their initial similarity and the rate of error growth in the chosen variable/region (Lorenz 1969).

The simplest analogue consists of finding the single most similar match in the observational

record for the (non-contemporaneous) observed data, where similarity is arbitrarily defined based on the variable and statistic deemed most relevant for the particular problem at hand. However, due to lack of long-term observations this natural analogue method is likely less skilful than a composite or constructed analogue approach (Dool 1994). In a constructed analogue, several observed fields are combined with potentially varying weights and the forecast is the similarly weighted combination of the fields' evolution. Similar methods are also used to probe physical drivers of variability by, for example, constructing atmospheric circulation analogues to determine the percentage of surface temperature variability driven by atmospheric dynamics (Deser *et al* 2016, O'Reilly *et al* 2017). Constructed analogues have proven skilful for both seasonal forecasts and on multiannual/decadal timescales (Dool 1994, Hawkins *et al* 2011, Ho *et al* 2013, Yiou and Déandréis 2019). However, on annual to decadal timescales, the skill in the NA SPG has often remained substantially lower than the surrounding regions using such constructed analogues (Newman 2013, Suckling *et al* 2017), despite the high skill in initialised predictions.

In terms of input data, as an alternative to probing the short observational record, model data can be used. Assuming that the model data is an adequate representation of reality, the increased amount of data increases the likelihood of finding similar/analogous climate states (Ding *et al* 2018, 2019, Brown and Caldeira 2020). Within the CMIP archives there exists a vast amount of catalogued climate model data; 3.3 PB in CMIP5 and projected to reach 18 PB in CMIP6 (Balaji *et al* 2018).

Here, we aim to design an updated model-analogue system to provide updated skill benchmarks of the SPG. We aim to do this by leveraging the vast CMIP5/6 archives. We focus on decadal timescales (i.e. lead years 2–10 combined) as the ability of simple models to largely capture the dynamics of the NA SPG on these timescales (Born *et al* 2015), should reduce the degrees of freedom and make the search for good analogues more attainable (Dool 1994). Nonetheless, our goal is not to downplay the value of initialised decadal prediction systems, for which the globally distributed skill in physically linked variables makes them an impressive scientific and technical achievement. Instead, we specifically aim to leverage the latent physical information within uninitialised climate model simulations that is not currently being used for predictions, but which could be. Such an approach could also avoid issues with initialisation shock as well as model mean state biases, which remain an issue for initialised prediction systems whether these biases are removed before making the predictions (anomaly assimilation) or in post-processing (Smith *et al* 2013).

When assessing our model-analogue based system, we compare against initialised and uninitialised simulations taken from the CMIP archive. In order to simplify interpretation of our results, we focus on the subset of models providing both uninitialised simulations and initialised predictions. We also discuss single-forcing experiments that include, for example, just the historical forcing due to anthropogenic aerosols, conducted as part of the Detection and Attribution Model Intercomparison Project (Gillett *et al* 2016). In the next section, we describe the design and construction of our model-analogue based system.

2. Methods

2.1. Designing the model-analogue system

The overall ambition of our model-analogue based forecast system is to analyse observed climate fields, determine the most similar simulated fields, and use these simulations to create forecasts of real-world SSTs, focussing on the NA SPG. This is different to some analogue or linear inverse modelling approaches, where one dimensional time series' of predictors are used, based on external forcings or the principal component time series' of fixed patterns of variability (Oldenborgh *et al* 2005, Zanna 2012, Newman 2013, Eden *et al* 2015, Suckling *et al* 2017). We use spatial inputs for two reasons: (a) to avoid predetermining the important patterns/modes of variability (within a given domain, e.g. the North Atlantic), in case these cannot be adequately ascertained given the limited observational record, and (b) the combined size of the CMIP5 and CMIP6 archives means we have very many simulated fields to choose from and so do not necessarily require significant reduction of the dimensionality of the problem. We furthermore choose to target lead times of 2–10 years to focus on decadal predictability and unless otherwise stated all forecasts refer to these combined lead times.

Our initial hypothesis is that time mean SSTs over some time window within some subregion of the North Atlantic are linearly related to SPG SSTs at a forecast lead time of 2–10 years (inclusive) combined. This hypothesis is based on the space-scales and mechanisms of observed and simulated long period variability in the North Atlantic (Knudsen *et al* 2011, Ba *et al* 2014); for example, the slow northward advection of SST anomalies (Vellinga and Wu 2004). This hypothesis is further motivated by the fact that, in order to compare observed and simulated fields, we require a well observed (in time and space) field, such as SSTs. In principle, our system could instead/also ingest an atmospheric variable such as surface air temperature, but we choose SSTs as oceanic memory is likely to provide greater skill (given the forecast variable is also SST). We could also ingest subsurface oceanic properties, which may provide improved physical linkages, but a lack of long-term data is again a drawback.

Such a model-analogue prediction system is based on the assumption that the mechanisms of variability that link past and future variations in SST within

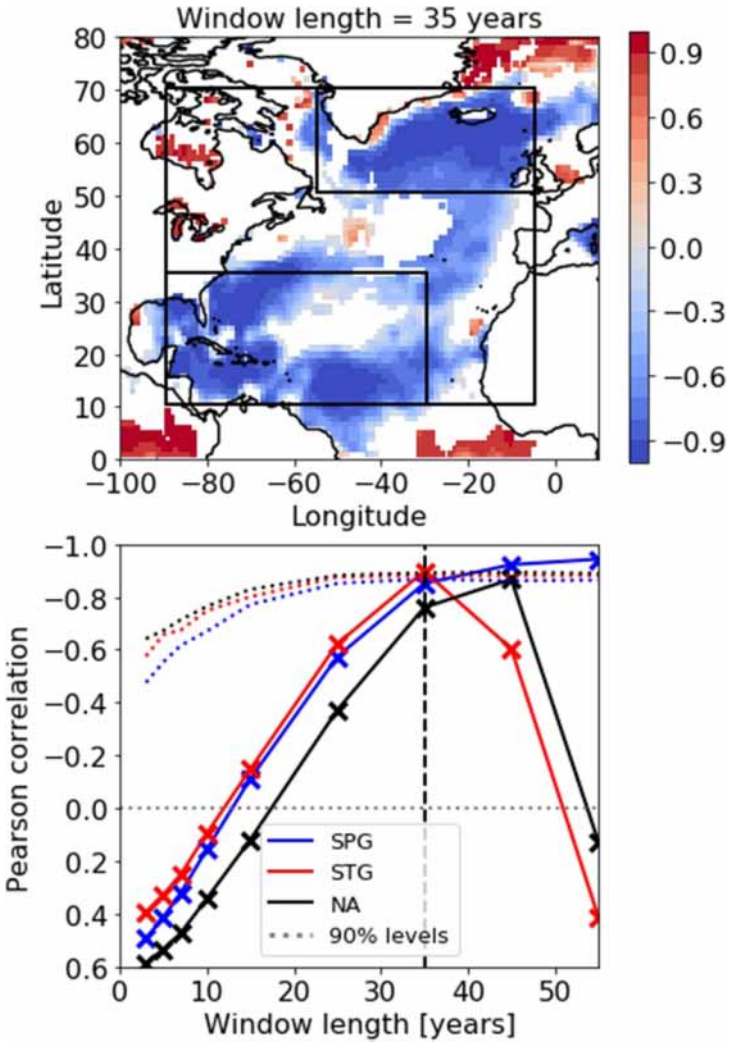

Figure 1. Observed relationships. Top: Pearson correlation between running 35 year means of observed, gridded SSTs and area-averaged SSTs in the SPG (top right box) at following lead years 2–10 inclusive (from HadISST1). Correlations are for the period 1960–2019 but are masked where the sign of the correlation over this period and the period 1870–1990 are different. Bottom: correlations between area-averaged SST in the SPG at lead times 2–10 and area-averaged SST in the SPG; subtropical gyre (STG, bottom left box); and North Atlantic (NA, large box) as a function of running mean window length. Significance estimated by bootstrapping random data with the same mean, variance, and auto-correlation as the SST time series and using the same running mean window lengths.

the North Atlantic are not systematically different between models and those seen in observations and that the patterns of these relationships are stationary in time. For example, the overall pattern of anomalous NA SSTs that leads to a particular SPG SST (forecast) anomaly is the same at the beginning of the 20th century as during the present day. However, as this is unlikely to be true for all models at all times, we aim to combine many models and epochs. In addition, our system makes no attempt to correctly time particular forcings and is *a priori* indifferent to interannual changes in external forcings. That is, any real world SPG SST variability that is directly driven by external forcings (e.g. a volcano), will not contribute to the predictions made by this system—unless, or until, this variability also has an imprint on NA SSTs. This aspect is unlike historical simulations or initialised (hindcast) predictions where the magnitude and timing of forcing due to, for example, a volcanic eruption can be prescribed.

The two key unknowns in this setup are the length of the time window to be used when searching for and choosing analogue fields, and the particular region (within the NA) over which to base our analogue choice. To address both these points, we regress observed, time-mean gridpoint SSTs in the North Atlantic against future area-averaged SPG SSTs (lead times 2–10 combined) using a variety of time-mean window lengths. We use optimally-interpolated observations from HadISST1 (Rayner *et al* 2003). Figure 1(a) shows an example of the correlation map when the time-mean moving window is 35 years. Note that, in order to satisfy our assumption of stationarity, we ignore all gridpoints where the sign of the regression slope differs between the period 1960–2019 (most well observed period) and

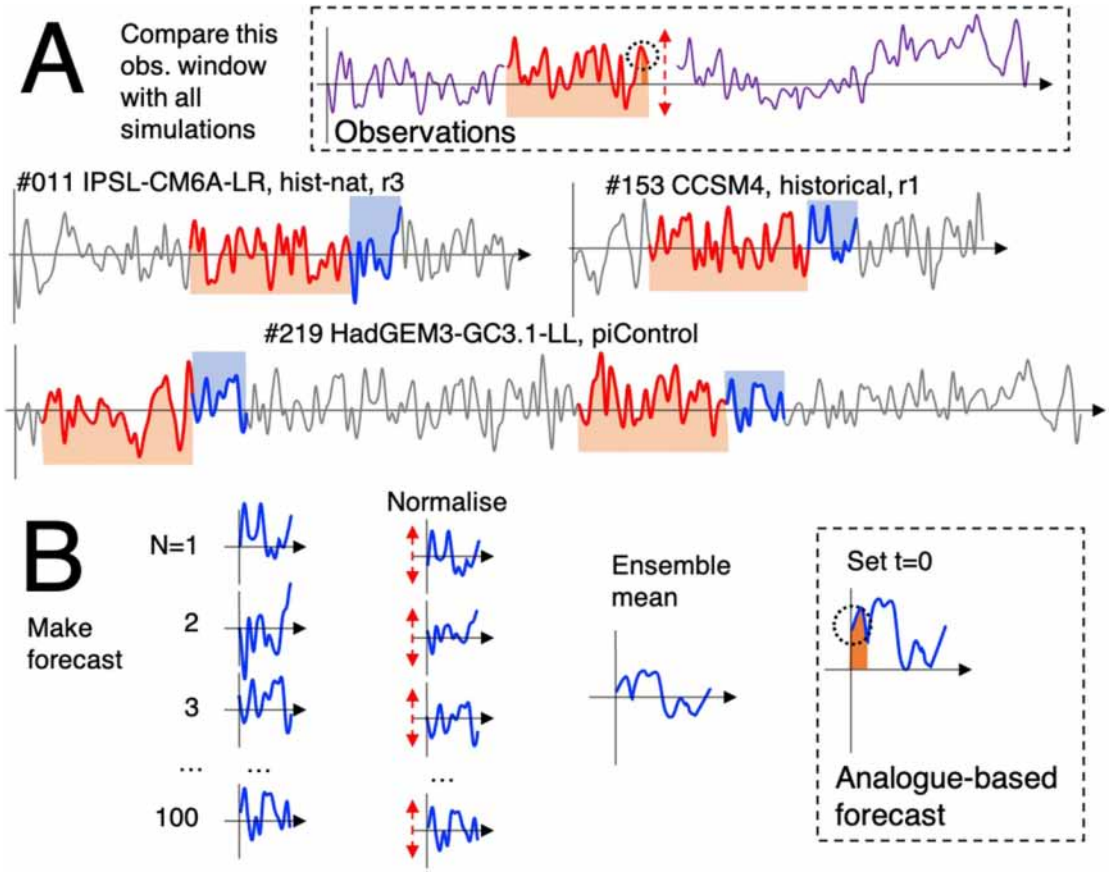

Figure 2. A schematic diagram of the steps taken to choose simulation segments and subsequently create the model-analogue forecast, as described in section 2. An arbitrary subset (numbers #11, #153, and #219) of the 668 possible simulations (totalling 100 630 simulated years) is shown.

1870–1990 (period not dominated by mid 90s SPG warming).

Figure 1(a) shows that there are strong correlations throughout the North Atlantic, particularly in the gyre regions. In terms of area averages, both the subtropical and subpolar regions show strong negative correlations when the time-mean window is 35 years in length and there is also a peak at this timescale for the whole North Atlantic (figure 1(b)). This peak suggests that 35 year mean SSTs are an optimal, simple prediction criterion for SPG SSTs on decadal timescales. Such a long timescale is perhaps not surprising given the long period variability observed within the SPG and wider North Atlantic. The 35 year window, negative correlation, and strong links to the subtropics, are also consistent with a role for Atlantic multidecadal variability (AMV, full period of around 70 years (Kerr 2000)) in providing predictability in the SPG (regardless of how AMV is driven). As such, we choose the whole North Atlantic as our input region.

2.2. Building the model-analogue system

Having chosen our input parameters based on analysis of observations, we now describe the method by which we compare observations and simulations (figure 2, part A) and subsequently create our forecasts (figure 2, part B), as follows:

Part A. Comparing observed and simulated SSTs

(a) All observations and simulations are regridded on to a regular $1° \times 1°$ grid. Annual data are used throughout.

(b) Time-mean maps of observed North Atlantic SST anomalies are produced for every possible 35 year window between 1870 and 2019 inclusive (116 total) using observations from HadISST (Rayner *et al* 2003). The 1960–1990 climatology is removed.

(c) We also create all possible 35 year time mean NA SST anomalies in all available models and experiments on the CMIP5 and CMIP6 archives (using up to the first ten ensemble members), denoted 'simulation segments'. For each simulation segment, a climatology is removed. This is the 1960–1990 time-mean, ensemble mean from that model's historical simulations regardless of the experiment. We use the following experiments from CMIP5: piControl, historical, rcp45, rcp85 (Taylor *et al* 2012). We use the following experiments from CMIP6: piControl, historical, hist-nat, hist-aer, hist-GHG, hist-stratO3, ssp126, ssp585 (Eyring *et al* 2016, Gillett

et al 2016, O'Neill et al 2016). After quality control (for example, removing models with incomplete metadata, such as basin masks), the total number of simulation segments is 100 630.

(d) For each of the 116 observed 35 year mean North Atlantic SST anomalies (created in #2) we compare to all 100 630 simulation segments (from #3) by computing the root mean square error (RMSE) between these two fields over our input domain (the North Atlantic as shown by the large box in figure 1(a)). Prior removal of the respective climatologies ensures this step does not merely return the same models that happen to have the smallest mean state biases, which can be as large as the signals we are trying to predict (Smith et al 2013, Menary et al 2015).

(e) From each of the best (i.e. lowest RMSE) 100 simulation segments we store the subsequent ten years of gridded SST data in order to create our forecast. Note we do not allow overlapping segments from the same simulation. For example, if a given simulation segment is chosen (e.g. HadGEM3-GC3.1-LL, piControl, 35 year window from year 305–339 inclusive) then we do not allow other segments from the same simulation that encompass any of the same years, although non-overlapping segments are allowed.

Part B. Creating the forecast

(a) Each of the 100, gridded, ten year forecast members are first normalised to the same standard deviation as the HadISST observations. This is to ensure that the final forecasts are not dominated by some models with very large variability. Normalisation is done for each gridpoint and is based on the standard deviations over the prior 35 year segment. That is, no information is allowed to leak from the successive observations. The means are also adjusted to match the observed mean over the previous five years. In the absence of any additional skill provided by our method, this step will result in forecasts with similar skill to persistence of the previous five year mean. The relative ensemble member densities before these steps are shown in figure 3(a).

(b) The ensemble mean of these 100 normalised forecasts is then taken.

(c) Finally, lead times 2–10 inclusive are averaged together.

This process results in SST forecast maps for each start year from 1904 to 2019. These can be compared with observations and other prediction methods to determine the relative skill of this model-analogue based system. Specifically, we compare to (a) ensemble mean hindcasts made with initialised decadal prediction systems (from 1960 onwards), (b) the nine year running mean of the ensemble mean of uninitialised (i.e. historical) simulations from the same modelling centres that provided hindcast simulations (from 1870 onwards), and (c) hindcasts from the MPI decadal prediction system that begin in 1900.

3. Results

3.1. The skill of the model-analogue system

Figure 3(a) shows the nine year running mean observed SPG SST along with the equivalent forecast using this model-analogue system. The system shows high skill (measured by the anomaly correlation coefficient, ACC) for the whole period ($r = 0.75$), which increases when just considering forecasts for the period since 1960 ($r = 0.82$). Both of these skill scores lie outside bootstrapped persistence (of previous five year mean) forecasts using pseudo-time series with the same mean, variance, and auto-correlation as observed (whole period: $r = 0.61$; since 1960: $r = 0.74$, 90% level).

In addition to the SPG, the skill of the model-analogue system remains high globally (figure 4(a)). This is despite the fact that the system is (a) only choosing simulation segments based on their spatial similarity to observations within the North Atlantic (highlighted region), (b) does not include any information about the timing of external forcings, and (c) that this analogue design was chosen for its potential skill in the NA SPG only. For this full period, we can also compare against initialised forecasts made using the unusually long initialised prediction dataset of MPI-ESM-LR (figure 4(b)) (Müller et al 2014) and un-initialised historical simulations taken from the combined CMIP5 and CMIP6 archives (figure 4(c)). The initialised predictions with MPI-ESM-LR and the CMIP5 + 6 combined uninitialised simulations also show globally high skill, with the exception of the Southern Ocean region. Nonetheless, within the NA SPG, there is a clear increase in skill in the model-analogue based system as compared to either of these baselines (figures 4(h) and (i)). For completeness, maps and comparisons of selected individual forecasts are provided in supplementary figure S1 (available online at stacks.iop.org/ERL/16/064090/mmedia).

To further test the robustness of the skill of the analogue system, we subsample our forecasts to just the period since 1960, which allows us to also compare against the combined CMIP5 and CMIP6 initialised prediction systems (figure 4(f)). This 'well-observed' period also gives us more confidence in the verifying observations that determine the skill of our system as well as the observed North Atlantic time mean SST patterns from which our system is built. For this period, much of the skill of the model-analogue based system exists in the North Atlantic (figure 4(d), grey hatching denotes skill is less than persistence), whilst the other systems continue to show high skill globally outside of the Southern

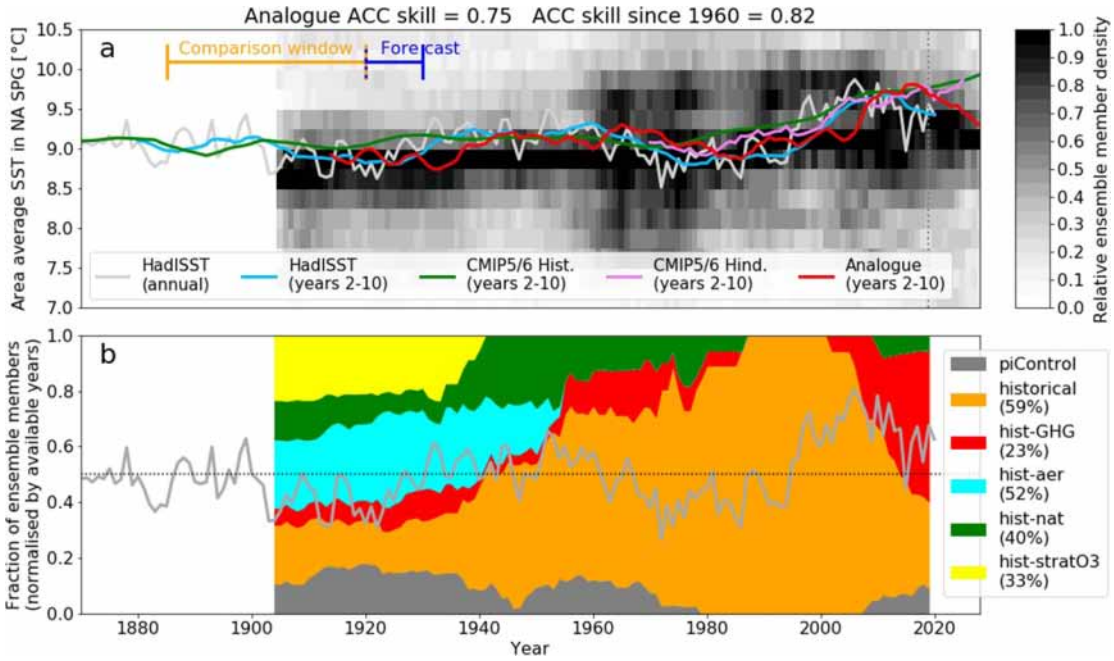

Figure 3. (a) Distribution of SPG area-averaged SSTs in all forecast ensemble members at all lead times in raw forecasts, i.e. after Part A of the model-analogue system construction, but before being combined in Part B (see section 2). Also shown are time series of annual mean (grey) and nine year running mean SPG SSTs from HadISST1 (blue; our forecast target). In addition, for the same validity (i.e. comparable) times we show the model-analogue based lead years 2–10 (inclusive) forecasts using the processed data (red), the nine year running mean from the uninitialised simulations (historical plus scenario; green) and the lead years 2–10 forecasts from the initialised predictions since 1960 (only those systems including start dates from 1960 to at least 2010, purple). The length of the 35 year analogue comparison window and subsequent forecast length are shown for reference. (b) Number of ensemble members for each start year that draw data from the labelled experiments, shown as a fraction of the total available simulation segments using that experiment. Numbers in brackets denote the percentage of those simulation segments that lie within 20 years of the associated observed segment (i.e. they are approximately contemporaneous). Grey line is observed SPG SSTs for reference (identical to panel a).

Ocean (figures 4(e)–(g)). Given that our system is designed based on observed variability in the North Atlantic and specifically targets skill in the NA SPG (cf figure 1), this is perhaps not surprising. Nonetheless, that our target region remains the region that compares most favourably against other methods, independently of the period chosen, suggests that our underlying hypotheses/design choices are valid. Furthermore, the model-analogue based system remains able to improve on the skill of initialised or uninitialised simulations within parts of the NA SPG (figures 4(h)–(l)). Nonetheless, the high skill of the uninitialised systems since 1960 (figure 4(l)) is consistent with a very strong role for external forcing in this region (Borchert *et al* 2021).

3.2. The behaviour of the model-analogue system

Following the high skill of our predictions within the North Atlantic SPG, it appears apposite to investigate which experiments our system is choosing, given the ten possible CMIP experiments totalling 100 630 simulation segments. For example, if some fraction of an observed SST pattern is externally forced at a given time, then we might expect to find simulation segments that incorporate that forcing to be preferentially chosen.

The most commonly chosen experiment, weighted by its relative availability, is the historical (figure 3(b)). Initially, the other commonly chosen experiments are the piControl and the hist-aer, the latter covering the same time-frame as the historical experiment but only including emissions of anthropogenic aerosols. For the first 50 years, hist-aer is approximately 5 times more likely to be chosen than the hist-GHG experiment, which includes instead only emissions of GHGs. The use of simulation segments from the hist-aer experiment up to around 1950 corresponds with a gradual warming of the SPG from 1900, which is sometimes associated with models that also encompass strong aerosol forcing (Booth *et al* 2012). Given also that 52% of the hist-aer simulation segment start times are within 20 years of the actual year (numbers in brackets in figure 3(b)), this gives us confidence that our method is matching simulated and observed SST patterns that are due to similar tendencies (i.e. an increase) in aerosol forcing. Nonetheless, this information alone does not mean that aerosols were necessarily the dominant forcing of North Atlantic variability up to 1950 in reality, merely that simulations with their inclusion provide time mean SSTs more similar to observed SST estimates than achieved by other experiments.

It is notable that the hist-nat experiment is chosen up to around 1985, but is not chosen during the period of particularly strong volcanic forcing associated with the eruption of Mount Pinatubo in 1991, for

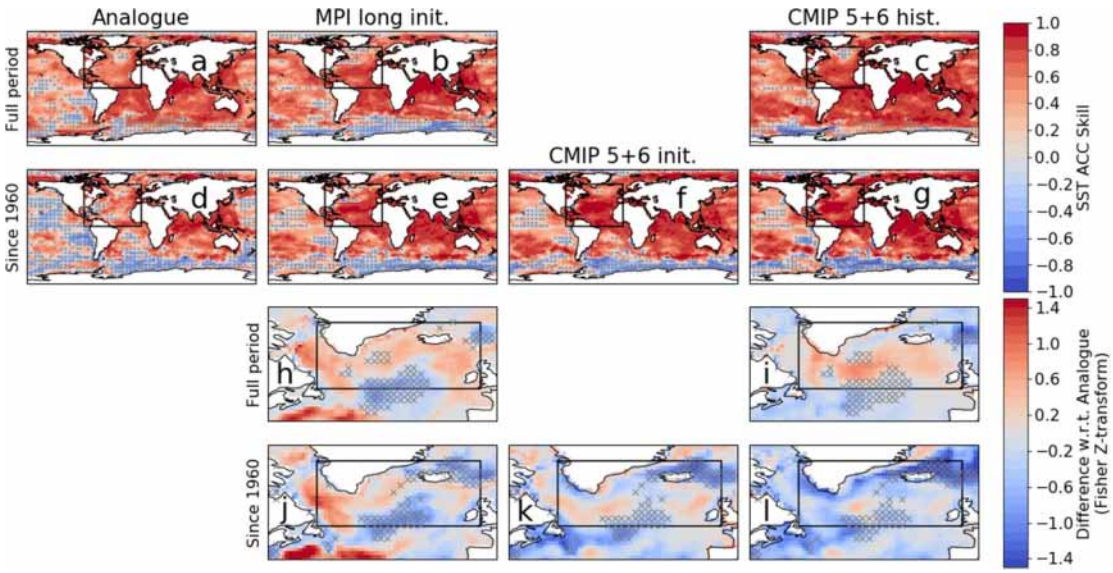

Figure 4. The anomaly correlation coefficient (ACC) skill of lead time 2–10 (inclusive) forecasts, using the model-analogue system (first column; (a), (d)), initialised predictions with MPI-ESM-LR (second column; (b), (e)), initialised predictions with a combination of available CMIP5 and CMIP6 prediction systems (third column; (f)), and uninitialised forecasts from similar CMIP5 and CMIP6 historical simulations (fourth column; (c), (g)). Top row shows skill over the full available period and second row the skill over the common and well observed period since 1960. Hatching shows regions where skill is not significantly greater than the 90% level of a persistence forecast of five year means to lead years 2–10 (estimated using pseudo-data with the same mean, variance, and auto-correlation as observed). Final two rows show North Atlantic focussed difference between model-analogue skill and the various baselines after first transforming ACC using the Fisher Z-transform. Here, hatching shows regions where skill in the model-analogue system is not significantly greater than persistence at the 90% level. In panels a-g, the black box denotes the domain over which the model-analogues were chosen, as shown also in figure 1 and described in section 2. In panels (h)–(l), the black box denotes the subpolar region used for the skill of area-averages in figures 3(a) and 5. CMIP5 + 6 initialised models used: MPI-ESM-LR, MIROC5, NorCPM1, CanESM5, GFDL-CM2p1, MIROC6, CanCM4, IPSL-CM6A-LR, bcc-csm1-1, GEOS-5. CMIP5 + 6 uninitialised models used: MPI-ESM-LR, MIROC5, NorCPM1, CanESM5, GFDL-CM2p1, MIROC6, CanESM2, IPSL-CM6A-LR, bcc-csm1-1.

example. Either the inclusion of volcanic forcing has too small an impact when part of a 35 year time mean, or the addition of the associated anthropogenic forcings (i.e. the historical experiment) provides a better match to observations (in terms of North Atlantic SSTs).

From the middle of the 20th century onwards, and in particular after the year 2000, the hist-GHG experiment becomes relatively more important (figure 3(b)). In the target observations, this period is characterised by an anomalous warming of the STG region and a relative lack of warming in the SPG, the pattern of which the hist-GHG experiment is the most able to reproduce. Although scenario experiments are included (e.g. RCP4.5, RCP8.5, SSP126, SSP585), none are chosen. This may be at least partly due to our deliberate choice to treat all experiments as separate entities, rather than attempt to concatenate data from, for example, the historical and subsequent scenario experiments. As such, the first 35 year mean for a CMIP5 (CMIP6) scenario experiment beginning in 2005 (2015) ends at the year 2040 (2050), for which the mean state difference to present-day observations is presumably larger than can be found in other, more contemporaneous, experiments. Future work will aim to widen the effective pool of source model data.

In summary, although a statistical prediction, the consistency of the potentially important forcings in reality with the timing of the choice of particular experiments gives us some additional confidence in our model-analogue based approach. However, our experimental design that does not separate the mean state from the variability (e.g. does not remove some of the warming from future scenario simulations to render them more plausible for earlier periods) may also be a factor.

We probe our approach further by investigating the sensitivity of our results (i.e. the skill in forecasts of SPG SSTs) to our choice of time-mean window (figure 5). If our system is in fact most skilful at window lengths that are different to our observationally derived window length (35 years) then this would imply either systematically different behaviour between models and observations, or that the skill of our system is not likely to be robust. However, as a function of window length, the skill of our system (figure 5) largely mirrors the correlation between observed North Atlantic SSTs and subsequent forecasts of area-averaged SPG SSTs seen in figure 1(b). That is, our system is more skilful when we choose a window length that is most consistent with the strongest observed relationships, giving us further confidence in our method.

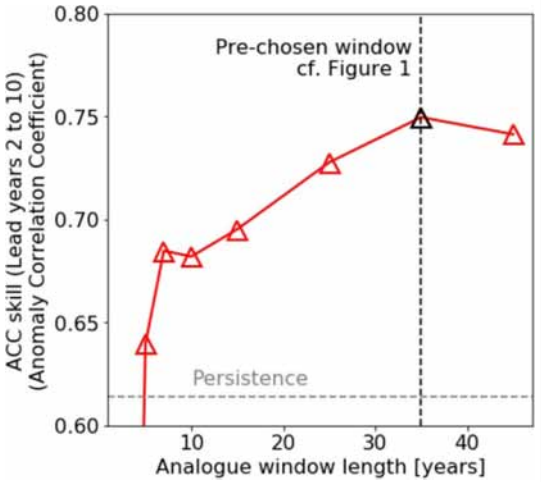

Figure 5. The sensitivity of the skill of our model-analogue based forecasts to the choice of running mean window length using the full period.

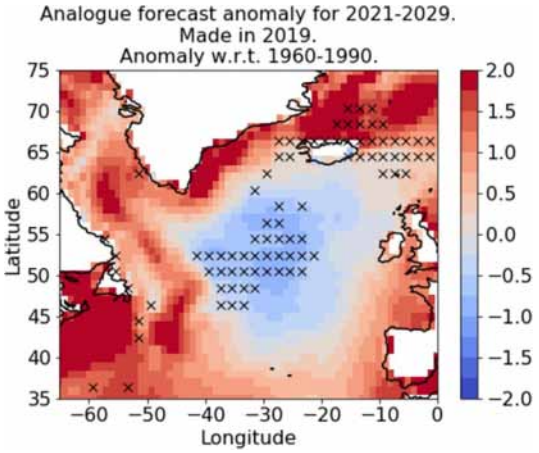

Figure 6. Actual anomaly forecast for the North Atlantic subpolar gyre region for the period 2021–2030 inclusive made using our model-analogue based system. The anomaly is with respect to an observed climatology over the period 1960–1990. Regions where the skill of the model-analogue system since 1960 is not significantly improved against persistence at the 90% level (see figure 4), and thus the forecast may be less reliable, are hatched.

3.3. Real SPG SST forecast with the model-analogue system

Finally, having demonstrated the skill of our model-analogue based prediction system and investigated the physical basis of the predictions, we use it to make a tentative real-world prediction of anomalous North Atlantic SPG SSTs at lead times 2–10 inclusive (figure 6). The system predicts an anomalously cool decade in the central SPG compared to climatology (i.e. 1960–1990), although the SPG as a whole is not forecast to be cooler than climatology (cf figure 3(a)). Note that hatched regions denote where the ACC skill of our system since 1960 is not significantly greater than persistence and thus less reliable. These cover a large part of the cool anomaly and so the overall cooling in the central SPG may not be reliable. This prediction is different to both the ensemble mean uninitialised models (cf figure 3(a), green lines) and initialised models (figure 3(a), purple lines), although the lead-time horizon for these is limited and some single model studies do forecast a cooling (Hermanson *et al* 2014). However, the analogue-model based prediction of cooler central-SPG SSTs would be consistent with the projected North Atlantic warming hole associated with a weakening of the Atlantic Meridional Overturning Circulation (AMOC) seen in most climate models (Menary and Wood 2018). It is also broadly consistent with the latest observations for 2020, which continue the downward trend in SPG SSTs (figure 3(a)). On the other hand, recent observational analyses forecast a possible short-term strengthening of the AMOC in the coming years (Desbruyères *et al* 2019, Moat *et al* 2020), which would likely eventually contribute to a warmer SPG.

4. Summary and discussion

We have presented a model-analogue based method to predict NA SPG SSTs on decadal timescales. We have shown that our system has a high level of skill as well as demonstrating that the behaviour of our system is consistent with our initial hypotheses. This system can serve as a useful benchmark in the NA SPG for future initialised predictions using models from CMIP6 and beyond. Combined with other empirical methods, such benchmarks can also help quantify the uncertainty in climate predictions (Suckling and Smith 2013, Brunner *et al* 2020).

Although the system we have designed does not require that dedicated coupled climate model simulations are performed, and can thus be considered a cheaper alternative, it does rely on these simulations already existing. In addition, the computational overheads involved in pre-processing 668 individual model simulations comprising a total of 100 630 simulated years are not trivial. We estimate that we used a total of 400 000 CPU hours on Jasmin, the UK's high performance storage and compute cluster (Lawrence *et al* 2013), which represents approximately 0.3% of the cost of, for example, the UK Met Office DCPP-A hindcast experiments (Leon Hermanson, *pers. comm.*). Compared to initialised prediction systems, our system remains relatively simple to modify and explore (cf figure 5).

We have demonstrated skilful predictions for both the period since 1904 (our first start year) and since 1960. The skill in predictions since 1960 is higher than in the full period, which may be related to the

broadly upwards trend in SPG SSTs since this time (figure 3(a)). Nonetheless, we also have more confidence in the skill of our system during this recent period. Further back in time, the gridded SST observations are increasingly sparse, which results in an increasing number of grid cells relaxing towards climatological values (Rayner *et al* 2003). Such 'relaxing to climatology' violates our initial hypotheses, namely that the pattern of SSTs in the North Atlantic is physically linked to the future evolution of SSTs in the NA SPG and that these relationships are stationary. To assess the stationarity of the SST pattern in the source models, we also computed this relationship in the piControl simulations. For some models, the SST pattern was similar to observed (figure 1(a)), but for others it was not, nor stationary within a given model over a multi-century simulation (not shown). As such, a future update to the system could be to pre-select the allowable models based on a metric of the similarity of this pattern. Nonetheless, in the current proof-of-concept study, we choose not to *a posteriori* change the experimental design. In addition, we tested using a correlation-based measure of similarity, rather than RMSE, in steps #4 and #5 of our model-building methodology (see section 2). This severely reduced the ultimate skill of our system. We hypothesise that this was due to the inclusion of too many simulation segments with large mean state offsets (despite similar anomalous patterns). If the mean state and variability are not independent, as has been demonstrated in the North Atlantic (Menary *et al* 2015), this would likely lead to poor forecasts and thus lower skill.

The CMIP initialised and uninitialised dynamical models we have compared against are generally more skilful than our model-analogue based method outside of the NA SPG, consistent with the construction of the analogue system. The uninitialised historical simulations contain the global pattern and precise temporal evolution of the external forcings, while the initialised systems also include global information in their initialisation, all of which provide skill. Nonetheless, despite these advantages, the model-analogue based system is able to demonstrate comparable skill to these systems within the NA SPG (figures 4(h)–(l)). This result demonstrates in principle how a targeted approach to the creation of a model-analogue system, involving the determination of real-world linkages between precursor and forecast variables, can provide high levels of prediction skill. In future, a hybrid approach may be possible, to synthesise the skill arising from different methods. Here, our goal was to design a system to provide skilful forecasts of NA SPG SST on decadal timescales in order to provide an updated benchmark for CMIP6 initialised prediction systems, which we have achieved, and to provide a jumping-off point for future analogue-based methods. This highlights the potential, and untapped, power of large multi-model archives of uninitialised simulations for making real-world predictions. Finally, it also suggests that that there remains yet more skill to be realised through initialised decadal predictions with dynamical models.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://esgf-index1.ceda.ac.uk/projects/esgf-ceda/.

Acknowledgments

MBM was supported by the EPICE project funded by the European Union's Horizon 2020 programme (Grant Agreement Number 789445) as well as the 4C project (Grant Agreement Number 821003) and the HARMONY project funded by France's Agence Nationale de la Recherche (ANR; Grant Agreement Number ANR-20-ERC9-0001). JM was supported by the EU-H2020 Blue Action (Grant Agreement Number 727852) and EUCP (Grant Agreement Number 776613) Research Programmes. JR was supported by the National Environmental Research Council (NERC) 'North Atlantic Climate System Integrated Study' (ACSIS) program (Grant Agreement Number NE/N018001/1) and NCAS. The climate model simulations are available via the Earth System Grid Federation (ESGF) archive of Coupled Model Intercomparison Project 6 (CMIP6) data, for instance on https://esgf-index1.ceda.ac.uk/projects/esgf-ceda/ . This work used JASMIN, the UK's collaborative data analysis environment (http://jasmin.ac.uk).

ORCID iDs

Matthew B Menary https://orcid.org/0000-0002-9627-2056
Juliette Mignot https://orcid.org/0000-0002-4894-898X
Jon Robson https://orcid.org/0000-0002-3467-018X

References

Ba J *et al* 2014 A multi-model comparison of Atlantic multidecadal variability *Clim. Dyn.* **43** 2333–48
Balaji V *et al* 2018 Requirements for a global data infrastructure in support of CMIP6 *Geosci. Model Dev.* **11** 3659–80
Bindoff N L *et al* 2013 Chapter 10: detection and attribution of climate change: from global to regional *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (available at: https://www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5_Chapter10_FINAL.pdf)
Boer G J *et al* 2016 The Decadal Climate Prediction Project (DCPP) contribution to CMIP6 *Geosci. Model Dev.* **9** 3751–777
Booth B B B, Dunstone N J, Halloran P R, Andrews T and Bellouin N 2012 Aerosols implicated as a prime driver of

twentieth-century North Atlantic climate variability *Nature* **484** 228–32
Borchert L F, Menary M B, Swingedouw D, Sgubin G, Hermanson L and Mignot J 2021 Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6 *Geophys. Res. Lett.* **48** e2020GL091307
Born A, Mignot J and Stocker T F 2015 Multiple equilibria as a possible mechanism for decadal variability in the North Atlantic Ocean *J. Clim.* **28** 8907–22
Brown P T and Caldeira K 2020 Empirical prediction of short-term annual global temperature variability *Earth Space Sci.* **7** e2020EA001116
Brunner L *et al* 2020 Comparing methods to constrain future European climate projections using a consistent framework *J. Clim.* **33** 8671–92
Collins M *et al* 2006 Interannual to decadal climate predictability in the North Atlantic: a multimodel-Ensemble Study *J. Clim.* **19** 1195–203
Desbruyères D G, Mercier H, Maze G and Daniault N 2019 Surface predictor of overturning circulation and heat content change in the Subpolar North Atlantic *Ocean Sci.* **15** 809–17
Deser C, Terray L and Phillips A S 2016 Forced and internal components of winter air temperature trends over North America during the past 50 years: mechanisms and implications *J. Clim.* **29** 2237–58
Ding H, Newman M, Alexander M A and Wittenberg A T 2018 Skillful climate forecasts of the Tropical Indo-Pacific Ocean using model-analogs *J. Clim.* **31** 5437–59
Ding H, Newman M, Alexander M A and Wittenberg A T 2019 Diagnosing secular variations in retrospective ENSO seasonal forecast skill using CMIP5 model-analogs *Geophys. Res. Lett.* **46** 1721–30
Dool H M V D 1994 Searching for analogues, how long must we wait? *Tellus* A **46** 314–24
Dunstone N J, Smith D M and Eade R 2011 Multi-year predictability of the Tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean *Geophys. Res. Lett.* **38** L14701
Eden J M, Van Oldenborgh G J, Hawkins E and Suckling E B 2015 A global empirical system for probabilistic seasonal climate prediction *Geosci. Model Dev.* **8** 3947–73
Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J and Taylor K E 2016 Overview of the coupled model Intercomparison Project phase 6 (CMIP6) experimental design and organization *Geosci. Model Dev.* **9** 1937–58
Gillett N P, Shiogama H, Funke B, Hegerl G, Knutti R, Katja Matthes K, Santer B D, Stone D and Tebaldi C 2016 The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6 *Geosci. Model Dev.* **9** 3685–97
Hawkins E, Robson J, Sutton R, Smith D and Keenlyside N 2011 Evaluating the potential for statistical decadal predictions of sea surface temperatures with a perfect model approach *Clim. Dyn.* **37** 2495–509
Hermanson L, Eade R, Robinson N H, Dunstone N J, Andrews M B, Knight J R, Scaife A A and Smith D M 2014 Forecast cooling of the Atlantic subpolar gyre and associated impacts *Geophys. Res. Lett.* **41** 5167–74
Ho C K, Hawkins E, Shaffrey L and Underwood F M 2013 Statistical decadal predictions for sea surface temperatures: a benchmark for dynamical GCM predictions *Clim. Dyn.* **41** 917–35
Kerr R A 2000 A North Atlantic climate pacemaker for the centuries *Science* **288** 1984–5
Knudsen M F, Seidenkrantz M-S, Jacobsen B H and Kuijpers A 2011 Tracking the Atlantic multidecadal oscillation through the last 8 000 years *Nat. Commun.* **2** 178
Kushnir Y *et al* 2019 Towards operational predictions of the near-term climate *Nat. Clim. Change* **9** 94–101
Lawrence B N, Bennett V L, Churchill J, Juckes M, Kershaw P, Pascoe S, Pepler S, Pritchard M and Stephens A 2013 Storing and manipulating environmental big data with JASMIN *2013 IEEE Int. Conf. Big Data* pp 68–75
Lorenz E N 1969 Atmospheric predictability as revealed by naturally occurring analogues *J. Atmos. Sci.* **26** 636–46
Menary M B, Hodson L R, Robson J I, Sutton R T, Wood R A and Hunt J A 2015 Exploring the impact of CMIP5 model biases on the simulation of North Atlantic decadal variability *Geophys. Res. Lett.* **42** 5926–34
Menary M B and Wood R A 2018 An Anatomy of the projected North Atlantic warming hole in CMIP5 models *Clim. Dyn.* **50** 3063–80
Moat B I *et al* 2020 Pending recovery in the strength of the meridional overturning circulation at 26° N *Ocean Sci.* **16** 863–74
Monerie P-A, Robson J, Dong B and Dunstone N 2018 A role of the Atlantic Ocean in predicting summer surface air temperature over North East Asia? *Clim. Dyn.* **51** 473–91
Müller W A, Pohlmann H, Sienz F and Smith D 2014 Decadal climate predictions for the period 1901–2010 with a coupled climate model *Geophys. Res. Lett.* **41** 2100–7
Newman M 2013 An empirical benchmark for decadal forecasts of global surface temperature anomalies *J. Clim.* **26** 5260–9
O'Neill B C *et al* 2016 The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6 *Geosci. Model Dev.* **9** 3461–82
O'Reilly C H, Woollings T and Zanna L 2017 The dynamical influence of the Atlantic multidecadal oscillation on continental climate *J. Clim.* **30** 7213–30
Oldenborgh J V, Geert M A B, Ferranti L, Stockdale T N and Anderson D L T 2005 Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years? *J. Clim.* **18** 3240–9
Rayner N A, Parker D E, Horton E B, Folland C K, Alexander L V, Rowell D P, Kent E C and 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. (1984–2012)* **108** 4407
Smith D M, Eade R and Pohlmann H 2013 A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction *Clim. Dyn.* **41** 3325–38
Suckling E B and Smith L A 2013 An evaluation of decadal probability forecasts from state-of-the-art climate models *J. Clim.* **26** 9334–47
Suckling E B, Van Oldenborgh G J, Eden J M and Hawkins E 2017 An empirical model for probabilistic decadal prediction: global attribution and regional hindcasts *Clim. Dyn.* **48** 3115–38
Sutton R T and Hodson D L R 2005 Atlantic Ocean forcing of North American and European summer climate *Science* **309** 115–8
Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design *Bull. Am. Meteorol. Soc.* **93** 485–98
Vellinga M and Wu P L 2004 Low-latitude freshwater influence on centennial variability of the Atlantic thermohaline circulation *J. Clim.* **17** 4498–511
Yiou P and Déandréis C 2019 Stochastic ensemble climate forecast with an analogue model *Geosci. Model Dev.* **12** 723–34
Zanna L 2012 Forecast skill and predictability of observed Atlantic Sea surface temperatures *J. Clim.* **25** 5047–56