Response of Arctic Freshwater to the Arctic Oscillation in Coupled Climate Models

Sam B. Cornish, Yavor Kostov, Helen L. Johnson, Camille Lique

Response of Arctic Freshwater to the Arctic Oscillation in Coupled Climate Models

SAM B. CORNISH

Department of Earth Sciences, University of Oxford, Oxford, United Kingdom

YAVOR KOSTOV[a]

Department of Physics, University of Oxford, Oxford, United Kingdom

HELEN L. JOHNSON

Department of Earth Sciences, University of Oxford, Oxford, United Kingdom

CAMILLE LIQUE

Université de Bretagne Occidentale, CNRS, IRD, Ifremer, Laboratoire d'Océanographie Physique et Spatiale, IUEM, Brest, France

(Manuscript received 13 September 2019, in final form 21 December 2019)

ABSTRACT

The freshwater content (FWC) of the Arctic Ocean is intimately linked to the stratification—a physical characteristic of the Arctic Ocean with wide relevance for climate and biology. Here, we explore the relationship between atmospheric circulation and Arctic FWC across 12 different control-run simulations from phase 5 of the Coupled Model Intercomparison Project. Using multiple lagged regression, we seek to isolate the linear response of Arctic FWC to a step change in the strength of the Arctic Oscillation (AO) as well as the second and third orthogonal modes of SLP variability over the Arctic domain. There is broad agreement among models that a step change to a more anticyclonic AO leads to an increase in Arctic FWC, with an *e*-folding time scale of 5–10 yr. However, models differ widely in the degree to which a linear response to SLP variability can explain FWC changes. Although the mean states, time scales, and magnitudes of FWC variability may be broadly similar, the physical origins of variability are highly inconsistent among models. We perform a robustness test that incorporates a Monte Carlo approach to determine which response functions are most likely to represent causal, physical relationships within the models and which are artifacts of regression. Convolution with SLP reanalysis data shows that the four most robust response functions have some skill at reproducing observed accumulation of FWC during the late 1990s and 2000s, consistent with the idea that this change was largely wind driven.

1. Introduction

The physical dynamics of the Arctic encompass atmospheric, oceanic, and cryospheric processes; systems that are interlinked at the surface of the ocean. The Arctic Ocean is stratified by salinity—it is a so-called β ocean (Carmack 2007)—and therefore freshwater in the Arctic represents an important control on these linkages, because its abundance modulates the connections between the surface of the ocean and the relatively warm Atlantic Water below. Changes to the freshwater reservoir, and therefore the stratification, have implications for sea ice stability and growth, vertical heat fluxes, mixing of nutrients, and the carbon cycle and have broad indirect implications for climate and biology [for an overview, see Carmack et al. (2016)]. Any attempt to understand the future evolution of the Arctic climate system would benefit from an improved representation of the processes governing freshwater variability (Lique et al. 2016). Furthermore, changes to Arctic freshwater content (FWC) should be closely related to changes in the supply and export of freshwater, though these links have proven difficult to pin down in observations

[a] Current affiliation: College of Life and Environmental Sciences, University of Exeter, Exeter, United Kingdom.

Corresponding author: Sam Cornish, sam.cornish@earth.ox.ac.uk

DOI: 10.1175/JCLI-D-19-0685.1

(e.g., Haine et al. 2015). Enhanced freshwater export to the North Atlantic has the potential to affect the Atlantic meridional overturning circulation (AMOC) by modifying the salinity and stratification in regions of deep water formation (Jahn and Holland 2013; Wang et al. 2018) and/or altering the densities along boundary currents in the subpolar gyre (Pickart and Spall 2007).

The Arctic Ocean maintains a relatively fresh state as a result of its position at the northerly node of the global hydrological cycle (Carmack et al. 2016). Fluxes of freshwater from Eurasian and North American rivers, flow of fresh Pacific Water through the shallow Bering Strait, and an imbalance of precipitation and evaporation supply freshwater to the Arctic at a net rate of approximately $10\,000\,\mathrm{km^3\,yr^{-1}}$ (Haine et al. 2015). The freshwater reservoir of the Arctic Ocean and Canadian Arctic Archipelago (CAA) comprises both liquid ($\sim$100 000 $\mathrm{km^3}$) and solid ($\sim$14 000 $\mathrm{km^3}$) components (Haine et al. 2015), unevenly distributed across the Arctic basins, shelf regions, and the CAA. Both the solid and liquid components show a pronounced seasonal cycle, and both the regional distribution of liquid freshwater and the total liquid Arctic FWC vary significantly on interannual to decadal time scales (e.g., Polyakov et al. 2008).

A prerequisite of improving simulated FWC variability is the ability to assess and evaluate existing simulations. Indeed, we must go beyond comparing the mean state and variability of the Arctic FWC reservoir in climate models and evaluate the physical relationships between drivers and responses. Here, we isolate what is considered to be an important mechanism in the natural variability of Arctic FWC across 12 different model simulations from phase 5 of the Coupled Model Intercomparison Project (CMIP5): the response of Arctic FWC to changes in the dominant patterns of large-scale atmospheric circulation over the Arctic domain. These atmospheric circulation patterns are described by the leading modes of sea level pressure (SLP) variability. In section 2 we define the time series we use and provide background to the relevant climatology in each model.

It has been understood for some time that winds can drive the redistribution and export of freshwater, which resides at the surface of the Arctic Ocean, whether as liquid or sea ice (Proshutinsky and Johnson 1997; Proshutinsky et al. 2009). Atmospheric circulation affects the pathways and fluxes of inflowing Pacific Waters (Alkire et al. 2007; Steele et al. 2004) and low-salinity waters from the Siberian shelves (Newton et al. 2008; Timmermans et al. 2011) as well as more saline Atlantic Waters (Morison et al. 2006; Muilwijk et al. 2019). These connections are mediated by the sea ice cover, and its decline over recent decades may have contributed to the increased salinity contrast between the (fresher) Amerasian basin and the Eurasian basin (Wang et al. 2019). Redistribution of liquid freshwater within the Arctic is itself important for establishing geostrophic circulation via the tilting of the halocline; these dynamics are perhaps most clear in the Beaufort Gyre, the largest reservoir of FWC in the Arctic Ocean (e.g., Manucharyan and Spall 2016).

As the leading mode of extratropical sea level pressure variability in the Northern Hemisphere, the Arctic Oscillation (AO; Thompson and Wallace 1998) captures the dominant variability in the large-scale wind forcing and therefore represents a natural starting point in identifying causal relationships between atmospheric and oceanic/cryospheric variability. Indeed, several studies have focused on determining the influence of the AO on patterns of sea ice drift (Rigor et al. 2002; Kwok 2009; Kwok et al. 2013; Armitage et al. 2018) and surface geostrophic circulation (Morison et al. 2012; Armitage et al. 2018). There has, however, been substantial difficulty in linking observed Arctic FWC changes to atmospheric variability (Rabe et al. 2014). Much of this difficulty arises from the fact that the ocean responds slowly to an atmospheric perturbation. Theory and idealized modeling suggest that FWC in the Arctic Ocean, or at least in its largest reservoir the Beaufort Gyre, bears a multiyear to decadal memory of past atmospheric forcing (Davis et al. 2014; Manucharyan and Spall 2016; Manucharyan and Isachsen 2019; Doddridge et al. 2019).

To appropriately capture a relationship where memory is important, we use linear response theory. Our method, following the approach of Kostov et al. (2017, 2018) and Johnson et al. (2018), yields impulse response functions and their time integrals, step response functions. In section 3, using the preindustrial control run of each model, we isolate the linear response of Arctic FWC to a 1-standard-deviation change in the strength of the AO, and the second and third orthogonal modes of SLP variability. Through convolution of the derived linear response functions with the original SLP, we attempt to reconstruct the FWC time series of each model. To the extent that the derived relationships are physically robust, we are then able to evaluate both the nature of the FWC–SLP relationship and its importance in determining overall FWC variability in each model. This intercomparison is the first objective of the paper.

The second objective of the paper is to probe the caveats associated with a regression-based technique and establish a technique for its assessment. In section 3b

TABLE 1. Models used and their characteristics. Grid resolutions are displayed as longitude × latitude, followed by the number of vertical depth levels. Sea ice grids match the ocean component, except for the sea ice grid of CanESM2, which matches the atmosphere component. Full definitions of model names can be found online (https://www.ametsoc.org/PubsAcronymList).

Model	Atmosphere grid	Ocean grid	Sea ice model	Reference
ACCESS1.0	1.88° × 1.25°	Tripolar; 1° × 1°; 50 levels	CICE, v4.1	Dix et al. (2013)
ACCESS1.3	1.88° × 1.25°	Tripolar; 1° × 1°; 50 levels	CICE, v4.1	Dix et al. (2013)
CanESM2	2.81° × 2.79°	1.41° × 0.94°; 40 levels	CanSIM1	Chylek et al. (2011)
CCSM4	1.25° × 0.94°	Dipolar; 1.11° × (0.27–0.54)°; North Pole in Greenland; 60 levels	CICE, v4.0	Gent et al. (2011)
CNRM-CM5	1.4° × 1.4°	ORCA-1°; tripolar; 42 levels	GELATO (v5)	Voldoire et al. (2013)
GFDL CM3	2.5° × 2°	Tripolar; ~1° × 1°; 50 levels; tripolar	SISp2	Griffies et al. (2011)
GFDL-ESM2M	2.50° × 2.02°	Tripolar; ~1° × 1°; 50 levels	SISp2	Dunne et al. (2012)
IPSLA-CM5A-LR	3.75° × 1.89°	ORCA-2°; tripolar; 31 levels	LIM2	Dufresne et al. (2013)
IPSLA-CM5A-MR	2.5° × 1.27°	ORCA-2°; tripolar; 31 levels	NEOM-LIM2	Dufresne et al. (2013)
MIROC5	1.41° × 1.4°	Shifted poles; ~1.4° × (0.5–1.4)°; 50 levels	COCO4.5	Watanabe et al. (2010)
MPI-ESM-LR	1.88° × 1.87°	Shifted poles; ~1.5° × 1.5°; 40 levels	MPIOM	Giorgetta et al. (2013)
MPI-ESM-MR	1.88° × 1.87°	Shifted poles; ~0.4° × 0.4°; 40 levels	MPIOM	Giorgetta et al. (2013)

we document a Monte Carlo approach that we employ to determine levels of statistical significance for our response functions. In section 4, we establish an external test on the robustness of the functions by utilizing the historical run of each model and applying further Monte Carlo tests. In section 5, after this evaluation, we select and describe the response functions that are most likely to represent physical relationships within the models from which they are derived.

The third and final objective is to evaluate whether the model-derived relationships selected in section 5 are at all an accurate reflection of the real-world relationship between Arctic FWC and SLP. In section 6, we convolve the model-derived impulse response functions with reanalysis SLP data from ERA-20C (Poli et al. 2016) and ERA-Interim (Berrisford et al. 2011). This yields a comparable FWC accumulation in the late 1990s–2000s to that reported by Rabe et al. (2014) and Polyakov et al. (2013). In this context, our method is a novel means for evaluating coupled models with freely evolving atmospheric components; this analysis provides information that cannot be obtained from historical simulations alone.

2. Models used and time series definitions

We use monthly mean data from 12 different model contributions to the CMIP5 ensemble (Table 1). For the principal analysis, we use the preindustrial control runs of each model because these are long, unforced runs in which the internal variability is most likely to be characterized by stationary processes. We later introduce the historical runs of each model as an independent test on the robustness of the derived relationships. Following the definitions below, we compute one-dimensional time series for our analysis: deseasonalized Arctic freshwater content and principal component time series of SLP variability north of 70°N.

a. Representation of Arctic freshwater

FWC in the Arctic Ocean is commonly defined relative to a reference salinity, S_{ref}, chosen as the mean salinity of the Arctic: 34.8, such that $S_{\text{ref}} = 34.8$ (Aagaard and Carmack 1989):

$$\text{FWC}(t) = \iiint_{z(S=S_{\text{ref}})}^{0} \frac{S_{\text{ref}} - S}{S_{\text{ref}}}\, dV. \tag{1}$$

This definition also physically corresponds to the salinity at a depth near the base of the Arctic halocline, meaning that it has real utility in quantifying the expansion and freshening of this upper fresh layer. Because the integration extends only as far down as the halocline, changes below the halocline—for instance, from changing Atlantic Water inflow—are not considered. In this analysis, we take the reference salinity as the mean Arctic Ocean salinity for each model in question (Table 2), in order to avoid the effects of salinity biases in the models relative to the real Arctic. In each model simulation, these reference salinities all lie near the base of the halocline (not shown).

In constructing our freshwater budget for each model, we exclude the Canadian Arctic Archipelago, which is represented very differently between models. As such, we use the same Arctic Ocean domain as chosen by Serreze et al. (2006) in their observational synthesis (Fig. 2). Note that the archipelago is often included in other Arctic freshwater budgets (e.g., Haine et al. 2015). The model simulations show a range of FWC mean states (Fig. 1), with a factor-of-2 difference between the ensemble members with the greatest and smallest FWC mean states. For reference, Serreze et al.

TABLE 2. Preindustrial-control-run Arctic FWC characteristics. The standard deviation is taken after deseasonalizing and linearly detrending the FWC time series. The decorrelation time scale is the *e*-folding time scale of the autocorrelation function, given to the nearest year.

Model simulation	S_{ref} (mean Arctic salinity)	Mean FWC (km^3)	Std dev (km^3)	Decorrelation time scale (yr)
ACCESS1.0	34.64	78 000	4100	14
ACCESS1.3	34.71	77 000	3500	21
CanESM2	34.71	109 000	3700	7
CCSM4	34.77	75 000	2500	6
CNRM-CM5	34.40	107 000	2800	19
GFDL CM3	34.65	90 000	2600	6
GFDL-ESM2M	34.73	76 000	2600	9
IPSLA-CM5A-LR	34.63	128 000	3000	9
IPSLA-CM5A-MR	34.68	136 000	3100	12
MIROC5	34.79	64 000	2800	7
MPI-ESM-LR	34.69	80 000	2300	13
MPI-ESM-MR	34.61	82 000	2400	8

(2006) estimate the pre-2000 climatological mean liquid FWC to be 74 000 ± 7400 km^3, based on hydrographic observations. Intermodel comparisons of absolute values of freshwater as commonly defined in the literature are ambiguous (Schauer and Losch 2019). To provide parity, we report our linear response functions in terms of the fractional change in each model's Arctic FWC reservoir.

Some of the preindustrial control simulations exhibit a clear trend in FWC through the run (CanESM2, CNRM-CM5, and MIROC5). In all cases, we linearly detrend the FWC time series before performing the analysis. We group model runs into lengths of 3600, 6000, and 9600 months to standardize our regression procedure (section 3a). Where we truncate model runs to fit this grouping scheme, we begin from the first month in all cases except MIROC5 and MPI-ESM-MR, which display trends in the early part of the runs that are inconsistent with the rest of the time series. In these two simulations we omit the first 1008 months.

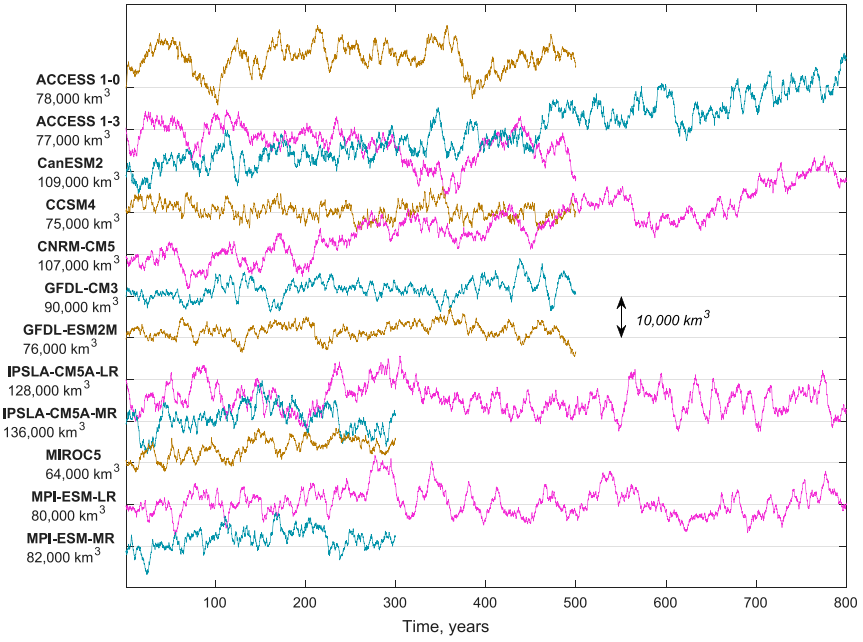

FIG. 1. Deseasonalized Arctic FWC time series in 12 CMIP5 preindustrial control runs. Model simulations are categorized into lengths of 3600, 6000 and 9600 months (see text). Mean FWC is given beneath the name of each simulation. The magnitude of variability can be seen by comparison with *y*-axis spacing of 10 000 km^3. The first value of each time series is set to this spacing; the lines do not denote the time series mean.

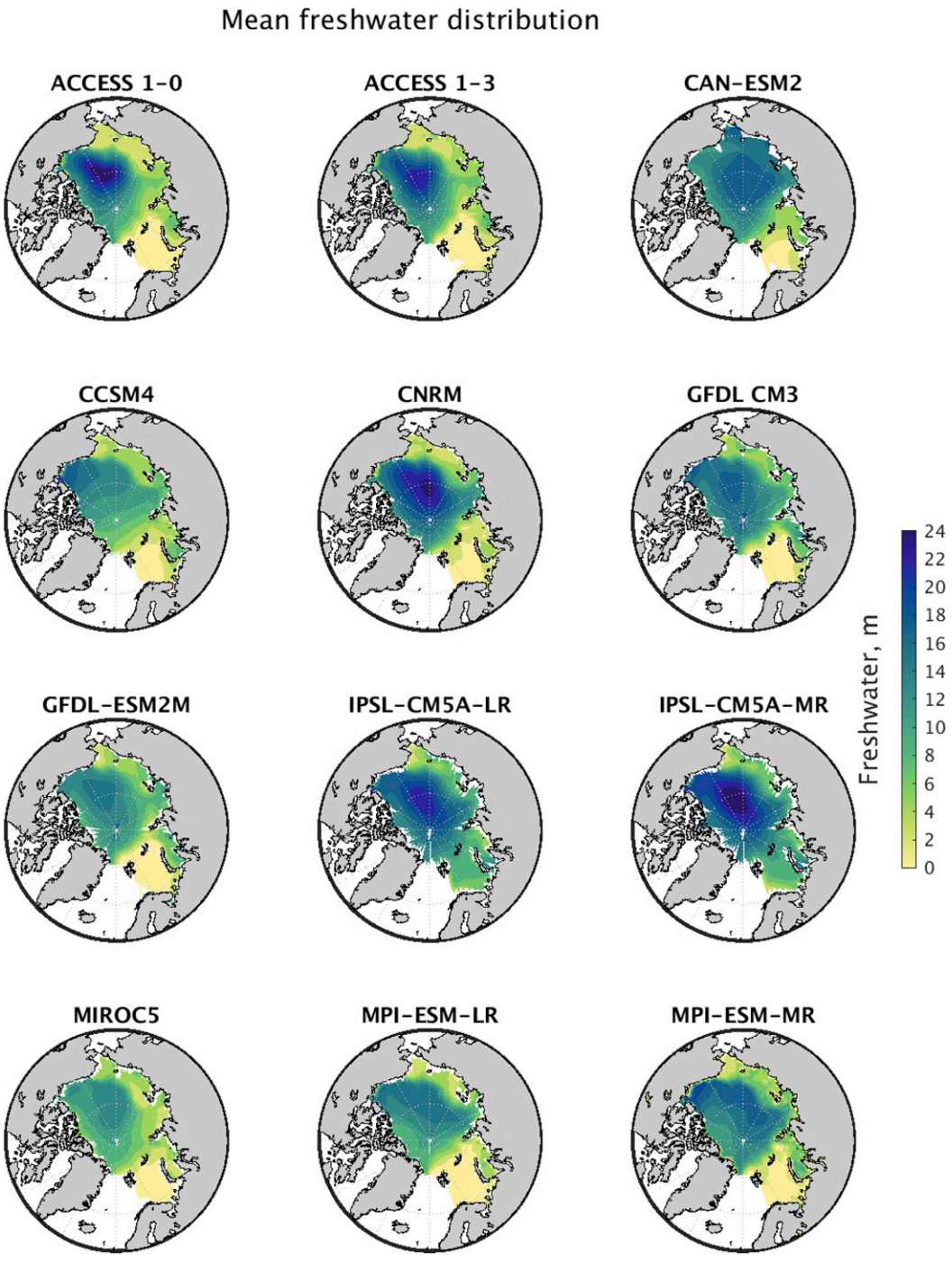

FIG. 2. Mean freshwater distribution (m) over the Arctic Ocean domain considered in this paper. The mean is taken over a 50-yr period that is selected randomly from within the preindustrial control runs.

The mean freshwater distribution (Fig. 2) and hydrography (not shown) within the Arctic also varies between model simulations. Broadly, however, the models capture the greater depth-integrated freshwater storage in the deep basins compared with the shelves, a greater FWC and deeper halocline in the Amerasian basin than the Eurasian basin, and a saline and weakly stratified Barents Sea. The models show some common biases with respect to observations (e.g., the Polar Science Center Hydrographic Climatology; Steele et al. 2001): the Beaufort Gyre is more spatially diffuse, situated farther from Canada and Alaska, and

the central Arctic is overly fresh. These biases are also clear in the CMIP5 multimodel mean freshwater distribution (Shu et al. 2018). Additionally, the IPSL models show an overly fresh Barents Sea (Fig. 2). Intercomparison of Arctic Ocean hydrography in a suite of coupled climate models was last undertaken by Holland et al. (2007), who examined 10 models contributing to the Intergovernmental Panel on Climate Change Fourth Assessment Report. Ding et al. (2016) examined the processes governing the seasonal freshwater cycle in CMIP5 models, but otherwise the only intercomparison of Arctic Ocean freshwater and/or hydrography in models of the same generation of the CMIP5 models thus far comes from Arctic Ocean CORE-II studies, which considered forced (noncoupled) model simulations (Wang et al. 2016; Ilicak et al. 2016). Shu et al. (2018) assess the projected changes of the CMIP5 multimodel mean freshwater content in the twenty-first century.

b. *Representation of atmospheric variability*

The Arctic Oscillation is a winter-intensified, annular mode of sea level pressure variability that is linked to the polar vortex (Thompson and Wallace 1998). The AO correlates strongly with the North Atlantic Oscillation (NAO), especially during winter (Deser 2000), but affects climate beyond the Arctic and indeed beyond the Atlantic sector (Thompson and Wallace 2001). The positive (negative) mode of the AO involves a strengthening of cyclonic (anticyclonic) atmospheric circulation over the Arctic. In this analysis, we restrict our attention to sea level pressure variability north of 70°N and define the AO as its leading empirical orthogonal function (EOF). Figure 3 shows the AO as defined above, using ERA-Interim reanalysis data provided by the European Centre for Medium-Range Weather Forecasts (Berrisford et al. 2011). As plotted, the positive phase of this EOF corresponds to the anticyclonic, negative mode of the Arctic Oscillation. The first EOF (EOF1) of each model simulation considered here shows a high degree of spatial correlation with this definition of the AO (Table 3), and also with EOF1 from the other models. The first principal component (PC1) time series for each model also show similar spectral densities to one another and to the PC1 time series from ERA-Interim and ERA-20C data (not shown). The spatial patterns of the second and third EOFs, by contrast, vary considerably among models (see also Cai et al. 2018). Furthermore, as the variance explained by PC1 is substantially higher than that explained by PC2 and PC3 in all cases (Table 3), EOF1 must be close to the true underlying leading mode. In contrast, the variances explained by PC2 and PC3 are closer together (not shown), raising the possibility that our estimated EOF2 and EOF3 might be linear combinations of the true underlying second and third modes (North et al. 1982). As such, we only compare the response functions for PC1 in this paper, though we also calculate the response functions for PC2 and PC3 as part of the analysis (section 3a).

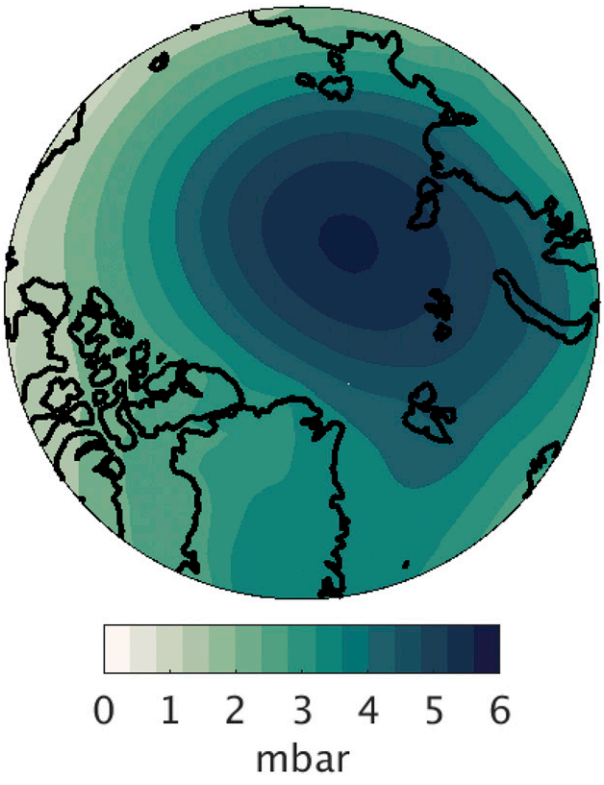

FIG. 3. The negative AO as represented by ERA-Interim (Berrisford et al. 2011). Calculated as the first EOF of SLP variability north of 70°N, this mode explains 53% of the variance in ERA-Interim SLP data spanning 1979–2018.

3. FWC–SLP relationships in CMIP5 preindustrial control runs

a. *Linear response function method*

We employ linear response theory (e.g., Hasselmann et al. 1993) as per Kostov et al. (2017) to determine the response of the Arctic FWC reservoir to a one-standard-deviation step change in the strength of the negative (anticyclonic) AO. We also determine the responses of FWC to one-standard-deviation step changes in the second and third EOFs of SLP. Although these functions are employed in our reconstructions, we do not directly compare them for reasons given in section 2b. Our approach is a statistical alternative to active model

TABLE 3. Preindustrial-control-run AO characteristics.

Model simulation	SLP % variance explained by EOF1	SLP % variance explained by first three EOFs	EOF1 spatial correlation (R^2) with ERA-Interim EOF1
ACCESS1.0	63	86	0.96
ACCESS1.3	65	87	0.96
CanESM2	67	87	0.93
CCSM4	73	90	0.87
CNRM-CM5	68	87	0.95
GFDL CM3	70	87	0.90
GFDL-ESM2M	68	89	0.93
IPSLA-CM5A-LR	73	90	0.97
IPSLA-CM5A-MR	70	89	0.96
MIROC5	68	87	0.94
MPI-ESM-LR	69	88	0.96
MPI-ESM-MR	71	89	0.97

perturbation experiments (see Marshall et al. 2017) and permits the investigation and comparison of fully coupled models.

First, we define our time series of interest. The findings of Johnson et al. (2018) indicate that all of the first three PCs of SLP are significant in explaining freshwater variability—in both a coupled model context and most likely the real Arctic. As a result, we choose to consider the first three PCs of SLP and by design allow the combination of these three orthogonal modes to maximize the Arctic FWC variance explained by SLP. For each model simulation, we treat the time series of Arctic FWC as the sum of convolutions of the first three PCs of atmospheric forcing, PC_i with $n = 3$, and their respective unknown impulse response functions G_i, integrated over time lags τ, from any given time t, to $\tau_{\max}$ years prior (with $d\tau = 1$ month), plus a residual $\varepsilon(t)$ that varies in time. We choose $\tau_{\max}$ as 20 years:

$$\mathrm{FWC}(t) = \sum_{i=1}^{n} \int_{\tau=0}^{\tau_{\max}} G_i(\tau)\mathrm{PC}_i(t-\tau)\, d\tau + \varepsilon(t). \qquad (2)$$

We solve for the three impulse response functions G_i simultaneously using multiple lagged regression. We can then attempt to reconstruct the original FWC time series by convolving the impulse response functions with the forcing; that is,

$$\mathrm{FWC}(t)^{\mathrm{recon}} = \sum_{i=1}^{n} \int_{\tau=0}^{\tau_{\max}} G_i(\tau)\mathrm{PC}_i(t-\tau)\, d\tau.$$

The fit between the reconstruction and the original time series indicates how much FWC variability in the model can be explained as a linear response to the first three modes of SLP, over lags of up to 20 years. (See appendix Fig. A1 for an example in the CCSM4 control run.) The misfit, on the other hand, gives us the residual term; that is, $\varepsilon(t) = \mathrm{FWC}(t) - \mathrm{FWC}(t)^{\mathrm{recon}}$. This residual noise is employed in the uncertainty estimation procedure [section 3b(2)].

For each model simulation, we attempt to minimize overfitting by assembling over 1000 different estimates for the impulse response functions. We do this by varying (i) the cutoff lag applied to our impulse response functions, between 20 and 30 years; (ii) the part of the run that we select for the regression; and (iii) the length of this part. We group the models by length and adopt a consistent approach to subsampling the runs for all models (see Table A1).

The final impulse response function is a mean of all estimates over time lags up to 20 years (we terminate impulses with greater cutoff lags at a lag of 20 years in order that all response functions used are the same length). We then integrate through time lags to obtain step response functions, also known as climate response functions (CRFs; e.g., Marshall et al. 2017; Muilwijk et al. 2018). We next evaluate the statistical robustness of these response functions using a Monte Carlo procedure (section 3b).

This technique follows that of Kostov et al. (2017, 2018) and Johnson et al. (2018) and the High-Resolution Global Environmental Model (HiGEM) contribution to the multimodel CRF study by Muilwijk et al. (2019). In this contribution, we develop the method through the addition of (i) a simultaneous approach to finding the impulse responses to each PC, (ii) Monte Carlo evaluations of response function error and significance, and (iii) the application of another run from the same model as an independent test of robustness.

b. *Monte Carlo procedure*

1) STATISTICAL SIGNIFICANCE

Regression techniques may produce relationships that do not reflect a causal mechanism but are simply a

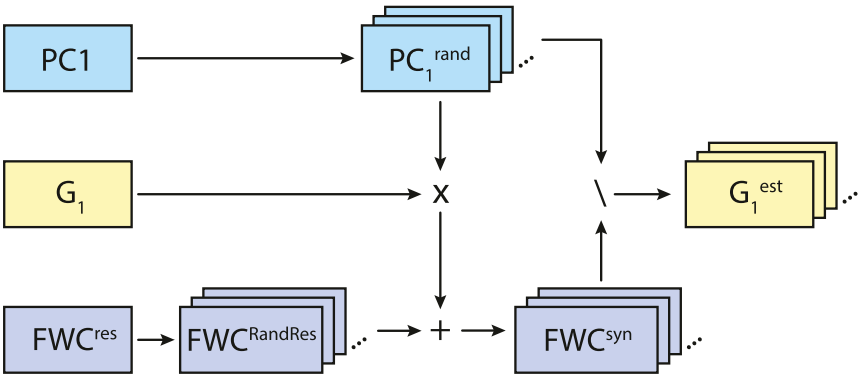

FIG. 4. Flowchart illustrating our Monte Carlo method for calculating the error on the response functions. Repeated boxes indicate the presence of 1000 versions of the given time series. $PC1^{rand}$ and $FWC^{RandRes}$ are generated through randomization of PC1 and FWC^{res}, respectively. The randomization process is described in section 3b(1). The multiplication sign indicates convolution [first term on the RHS of Eq. (3)], the plus sign refers to the addition and rescaling process that completes the RHS of Eq. (3), and the backslash is shorthand for the multiple lagged regression used to estimate impulse responses.

statistical artifact of the attempt to best fit the target data. This is an issue because it is unknown a priori how much of the variance the forcing should be able to explain; the impulse response functions will inevitably fit variability that is unrelated to the forcing, a problem that becomes exacerbated as forcings explain progressively less variance in the target data.

Monte Carlo methods have become a standard tool for the statistical testing of regression-derived relationships in the climate sciences (e.g., Lund 1970) and beyond. Here, we use a Monte Carlo approach to help establish the statistical significance of the relationships derived by our linear response function approach.

We repeat our method for finding the response functions and computing reconstructions but exchange the models' PC1 for a randomized forcing time series. The randomized, or surrogate, time series are generated by phase randomization as described in Schreiber and Schmitz (2000), and are designed to have the same spectral properties as the models' PC1 time series. Periodicity artifacts are minimized by first making each time series periodic by appending its reflection in time, then using only half the length of the resulting surrogate time series. We use the same procedure on every occasion that we generate randomized or surrogate time series. We perform 1000 such trials, and also repeat the procedure for the case where we have three simultaneous random forcing time series (emulating our three PCs).

We can then ask the following question for both the case of PC1 alone and all three PCs combined: Does the FWC reconstruction for each model (calculated as in section 2a) explain more variance than that which could be explained using random forcing, at the 95% certainty level?

If a response function derived from the original model SLP forcing explains more FWC variance than the 95th-percentile value from the trials with random forcing, it is likely that the response *is* physical and not an artifact of the regression method. The true relationship between a given mode of atmospheric forcing and the FWC response in any given model may explain a low degree of variance (lower than the significance level) in the FWC time series and still be physical. Note, however, that if a low-variance explaining relationship exists, its representation in the impulse response function may be distorted by the attempt to best fit the residual noise.

2) RESPONSE FUNCTION UNCERTAINTY

To establish uncertainty estimates for our linear response functions, we again employ a Monte Carlo approach. We start with the expectation that a given response function *is* physical. We find the fraction of the model simulation's FWC variance [R^2 in Eq. (3)] that is explained by this response function (as per section 2a) and determine the spectral density of the residual noise. We then create an ensemble of 1000 synthetic FWC time series, which is generated in the following manner. Figure 4 also illustrates this method.

As shown in Eq. (3), each synthetic time series contains a mixture of the FWC time series derived from (i) the convolution of our original response function G_1 with randomized SLP forcing PC_1^{rand}, the resulting time series of which has a standard deviation $\sigma_{G\times PC}$; and (ii) a random time series $FWC(t)^{RandRes}$

that has the same spectral density as the residual noise $\varepsilon(t)$ and has a standard deviation σ_{RandRes}. Note that the abbreviation RandRes stands for randomized residual. We scale the two time series in accordance with the appropriate division of variance explained and sum them to create the synthetic FWC time series $\text{FWC}(t)^{\text{syn}}$:

$$\text{FWC}(t)^{\text{syn}} = \int_{\tau=0}^{\tau_{\max}} G_1(\tau)\text{PC}_1^{\text{rand}}(t-\tau)\,d\tau + \sqrt{\frac{1-R^2}{R^2}} \times \left[\frac{\text{FWC}(t)^{\text{RandRes}}}{\sigma_{\text{RandRes}}/\sigma_{G\times\text{PC}}}\right]. \tag{3}$$

Taking this ensemble of 1000 synthetic FWC time series, $\text{FWC}(t)^{\text{syn}}$, and 1000 corresponding randomized forcing time series, $\text{PC}_1^{\text{rand}}$, we use multiple lagged regression to estimate an impulse response function, G_1^{est}, that best relates the two time series, as per $\text{FWC}(t)^{\text{syn}} = \int_{\tau=0}^{\tau_{\max}} G_1^{\text{est}}(\tau)\text{PC}_1^{\text{rand}}(t-\tau)\,d\tau$.

The resulting ensemble of 1000 response functions, G_1^{est}, answers the motivating question: How reliably can we "back out" a physical response function $G_1(\tau)$when it only explains a certain degree of variance R^2 and the residual noise $\text{FWC}(t)^{\text{RandRes}}$ has a given spectral density? In each case, the mean of the 1000 estimates faithfully reproduces the original response function $G_1(\tau)$. However, the spread may be large. The 1-standard-deviation error bars for the response functions that we derive from the preindustrial control runs are then given by the standard deviation of the 1000 response functions in this ensemble.

The magnitude of the 1-standard-deviation uncertainty on each response function is related to both the variance explained and the spectral properties of the residual noise with which we dilute the target time series (appendix Fig. C1). For more information on how the spectral properties affect error, see appendix C.

c. Results

In Fig. 5 we show the estimated responses of Arctic FWC in each model simulation to a step change in the strength of the negative AO. Though confidence in the results varies across the ensemble, we note the following. While the simulations exhibit a range of responses, there is broad agreement that the freshwater reservoir grows in size following a step change to a more negative (anticyclonic) AO. All models show a positive response after 10 years. Only GFDL-ESM2M and IPSL-CM5A-LR show negative mean responses by the end of the 20-yr cutoff lag; however, these results show large uncertainty windows that encompass positive responses. The ensemble mean shows an ~7.5% inflation of the freshwater reservoir after 20 years.

The time scales of response are also broadly similar across the 12 model simulations. On the whole, response functions show a quasi-exponential form, approaching a new equilibrium after 10 years. Similarly, Johnson et al. (2018) show that in the coupled climate model HiGEM (Shaffrey et al. 2009), Arctic FWC increases by ~8% of the reservoir size, reaching a new state that approximates equilibrium 15 years after a hypothetical step change to a more anticyclonic AO. We save more detailed analysis of response function forms for section 5, in which we select the response functions that we judge to be the most robust representations of the AO–FWC relationship on the basis of the results from both the preindustrial control and historical runs (section 4).

It is important to note that, in the frequency domain, the impulse response function is a transfer function relating the spectral densities of the forcing and response. A transfer function that acts as a filter mapping approximately white noise (SLP) onto red noise (FWC) will have a slow adjustment time scale in the time domain. Even in the absence of a genuine physical connection between the forcing and response time series, their spectral densities will dictate that of the transfer function, because the regression method seeks the best fit. As such, the spectral density/adjustment time scale of the transfer/impulse response function alone is not decisive evidence that the ocean has a decadal memory of a particular mode of atmospheric variability. A more convincing case would incorporate (i) strong theoretical grounds or (ii) statistical approaches that robustly reproduce such a relationship. We note the existence of the former for certain atmospheric modes (e.g., Davis et al. 2014; Manucharyan and Spall 2016), and test the second condition in these models here using Monte Carlo analyses.

In Fig. 6 we show the FWC reconstruction skill across models using the linear response functions (crosses), alongside the significance levels [bars; see section 3b(1) for method]. There are major differences across models in the extent to which these linear response functions can explain the FWC variability, with the lowest PC1 reconstruction R values around 0.2 and the highest around 0.6. We find that only four model simulations show AO–FWC relationships that are sufficiently strong to exceed the significance levels: CanESM2, CCSM4, GFDL CM3, and IPSL-CM5A-MR (see blue bars and crosses in Fig. 6). Note that the addition of the second and third EOFs must, by construction, increase the variance explained [Eq. (2); see pink bars and crosses in Fig. 6]. For some

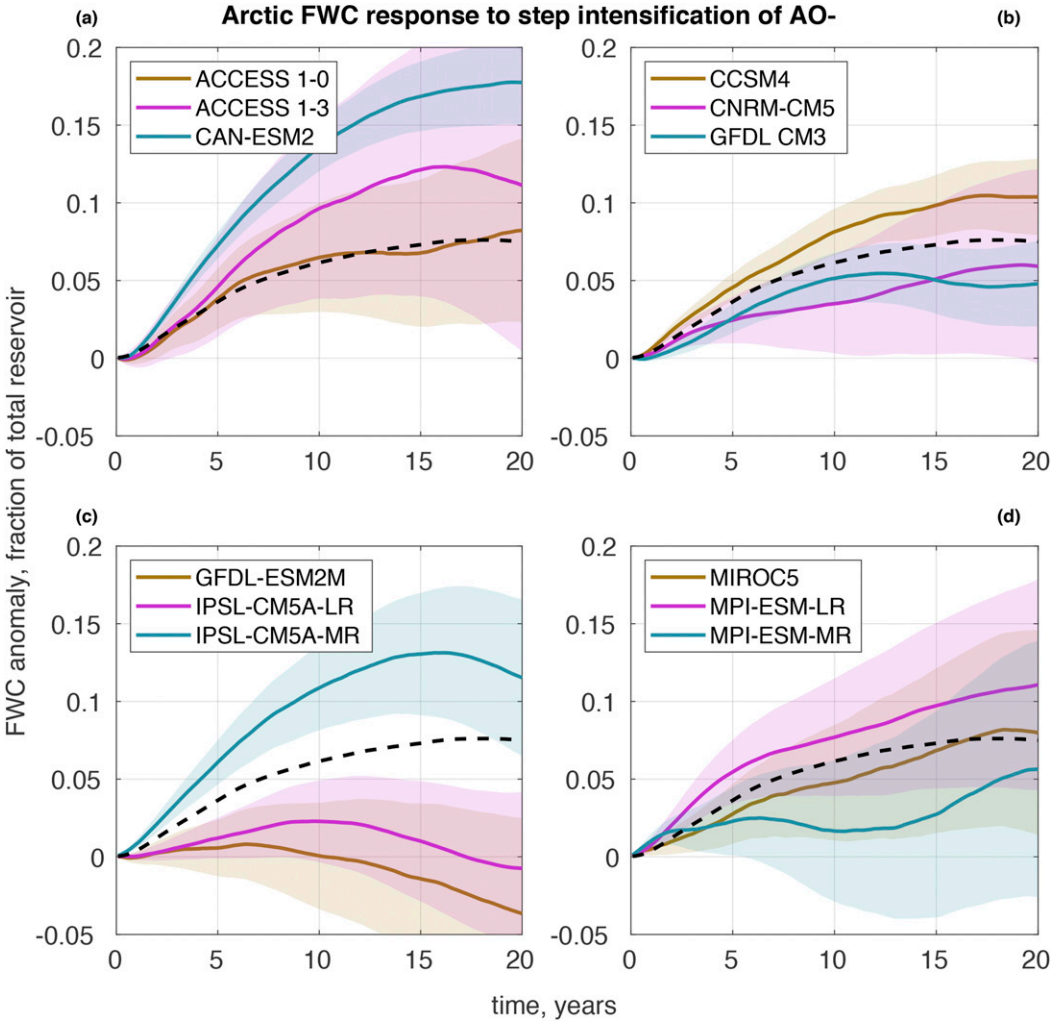

FIG. 5. Response of Arctic FWC to a step change to a more anticyclonic (more negative) AO in 12 CMIP5 preindustrial control runs. Shaded areas show the 1 std dev uncertainty, calculated using the Monte Carlo method as in section 3b(2). Response functions by model are divided into (a)–(d) for legibility, but the black dotted curve is the ensemble mean step response across all 12 models.

simulations (CNRM, IPSL-CM5A-LR, MIROC5) the combination of the first three PCs explains a statistically significant degree of variance, where the first PC alone does not. This implies that there is a physical relationship between SLP as expressed through the combination of PCs 1–3 and FWC in these simulations, but that the AO–FWC relationship is weak, indeed the estimated AO–FWC response function might not represent a physical relationship in the model. Note that the opposite occurs when the second and third modes are included for GFDL CM3; the combined response functions are not significant at the 95% level, indicating that in this model simulation, the second and third PCs exert limited control over the evolution of FWC.

There are a number of possible physical reasons why the models might show apparently weak relationships between FWC and SLP. These reasons include nonlinearity of the relationship; nonstationarity (for instance, the relationship might be modulated by a changing ice regime); a dominance of other forcing mechanisms, whether far field or local; and compensation from competing mechanisms of FWC change. Different sea ice representations across the models (the sea ice model used in each case is listed in Table 1) have the potential to modulate the nature and strength of the SLP–FWC relationship. We review evidence for this from the real world and evaluate model differences in sea ice representations in appendix B. We find, however, that there is no conclusive link across models between the

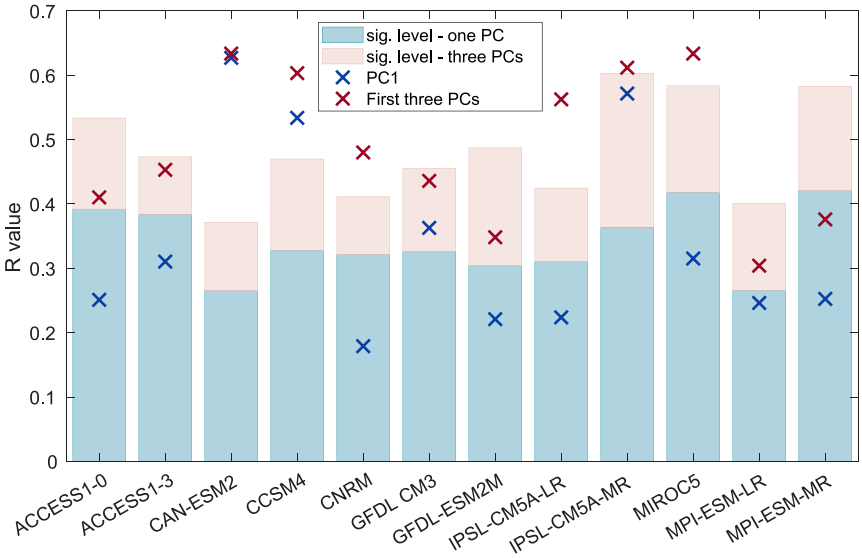

FIG. 6. Reconstruction skill of response functions for the preindustrial control run. Blue and red bars indicate the significance levels for PC1 and all three PCs, respectively, derived using a Monte Carlo technique. Blue crosses indicate the *R* value from the correlation of reconstructed FWC using PC1 and the original control run FWC time series, after removal of the reconstructions using PC2 and PC3. Red crosses give the *R* value for the correlation of the original FWC time series and the reconstruction using all three PCs.

different sea ice representations and the derived SLP–FWC relationships.

The FWC variance explained in each model is not evenly shared between the modes of forcing. For example, in MIROC5, the response to PC2 dominates the FWC variance explained by SLP in this model. By contrast, in CanESM2, the AO is dominant in explaining FWC variability. These differences may be related to the different FWC distributions between models, onto which the EOFs of SLP will project differently, hence producing a range of responses.

While it is beyond the scope of this paper to determine the physical mechanisms by which freshwater adjustments in response to the AO occur, we note that dedicated modeling and observational studies provide insight as to how the AO− might lead to Arctic FWC accumulation, and we suspect similar mechanisms to be at play in the CMIP5 models. Lique et al. (2010) showed using a Lagrangian analysis applied to a high-resolution ocean ice model that the Beaufort Gyre expands and intensifies under AO− forcing relative to AO+, while the export of freshwater through the Fram Strait decreases. The latter result is mirrored in the Zhang et al. (2003) forced model study. Observational studies also report expansion of the Beaufort Gyre under AO− forcing, and a more Eurasian position of the Transpolar Drift (Rigor et al. 2002; Steele et al. 2004; Kwok et al. 2013). These changes are largely connected to the impact on Ekman pumping within the Arctic (e.g., Ma et al. 2017), which influences the surface geostrophic circulation. Armitage et al. (2018) show with sea surface height data that the AO forcing drives cross-shelf Ekman transport, which excites along-shelf flow. Armitage et al. (2018) find that the AO− is associated with enhanced Bering Strait inflow (of relatively fresh waters) and decreased Atlantic inflow through the Barents Sea opening.

Different physical mechanisms have different time scales of adjustment that may be detectable in the form of the response functions. Thermodynamic sea ice changes should adjust rapidly to a change in the atmospheric circulation (Wernli and Papritz 2018). Sea ice growth perturbations can remain important on longer time scales, however, if changes to the sea ice export result in sustained anomalies to the ice-free area, leading to sustained sea ice growth anomalies. Mixing across the halocline is low (e.g., Zhang and Steele 2007), in the main due to strong stratification rather than sea ice cover (Guthrie et al. 2013; Lincoln et al. 2016). However, vertical mixing might alter with SLP-driven changes to the stratification, especially in shelf regions, and with changes in eddy activity. Liquid freshwater fluxes through the Arctic gateways adjust to SLP anomalies via both barotropic and baroclinic flow. Sea surface height gradients should respond quickly to SLP anomalies by Ekman pumping. In the interior of

the Arctic Ocean, the tilts of isopycnal surfaces are expected to adjust more slowly, whether by processes related to planetary Rossby waves (Yang et al. 2016) or by eddy activity (e.g., Manucharyan and Spall 2016). In the Beaufort Gyre, the time scale of adjustment is affected by a combination of eddy diffusivity and seasonally occurring drag on the geostrophic circulation by overlying sea ice, yielding a multiyear time scale that is shorter than that due to eddy diffusivity alone (Doddridge et al. 2019).

4. CMIP5 historical runs

We attempt to reconstruct Arctic FWC in each model's historical run (spanning 1850–2005) using the linear response functions described above, with several objectives in mind. First, we seek an independent test of whether a given response function is physical. Our hypothesis is that, providing the nature of the SLP–FWC relationship is largely unchanged between the preindustrial control run and the historical run of a given model, response functions trained on the control run should be capable of reproducing some of the FWC variability in the historical run, but only if they capture physical relationships in the control run. Because the historical runs are relatively short, response functions trained directly on the historical runs would be particularly vulnerable to overfitting.

Our second objective is to probe the strength of the SLP–FWC relationship, independent of the impulse response functions that were derived from the preindustrial control. We use a Monte Carlo approach with synthetic response functions to estimate the upper bounds on reconstruction skill in the historical runs (see section a of appendix A).

a. Reconstruction

First, we again define our time series of FWC and PCs of SLP in the same way as in section 2 but based on the historical-run data. We detrend the FWC time series as before, but do not attempt to decipher and remove any influence of radiative forcing a priori. We regress the SLP variability of the historical run onto the first three EOFs of SLP from the control run, then attempt to reconstruct the historical FWC time series by convolving the impulse response functions derived from the control run G_i^{PIC} with the SLP forcing from the historical run $\mathrm{PC}_i^{\mathrm{HIST}}$:

$$\mathrm{FWC}^{\mathrm{HIST}}(t) = \sum_{i=1}^{n} \int_{\tau=0}^{\tau_{\max}} G_i^{\mathrm{PIC}}(\tau)\mathrm{PC}_i^{\mathrm{HIST}}(t-\tau)\, d\tau. \quad (4)$$

We then calculate the correlation coefficient between this estimate of the historical FWC and the original historical FWC time series, which gives us the reconstruction skill.

b. Results

In Fig. 7, we show that convolution of historical-run forcing with impulse responses estimated from the preindustrial control run yields a wide range of FWC reconstruction skill across the 12 models. In some cases, all reconstruction skill disappears, and/or the correlation with the original FWC time series becomes negative. In these cases, we must assume either that (i) the original response functions were not representative of a physical relationship, or (ii) the relationship between SLP and FWC may differ between the control and historical runs.

Comparison of the R value from our reconstructions (crosses in Fig. 7) with the upper bounds on reconstruction skill derived using random response functions (squares in Fig. 7; see section a of appendix A for the method) indicates whether the impulse response functions derived in the control are also the most skillful possible linear description of the SLP–FWC relationship in the historical simulations. If the reconstruction skill is similar to, or higher than, the Monte Carlo estimate of the upper bound, then the response functions estimated from the control run are also appropriate at relating SLP and FWC variability in the historical run. Both of these metrics link the model's original SLP to its original FWC variability. By contrast, the blue bars in Fig. 7 indicate the 95th-percentile level of R values that can be achieved with random forcing and new impulse response functions trained on the historical run with this random forcing [the method is the same as in section 3b(1), but with the historical-run data]. Note that because the historical runs are short, these new response functions can be easily overfitted and will therefore produce a high bar for the significance levels. We suggest that they should be seen as a guide rather than absolute markers for significance. Comparison of these significance levels (blue bars) with the previous two metrics (squares and crosses) helps to establish how likely it is that the forcing time series itself can causally explain FWC variability in the historical run.

To aid with the interpretation of this combination of metrics, we provide two examples. The linear response functions for GFDL-ESM2M lose all reconstructive skill when convolved with the historical SLP variability; the R values are low or even negative (crosses in Fig. 7). Clearly, if a relationship between SLP and FWC exists in this run, it is not well described by these linear response functions. To test whether a relationship between SLP and FWC does exist, we next look to the Monte Carlo estimate of the upper bound on

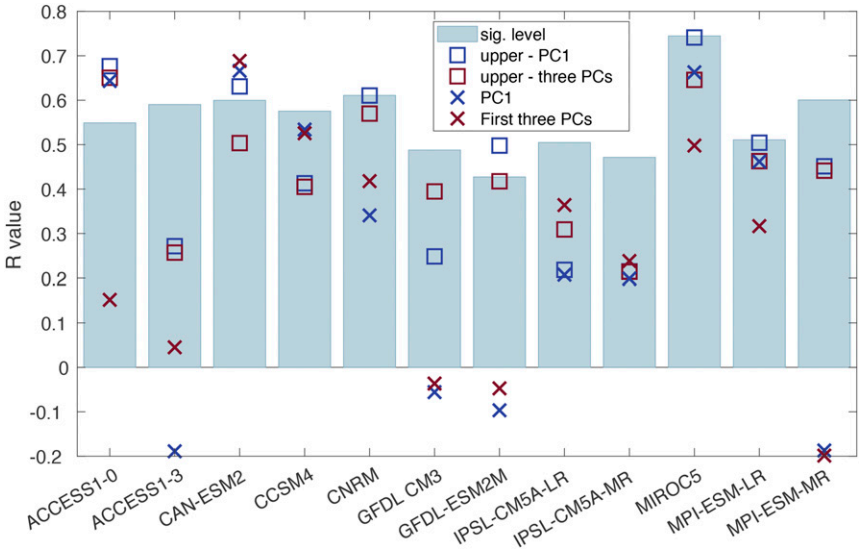

FIG. 7. Monte Carlo evaluation for the historical runs. Blue bars indicate the significance levels for a single forcing component. Crosses show the reconstruction skill using linear response functions derived from the preindustrial control run, calculated after removal of the FWC contributions from PC2 and PC3. Squares represent the 95th-percentile reconstruction skill using a Monte Carlo ensemble of synthetic response functions—an *upper* estimate of possible reconstruction skill with linear responses.

reconstruction skill (square) and compare this against the significance level (bar). The fact that the estimated upper bound on reconstruction skill for PC1 is greater than the significance level indicates that there may be a physical relationship between PC1 and FWC in the historical run, which explains a notable portion of the FWC variance ($R = 0.5$), but that relationship is different to the one obtained in the preindustrial control run.

The reconstruction skill in the CanESM2 historical run is high when using the response functions derived from the preindustrial control run (crosses, Fig. 7). Indeed, comparison with the R values from the Monte Carlo ensemble (squares) shows that these relationships are highly skillful in relating SLP and FWC variability, even though they were trained on a different run. This consistency with the significant result in the preindustrial control run convinces us that this is most likely a physical relationship in both runs.

In Table 4, we evaluate the likelihood that response functions represent causal, physical relationships between SLP and FWC variability in their respective models. While we can never have complete certainty of causal, physical relationships using this statistical method, we can establish the confidence that we may have in the results. To assist with this, we introduce a ranking procedure to compare our estimated response functions with an ensemble of synthetic response functions in both control and historical runs (for details see section b of appendix A). If the rank position goes up when including the historical run as an independent test, this increases the likelihood that our estimated response function represents a causal, physical relationship. If the rank goes down, it decreases that likelihood. As a second test, we compute the conditional probability that a synthetic response function that is more skillful than our estimated response function in the control run will also be more skillful in reconstructing historical variability. If the conditional probability is small, we may have more confidence in the robustness of the estimated response functions than if it is large. We indicate qualitatively how likely the control-run response functions are to represent physical model relationships in the last column in Table 4.

Again, to aid with interpretation, we provide two examples. The reconstruction skill for GFDL CM3 in the control run is statistically significant for PC1 and almost so for the combination of three PCs. Application of the same response function to the historical run, however, yields effectively zero reconstruction skill. Indeed, the rank of the response functions for this model drops markedly when including the historical run. Furthermore, the conditional probability of achieving higher reconstruction skill in the historical run with the synthetic response functions that were more skillful in the control run with random forcing is high. As a result, we must conclude either that a statistically significant result in the

TABLE 4. Response function evaluation by model. The *R*-value rank, given as a percentile, is explained in section b of appendix A. The conditional probability is explained in section 4b. Note that it cannot be computed when the estimated response function outperforms all synthetic responses in the control run. In the final column, check marks indicate that we can have confidence that a response function is a consistent representation of model physics rather than a statistical artifact, X is used when we suspect that response functions are a statistical artifact, and question marks indicate uncertainty in these judgements.

Model simulation		*R*-value rank as %		Conditional probability: random skill > HIST reconstruction	Physical?
		PIC	Change: PIC + HIST		
ACCESS1.0	PC1	55.1	+32.9	0.15	✓?
	All PCs	64.2	−1.6	0.41	X?
ACCESS1.3	PC1	84.1	−49.4	0.78	X?
	All PCs	92.0	−7.6	0.4	X?
CanESM2	PC1	100	0	-	✓
	All PCs	100	0	-	✓
CCSM4	PC1	100	0	-	✓
	All PCs	99.9	+0.1	0	✓
CNRM-CM5	PC1	64.8	+4.6	0.44	✓?
	All PCs	99.1	−3.4	0.11	✓
GFDL CM3	PC1	96.0	−25.6	0.65	?
	All PCs	91.3	−29.9	0.54	?
GFDL-ESM2M	PC1	67.8	−18.6	0.48	X
	All PCs	51.2	−2.6	0.51	X
IPSL-CM5A-LR	PC1	73.0	+18.3	0.12	✓?
	All PCs	100	0	-	✓
IPSL-CM5A-MR	PC1	99.6	+0.3	0	✓
	All PCs	95.4	+4	0.04	✓
MIROC5	PC1	68.0	+19.8	0.23	✓?
	All PCs	98.7	−3.3	0.38	✓
MPI-ESM-LR	PC1	90.1	+4.6	0.22	✓
	All PCs	67.7	+18.5	0.23	✓?
MPI-ESM-MR	PC1	38.5	−4.5	0.57	X
	All PCs	21.6	+1.7	0.67	X

control run was achieved by chance, not as a result of capturing a physical relationship, or that the nature of the SLP–FWC relationship is simply different in the historical run. It is not clear which is the case, so we indicate a question mark for this model in Table 4, but do not include the statistically significant AO–FWC result in the analysis carried out in the following section. By contrast, the rank of the response functions for model MPI-ESM-LR improves when including the historical run, and the conditional probability of being outperformed twice by the same random functions is moderately low. The historical-run results provide more confidence in the robustness of the MPI-ESM-LR control-run response functions, and so we include these results in the next section.

5. Most robust AO–FWC relationships

After statistical testing in both the control-run context and through the application of the historical run as an independent test, we select the models with AO–FWC response functions that can most confidently be judged to represent causal, physical relationships in both the preindustrial control and historical runs. These models are CanESM2, CCSM4, IPSL-CM5A-LR, and MPI-ESM-LR, and we show their responses to a step change to a more negative AO in Fig. 8. These functions bear certain similarities. They are all positive, with adjustment magnitudes greater than the previous ensemble mean (gray curve). They all have similar magnitudes relative to the whole ensemble and they show a broadly similar, quasi-exponential form. After a relatively slow first-year adjustment, the FWC reservoir grows rapidly within the first 10 years, approaching a new equilibrium between 15 and 20 years. Overinterpretation of the precise shapes of these response functions should be avoided—note how the error bars would accommodate a range of shapes. The ensemble mean of these four functions (black curve, Fig. 8) has an *e*-folding time scale of 7.4 yr and a maximum freshwater change of 12.8% of the total Arctic FWC reservoir.

6. Reconstructing observed FWC variability

Having extracted SLP–FWC relationships in each model, we attempt to evaluate whether these relationships are reflective of real-world dynamics, following

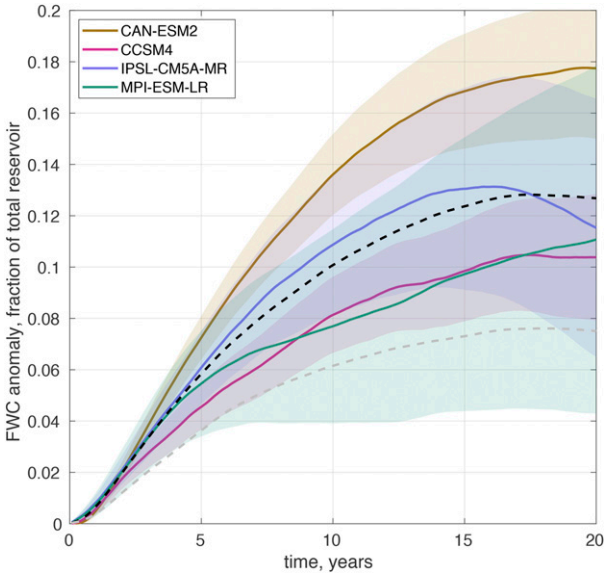

FIG. 8. Selected response functions: response of Arctic FWC to a step change to a more negative AO in preindustrial control runs. Response functions chosen as the most likely to represent physical, causal relationships between SLP and FWC after assessment are described in the text and in Table 4. The black dashed curve shows the ensemble mean of selected response functions; the gray dashed line shows the ensemble mean of all 12 response functions presented in this paper.

Johnson et al. (2018). Our impulse responses can be convolved with real-world atmospheric time series to attribute observed variability. We use ERA-Interim (Berrisford et al. 2011) and ERA-20C (Poli et al. 2016) reanalysis data as the forcing. The higher fidelity ERA-Interim data are available from 1979 onward, so we use the ERA-20C record only up to 1979.

In this analysis, we focus on the AO, because it is consistently represented across the ensemble of models and the reanalysis SLP data (Table 3). We select the four model-based response functions that most likely reflect causal relationships between the AO and FWC in the simulations (Fig. 8). We regress the reanalysis SLP data onto the first three EOFs of SLP for each model then convolve the resulting PC time series with their respective response functions. This yields SLP-driven time series of FWC through the historical period based on linear model-derived relationships. In Fig. 9, we plot reconstructions using PC1 alone in bold and reconstructions using the first three PCs as a dotted line. We show the period 1980–2017 and compare with observations of Arctic FWC from Rabe et al. (2014). Note that we require a full 20 years of forcing before the results reflect the full memory in the response functions, and as such we require reanalysis forcing from 1960 onward to build these reconstructions. To calculate the error bars, we use the ensemble of 1000 impulse response functions described in section 3b(2). Convolving these 1000 response functions with the atmospheric forcing yields 1000 different reconstructions. We plot the standard deviation of these reconstructions as our error bars.

Changes in the AO dominate the FWC variability; including the second and third modes does not significantly change the reconstruction. A sharp downward trend in FWC during approximately 1988–95 is evident in all reconstructions and is the result of a strong and sustained shift toward a positive phase of the AO at the start of that period. All four model-based reconstructions show an increase in Arctic FWC coincident in time with that reported by Rabe et al. (2014) during 1995–2012, due to a sustained shift toward a negative phase of the AO at the start of that period. The magnitude of change is somewhat smaller (~30%–80%) in the reconstructions than the observations, with CanESM2 and IPSL-CM5A-MR coming the closest to reproducing the observed increase. Note that differences in the magnitudes of variability also reflect the different mean states of Arctic FWC in each model (Fig. 1, Table 2).

For comparison, Shu et al. (2018) show that the CMIP5 historical-run multimodel mean captures ~50% of the observed freshening during the 1992–2012 observational period. This signal can be assumed to average across stochastic SLP forcing and represent the effects of longer-term anthropogenic climate change, relating to sea ice melt and an enhanced hydrological cycle, among other possible changes (which could include changes to SLP variability).

7. Concluding discussion

Theory, idealized modeling, and observations indicate that subsystems of Arctic liquid FWC (e.g., the Beaufort Gyre) and solid FWC (e.g., the Transpolar Drift sea ice stream) are controlled, at least in large part, by SLP variability (see the references in the introduction).

In another coupled climate model, HiGEM, Johnson et al. (2018) found that linear responses to the first three modes of SLP forcing could explain FWC variability with a reconstruction skill of $R = 0.93$. None of the CMIP5 models studied here exhibit SLP–FWC relationships that are as strong as that found in HiGEM. The relationship varies considerably in strength from model to model. In many cases the variance explained is insufficiently high for us to be confident that the result is physical. A major caveat of the regression technique is that it cannot be known a priori what portion of the

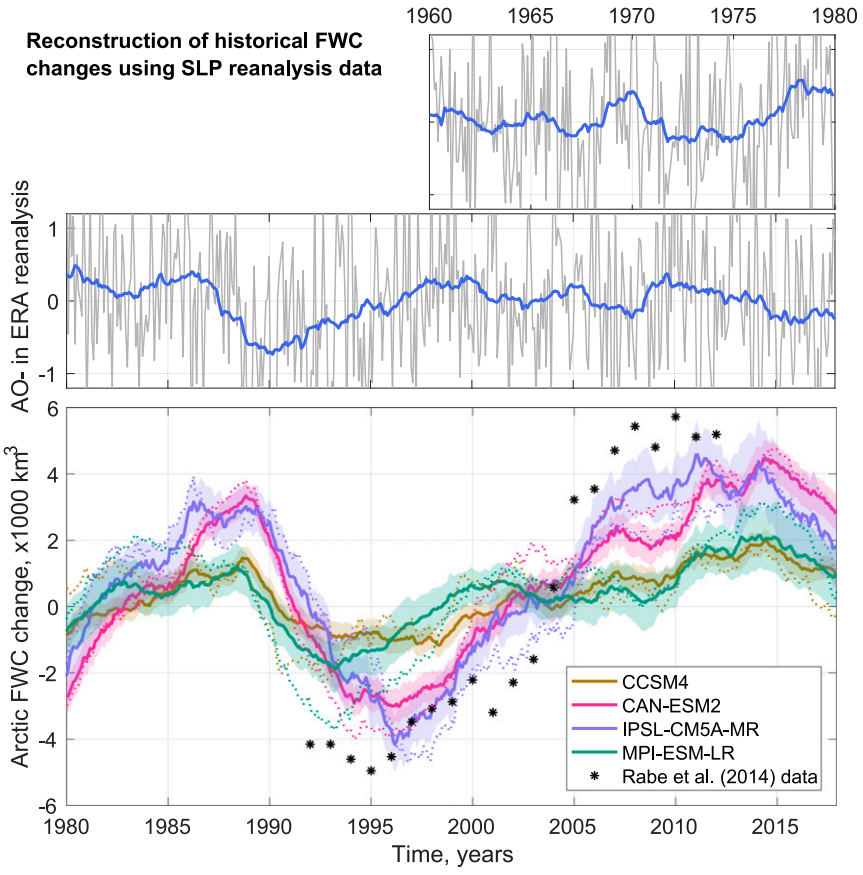

FIG. 9. Reconstruction of historical FWC changes in the real world from 1980, calculated through convolution of SLP reanalysis data from 1960 onward with response functions estimated from preindustrial control runs of each model: the negative AO in ERA-Interim reanalysis SLP data, normalized to 1 std dev in the monthly data for (top) the period 1960–80 and (middle) 1980–2018. Note that positive values indicate a more negative AO. The gray line shows the monthly time series, and the blue line shows a 3-yr moving mean. (bottom) The solid lines denote reconstruction using PC1 alone, and the dotted lines are for all three PCs.

FWC variability should be explainable by a given forcing, so if the selected mode of forcing does not dominate the variability, the response functions will inevitably be trained to best fit residual noise. Error increases as the decorrelation time scale of this residual increases (appendix C). As such, our technique is most effective when there are strong relationships between the variables of interest. Even with random forcing time series, the regression technique will yield response functions that will find some level of reconstruction skill, which may vary widely on a random basis. We outline here a Monte Carlo approach to fairly assess the level of confidence that we therefore can have in whether a regression-derived response function consistently represents the model physics.

The utility of this technique, on the other hand, is evident in several aspects. The technique allows us to cleanly isolate relationships that involve memory in a computationally efficient manner. These relationships can then be easily compared and evaluated in terms of their nature and strength across a range of coupled climate models. Here, we establish a baseline for the Arctic SLP–FWC relationship in CMIP5 models against which the effects of future model development in coupled climate models can be assessed. Further, the technique provides a novel means for model evaluation against observations, through the convolution of estimated response functions with forcing time series from observations or reanalysis data.

In section 6, we perform model evaluation against observations of FWC changes that are thought to have been driven largely by SLP variability (Johnson et al. 2018; Wang et al. 2019). The convolution of SLP reanalysis data and response functions from selected

models yields between 30% and 80% of the observed freshwater accumulation 1992–2012, with similar timing of changes. While coupled climate models may be able to reproduce the amplitude and frequency of internal atmospheric variability (e.g., the AO) we do not expect model SLP to vary with the same phase as in the real atmosphere either in the historical runs or the future, meaning we should exercise caution in interpreting FWC changes exhibited in these runs. As such, our model evaluation approach provides one solution to attributing historical changes, and while we cannot predict SLP variability years in advance, memory in the system may afford a few years of potential predictability for Arctic FWC (Johnson et al. 2018). As CMIP6 data become available, we suggest that this approach has utility in evaluating the representation of CMIP6 internal variability against the real world in cases where memory is important and the forcing process is freely evolving in the model run. Such an approach can also help to distinguish changes arising from forced versus internal variability, crucial in the context of climate change.

The results for the response of Arctic FWC to a step change in the AO that are most likely to represent causal, physical relationships in the model simulations all indicate that one-standard-deviation enhancement of the anticyclonic AO leads to freshwater accumulation, with a magnitude of ~13% of the total Arctic FWC reservoir, and with an *e*-folding time of ~7 years. These results show an appreciable correspondence with recent observations of Arctic FWC changes (section 6), providing some confidence that they are useful representations of real-world dynamics. FWC accumulation under AO− forcing is consistent with modeling and observation results that show an intensified Beaufort Gyre, weakened Fram strait freshwater export, and increased Bering Strait inflow under the AO− (e.g., Lique et al. 2010; Armitage et al. 2018).

The fact that CMIP5 models exhibit such different strengths of SLP–FWC relationships is a compelling indicator that, while the mean states and spectral densities of FWC variability might be broadly similar, the physical origins of this variability are inconsistent between models. This, in turn, calls into question our ability to understand the sources of future Arctic FWC variability in climate change runs of these models and our confidence in their predictions.

Acknowledgments. Author Cornish was funded by a DTP studentship through U.K. Natural Environment Research Council (NERC) Grant NE/L002612/1. Authors Kostov and Johnson are grateful for funding from the U.K.-OSNAP project through NERC Grant NE/K010948/1. Kostov was also funded by the TICTOC project through NERC Grant NE/P019064/1. Author Lique acknowledges funding from the French LEFE/INSU program through the project FREDY. We are grateful for input from Chris Hughes, who encouraged the use of a Monte Carlo approach and provided the script for generating surrogate time series. We are grateful also to Emma Beer, who wrote some scripts during a summer placement that we adapted for use in this project. We planned this paper during a meeting of the Forum for Arctic Modelling and Observational Synthesis (FAMOS). CMIP5 model data are available online (https://esgf-node.llnl.gov/projects/cmip5/). We are grateful to three anonymous reviewers and editor Rong Zhang, whose comments helped to substantially improve the paper.

APPENDIX A

Methods

a. *Monte Carlo estimate of upper bound on reconstruction skill in historical runs*

The Monte Carlo procedure in section 3b(1) yields an ensemble of 1000 synthetic impulse response functions for each model, symmetrically distributed about zero. We use this ensemble here to estimate bounds on the potential reconstruction skill in the historical runs through convolving synthetic response functions and historical-run SLP. The 1000 convolutions yields a histogram of R values, symmetrical about $R = 0$. Because the ensemble of 1000 response functions spans the plausible space of possible impulse responses, some should be skillful by chance. The maximum R value is likely the product of a relationship that is overfitted by chance. The level of reconstruction skill is less sensitive to small changes to the response functions at the 95th percentile level, however. We take the 95th percentile value as the estimated upper bound on reconstruction skill; this also enables an equivalent comparison against the significance level.

b. *Ranking procedure using historical run as independent test*

We rank the performance of the preindustrial-control-run reconstructions described in section 3 against the ensemble of reconstructions that are used to establish the significance levels [section 3b(1)]. The exact same 1000 synthetic impulse response functions used in the significance evaluation are carried forward to the analysis that yields the upper bound on reconstruction skill in the historical run (section a of this appendix). The forcing for these reconstructions is the historical-run SLP variability.

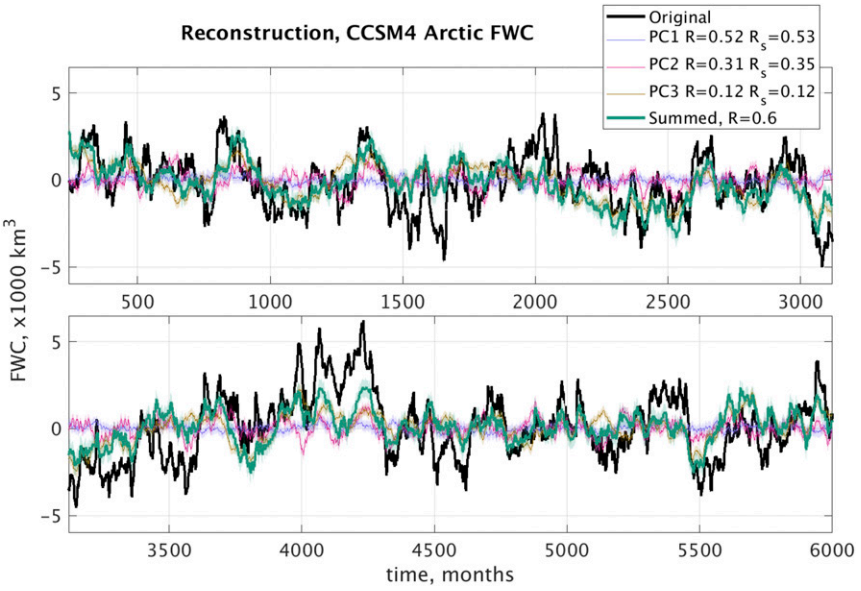

FIG. A1. Preindustrial-control-run reconstruction of Arctic FWC in CCSM4 using SLP and estimated response functions; R denotes the correlation of the reconstructed time series with the original FWC time series, and R_S denotes the correlation after removal of the FWC contributions from the other components.

For each unique synthetic response function, we then sum the FWC reconstruction R value from the control and historical run. Summing the reconstruction R value from our reconstructions of historical and control FWC variability, we again rank our estimated response function within the ensemble of synthetic response functions. In Table 4, we show side by side the control-run rank (>95% corresponds to statistical significance) and the change in rank when results for the control run and historical runs are combined. The combined rank is a qualitative metric; it is the change in rank (whether up or down) that is more informative than the absolute value.

c. *Linear response function method*

Table A1 illustrates the segmentwise approach used to subsample the model runs in the regression process—an attempt to minimize overfitting (section 3a). Figure A1 compares the reconstructed FWC time series and the original FWC time series in the CCSM4 control run. Reconstructions use the linear response functions described in section 3a.

APPENDIX B

Sea Ice Relationships

The momentum transfer from atmosphere to ocean depends on sea ice characteristics in a complex manner (Martin et al. 2014; Tsamados et al. 2014; Petty et al. 2016; Cole et al. 2017), with implications for ocean circulation beneath (Davis et al. 2014; Dewey et al. 2018; Meneghello et al. 2018) and for sea ice drift and export from the Arctic (Haas et al. 2008; Petty et al. 2016). It is therefore important to consider model representations of sea ice when discussing SLP–FWC relationships.

Analysis of Arctic sea ice variability in the historical runs of CMIP5 models reveals a wide range of biases in sea ice area and export (Langehaug et al. 2013), some of which may originate from model tuning aimed at reducing other biases such as hemispheric sea ice area (Ivanova et al. 2016). We show the mean sea ice volume and extent over our Arctic region in Fig. B1. Across the 12 preindustrial simulations considered here, no significant relationship (at a confidence level of 90%) exists between the strength of the SLP–FWC relationship (the reconstruction R value for all three EOFs) and these mean sea ice metrics. Comparison of simulations from the same modeling group also indicates inconsistent effects of different sea ice representations (e.g., compare IPSL models, MPI models).

Next we ask whether the growth and melt of sea ice is important in attributing liquid FWC changes unexplained by SLP variability. Here we analyze the seven model simulations in our selection for which ice-to-ocean freshwater flux is archived. After subtracting the contribution of SLP to the variability in the FWC time

TABLE A1. Explanation of the four distinct segmenting schemes used to subsample the model runs in the regression process. Each is made up of 20 segments, of the length shown, and starting on the month shown. The starting months are separated by the spacing between segments, also shown.

	1	2	3	4
3600-month runs				
Length of each segment (months)	3581	3353	3239	3125
Spacing between segments (months)	1	13	19	25
Starting month, segment no.				
1	1	1	1	1
2	2	14	20	26
3	3	27	39	51
⋮	⋮	⋮	⋮	⋮
20	20	248	362	476
6000-month runs				
Length of each segment (months)	5981	5354	5031	4708
Spacing between segments (months)	1	34	51	68
Starting month, segment no.				
1	1	1	1	1
2	2	35	52	69
3	3	69	103	137
⋮	⋮	⋮	⋮	⋮
20	20	647	970	1293
9600-month runs				
Length of each segment (months)	9581	8365	7738	7111
Spacing between segments (months)	1	65	98	131
Starting month, segment no.				
1	1	1	1	1
2	2	66	99	132
3	3	131	197	263
⋮	⋮	⋮	⋮	⋮
20	20	1236	1863	2490

series, we compute correlations with accumulated ice-to-ocean freshwater fluxes (I2O) over the same domain (Table B1). The expected correlation is positive. I2O is somewhat important in explaining non-SLP-related variability in ACCESS1.0, ACCESS1.3, and MIROC5. Weak, and negative, relationships exist between I2O and FWC in the GFDL and IPSL models. I2O variability is particularly muted in the IPSL models relative to FWC variability (not shown).

APPENDIX C

Spectral Properties and Sources of Error

There is a positive correlation between the decorrelation time scale of the FWC time series and the error on the response function that we calculate, and a negative correlation between error and the variance explained by the FWC reconstructions. Table C1 shows the variance in the error explained by key regressors.

To determine the origin of the error associated with the decorrelation time scale, we perform synthetic tests.

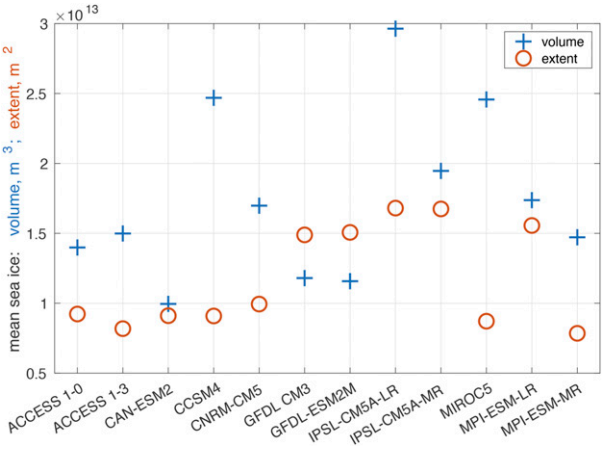

FIG. B1. Mean sea ice volume and extent in preindustrial control runs of 12 CMIP5 model simulations.

We construct predictable FWC time series using a known impulse function and known (white noise) forcing time series. We can then mix these predictable signals with residual noise, as per section 3b(2). In this method, however, we construct residual noise time series as first-order autoregressive [AR(1)] processes using convolution with an exponential impulse response function, $G(\tau) = e^{-\tau/T}$. This yields AR(1) time series with an autocorrelation function, $r(\tau) = a^{\tau}$ that has an e-folding decorrelation time scale $T = -1/\ln a$.

We then undertake the regression method to attempt to find the original response functions. We span a reasonable parameter space and find the errors as the standard deviation of the 500 response function estimates for each choice of parameters. Error increases with both the residual noise decorrelation time scale and the inverse of variance explained (Fig. C1).

An additional source of error associated with the spectral properties of the FWC time series arises from the use of a relatively short response function cutoff

TABLE B1. Correlation of cumulative ice-to-ocean freshwater fluxes (I2O) with the residual of Arctic FWC after subtraction of the reconstruction using SLP and response functions. I2O is accumulated from the beginning of the run. Only model simulations in which the I2O are archived are shown. All values are for the preindustrial control runs.

Model simulation	R value: I2O and FWC residual
ACCESS1.0	0.66
ACCESS1.3	0.49
GFDL CM3	−0.09
GFDL-ESM2M	0.27
IPSLA-CM5A-LR	−0.14
IPSLA-CM5A-MR	−0.28
MIROC5	0.43

TABLE C1. Percent variance explained in the response function error across models by different regressors. *Rescaled* refers to rescaling by the length of the time series. *Inverse variance* refers to the inverse of the variance explained by the response function reconstruction (section 3c).

Regressor(s)	Intermodel response function error % variance explained
FWC decorrelation time scale	42
Rescaled FWC decorrelation time scale	51
Rescaled FWC decorrelation time scale and inverse variance explained	76
Rescaled FWC residual decorrelation time scale and inverse variance explained	77

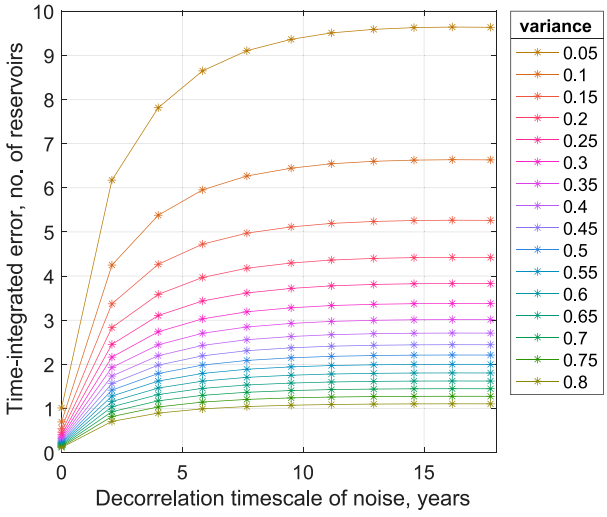

FIG. C1. Response function error associated with residual noise—characterized by different spectral properties—in the target FWC time series. Error is integrated through time. The legend heading *variance* refers to the variance explained by the predictable part of the signal. As this variance decreases, the residual noise becomes increasingly dominant. The PC1 response function for ACCESS1.0 was used as the initial response function for these synthetic tests.

lag. As the decorrelation time scale of the FWC time series approaches and exceeds the cutoff lag time scale, error grows. For a 20-yr cutoff lag, the error follows a roughly linear trend over the range of time scales we consider, from close to zero additional error at a decorrelation time scale of 5 years, to 0.7 reservoir fractions of integrated response function error at 20 years (not shown).

Given that (i) red residual noise increases the error in the determination of impulse response functions, (ii) the residual noise generally has similar spectral properties to the original FWC time series, and (iii) a (relatively small) error grows as the decorrelation time scale of FWC approaches and exceeds the response function cutoff lag, it might be considered desirable to high-pass filter the data. It is conceivable that other sources of variability possibly unrelated to Arctic SLP add noise on longer time scales and contribute to error. However, we choose not to include analysis of high-pass filtered data here for three main reasons. First, there is not an obvious physically motivated time scale to choose as the filter threshold. Second, the SLP forcing has roughly equal power at all frequencies; via an impulse response function, it should be able to produce long-time-scale variability. Third, high-pass filtering a red time series, for which power increases with decreasing frequency, creates a new peak in the power spectrum at frequencies immediately above the threshold, effectively forcing the time series to become periodic on these time scales.

Repeating the analysis in section 3 with a 40-yr cutoff high-pass filter leads to step responses that are qualitatively similar to the originals, though with shorter adjustment time scales and a pronounced oscillatory component (not shown). The number of significant results for the AO response functions increases to 10 of 12 models (not shown). This lends credence to the robustness of the results, but these response functions portray a distorted SLP–FWC relationship, so we do not include them here.

REFERENCES

Aagaard, K., and E. C. Carmack, 1989: The role of sea ice and other fresh water in the Arctic circulation. *J. Geophys. Res.*, **94**, 14 485–14 498, https://doi.org/10.1029/JC094iC10p14485.

Alkire, M. B., K. K. Falkner, I. Rigor, M. Steele, and J. Morison, 2007: The return of Pacific Waters to the upper layers of the central Arctic Ocean. *Deep-Sea Res. I*, **54**, 1509–1529, https://doi.org/10.1016/j.dsr.2007.06.004.

Armitage, T. W., S. Bacon, and R. Kwok, 2018: Arctic sea level and surface circulation response to the Arctic Oscillation. *Geophys. Res. Lett.*, **45**, 6576–6584, https://doi.org/10.1029/2018GL078386.

Berrisford, P., and Coauthors, 2011: The ERA-Interim Archive: Version 2.0. ERA Rep. Series 1, 23 pp., https://www.ecmwf.int/sites/default/files/elibrary/2011/8174-era-interim-archive-version-20.pdf.

Cai, L., V. A. Alexeev, J. E. Walsh, and U. S. Bhatt, 2018: Patterns, impacts, and future projections of summer variability in the Arctic from CMIP5 models. *J. Climate*, **31**, 9815–9833, https://doi.org/10.1175/JCLI-D-18-0119.1.

Carmack, E. C., 2007: The alpha/beta ocean distinction: A perspective on freshwater fluxes, convection, nutrients and productivity in high-latitude seas. *Deep-Sea Res. II*, **54**, 2578–2598, https://doi.org/10.1016/j.dsr2.2007.08.018.

——, and Coauthors, 2016: Freshwater and its role in the Arctic marine system: Sources, disposition, storage, export, and physical and biogeochemical consequences in the Arctic and

global oceans. *J. Geophys. Res. Biogeosci.*, **121**, 675–717, https://doi.org/10.1002/2015JG003140.

Chylek, P., J. Li, M. K. Dubey, M. Wang, and G. Lesins, 2011: Observed and model simulated 20th century Arctic temperature variability: Canadian Earth System Model CanESM2. *Atmos. Chem. Phys. Discuss.*, **11**, 22 893–22 907, https://doi.org/10.5194/acpd-11-22893-2011.

Cole, S. T., and Coauthors, 2017: Ice and ocean velocity in the Arctic marginal ice zone: Ice roughness and momentum transfer. *Elementa Sci. Anthropocene*, **5**, 55, https://doi.org/10.1525/elementa.241.

Davis, P. E., C. Lique, and H. L. Johnson, 2014: On the link between Arctic sea ice decline and the freshwater content of the Beaufort Gyre: Insights from a simple process model. *J. Climate*, **27**, 8170–8184, https://doi.org/10.1175/JCLI-D-14-00090.1.

Deser, C., 2000: On the teleconnectivity of the ''Arctic Oscillation.'' *Geophys. Res. Lett.*, **27**, 779–782, https://doi.org/10.1029/1999GL010945.

Dewey, S., J. Morison, R. Kwok, S. Dickinson, D. Morison, and R. Andersen, 2018: Arctic ice-ocean coupling and gyre equilibration observed with remote sensing. *Geophys. Res. Lett.*, **45**, 1499–1508, https://doi.org/10.1002/2017GL076229.

Ding, Y., J. A. Carton, G. A. Chepurin, M. Steele, and S. Hakkinen, 2016: Seasonal heat and freshwater cycles in the Arctic Ocean in CMIP5 coupled models. *J. Geophys. Res. Oceans*, **121**, 2043–2057, https://doi.org/10.1002/2015JC011124.

Dix, M., and Coauthors, 2013: The ACCESS coupled model: Documentation of core CMIP5 simulations and initial results. *Aust. Meteor. Oceanogr. J.*, **63**, 83–99, https://doi.org/10.22499/2.6301.006.

Doddridge, E. W., G. Meneghello, J. Marshall, J. Scott, and C. Lique, 2019: A three-way balance in the Beaufort Gyre: The ice-ocean governor, wind stress, and eddy diffusivity. *J. Geophys. Res. Oceans*, **124**, 3107–3124, https://doi.org/10.1029/2018JC014897.

Dufresne, J.-L., and Coauthors, 2013: Climate change projections using the IPSL-CM5 Earth System Model: From CMIP3 to CMIP5. *Climate Dyn.*, **40**, 2123–2165, https://doi.org/10.1007/s00382-012-1636-1.

Dunne, J. P., and Coauthors, 2012: GFDL's ESM2 global coupled climate–carbon Earth System Models. Part I: Physical formulation and baseline simulation characteristics. *J. Climate*, **25**, 6646–6665, https://doi.org/10.1175/JCLI-D-11-00560.1.

Gent, P. R., and Coauthors, 2011: The Community Climate System Model version 4. *J. Climate*, **24**, 4973–4991, https://doi.org/10.1175/2011JCLI4083.1.

Giorgetta, M. A., and Coauthors, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *J. Adv. Model. Earth Syst.*, **5**, 572–597, https://doi.org/10.1002/jame.20038.

Griffies, S. M., and Coauthors, 2011: The GFDL CM3 coupled climate model: Characteristics of the ocean and sea ice simulations. *J. Climate*, **24**, 3520–3544, https://doi.org/10.1175/2011JCLI3964.1.

Guthrie, J. D., J. H. Morison, and I. Fer, 2013: Revisiting internal waves and mixing in the Arctic Ocean. *J. Geophys. Res. Oceans*, **118**, 3966–3977, https://doi.org/10.1002/jgrc.20294.

Haas, C., A. Pfaffling, S. Hendricks, L. Rabenstein, J. L. Etienne, and I. Rigor, 2008: Reduced ice thickness in Arctic Transpolar Drift favors rapid ice retreat. *Geophys. Res. Lett.*, **35**, L17501, https://doi.org/10.1029/2008GL034457.

Haine, T. W., and Coauthors, 2015: Arctic freshwater export: Status, mechanisms, and prospects. *Global Planet. Change*, **125**, 13–35, https://doi.org/10.1016/J.GLOPLACHA.2014.11.013.

Hasselmann, K., R. Sausen, E. Maier-Reimer, and R. Voss, 1993: On the cold start problem in transient simulations with coupled atmosphere-ocean models. *Climate Dyn.*, **9**, 53–61, https://doi.org/10.1007/BF00210008.

Holland, M. M., J. Finnis, A. P. Barrett, and M. C. Serreze, 2007: Projected changes in Arctic Ocean freshwater budgets. *J. Geophys. Res.*, **112**, G04S55, https://doi.org/10.1029/2006JG000354.

Ilicak, M., and Coauthors, 2016: An assessment of the Arctic Ocean in a suite of interannual CORE-II simulations. Part III: Hydrography and fluxes. *Ocean Modell.*, **100**, 141–161, https://doi.org/10.1016/j.ocemod.2016.02.004.

Ivanova, D. P., P. J. Gleckler, K. E. Taylor, P. J. Durack, and K. D. Marvel, 2016: Moving beyond the total sea ice extent in gauging model biases. *J. Climate*, **29**, 8965–8987, https://doi.org/10.1175/JCLI-D-16-0026.1.

Jahn, A., and M. M. Holland, 2013: Implications of Arctic sea ice changes for North Atlantic deep convection and the meridional overturning circulation in CCSM4-CMIP5 simulations. *Geophys. Res. Lett.*, **40**, 1206–1211, https://doi.org/10.1002/grl.50183.

Johnson, H. L., S. B. Cornish, Y. Kostov, E. Beer, and C. Lique, 2018: Arctic Ocean freshwater content and its decadal memory of sea-level pressure. *Geophys. Res. Lett.*, **45**, 4991–5001, https://doi.org/10.1029/2017GL076870.

Kostov, Y., J. Marshall, U. Hausmann, K. C. Armour, D. Ferreira, and M. M. Holland, 2017: Fast and slow responses of Southern Ocean sea surface temperature to SAM in coupled climate models. *Climate Dyn.*, **48**, 1595–1609, https://doi.org/10.1007/s00382-016-3162-z.

——, D. Ferreira, K. C. Armour, and J. Marshall, 2018: Contributions of greenhouse gas forcing and the southern annular mode to historical Southern Ocean surface temperature trends. *Geophys. Res. Lett.*, **45**, 1086–1097, https://doi.org/10.1002/2017GL074964.

Kwok, R., 2009: Outflow of Arctic Ocean sea ice into the Greenland and Barents Seas: 1979–2007. *J. Climate*, **22**, 2438–2457, https://doi.org/10.1175/2008JCLI2819.1.

——, G. Spreen, and S. Pang, 2013: Arctic sea ice circulation and drift speed: Decadal trends and ocean currents. *J. Geophys. Res. Oceans*, **118**, 2408–2425, https://doi.org/10.1002/jgrc.20191.

Langehaug, H. R., F. Geyer, L. H. Smedsrud, and Y. Gao, 2013: Arctic sea ice decline and ice export in the CMIP5 historical simulations. *Ocean Modell.*, **71**, 114–126, https:/doi.org/10.1016/j.ocemod.2012.12.006.

Lincoln, B. J., T. P. Rippeth, Y.-D. Lenn, M. L. Timmermans, W. J. Williams, and S. Bacon, 2016: Wind-driven mixing at intermediate depths in an ice-free Arctic Ocean. *Geophys. Res. Lett.*, **43**, 9749–9756, https://doi.org/10.1002/2016GL070454.

Lique, C., A. M. Treguier, B. Blanke, and N. Grima, 2010: On the origins of water masses exported along both sides of Greenland: A Lagrangian model analysis. *J. Geophys. Res.*, **115**, C05019, https://doi.org/10.1029/2009JC005316.

——, M. M. Holland, Y. B. Dibike, D. M. Lawrence, and J. A. Screen, 2016: Modeling the Arctic freshwater system and its integration in the global system: Lessons learned and future challenges. *J. Geophys. Res. Biogeosci.*, **121**, 540–566, https://doi.org/10.1002/2015JG003120.

Lund, I. A., 1970: A Monte Carlo method for testing the statistical significance of a regression equation. *J. Appl. Meteor.*, **9**,

330–332, https://doi.org/10.1175/1520-0450(1970)009<0330:AMCMFT>2.0.CO;2.

Ma, B., M. Steele, and C. M. Lee, 2017: Ekman circulation in the Arctic Ocean: Beyond the Beaufort Gyre. *J. Geophys. Res. Oceans*, **122**, 3358–3374, https://doi.org/10.1002/2016JC012624.

Manucharyan, G. E., and M. A. Spall, 2016: Wind-driven freshwater buildup and release in the Beaufort Gyre constrained by mesoscale eddies. *Geophys. Res. Lett.*, **43**, 273–282, https://doi.org/10.1002/2015GL065957.

——, and P. E. Isachsen, 2019: Critical role of continental slopes in halocline and eddy dynamics of the Ekman-driven Beaufort Gyre. *J. Geophys. Res. Oceans*, **124**, 2679–2696, https://doi.org/10.1029/2018JC014624.

Marshall, J., J. Scott, and A. Proshutinsky, 2017: Climate response functions for the Arctic Ocean: A proposed coordinated modelling experiment. *Geosci. Model Dev.*, **10**, 2833–2848, https://doi.org/10.5194/gmd-10-2833-2017.

Martin, T., M. Steele, and J. Zhang, 2014: Seasonality and long-term trend of Arctic Ocean surface stress in a model. *J. Geophys. Res. Oceans*, **119**, 1723–1738, https://doi.org/10.1002/2013JC009425.

Meneghello, G., J. Marshall, M.-L. Timmermans, and J. Scott, 2018: Observations of seasonal upwelling and downwelling in the Beaufort Sea mediated by sea ice. *J. Phys. Oceanogr.*, **48**, 795–805, https://doi.org/10.1175/JPO-D-17-0188.1.

Morison, J., M. Steele, T. Kikuchi, K. Falkner, and W. Smethie, 2006: Relaxation of central Arctic Ocean hydrography to pre-1990s climatology. *Geophys. Res. Lett.*, **33**, L17604, https://doi.org/10.1029/2006GL026826.

——, R. Kwok, C. Peralta-Ferriz, M. Alkire, I. Rigor, R. Andersen, and M. Steele, 2012: Changing Arctic Ocean freshwater pathways. *Nature*, **481**, 66–70, https://doi.org/10.1038/nature10705.

Muilwijk, M., L. H. Smedsrud, M. Ilicak, and H. Drange, 2018: Atlantic Water heat transport variability in the 20th century Arctic Ocean from a global ocean model and observations. *J. Geophys. Res. Oceans*, **123**, 8159–8179, https://doi.org/10.1029/2018JC014327.

——, and Coauthors, 2019: Arctic Ocean response to Greenland Sea wind anomalies in a suite of model simulations. *J. Geophys. Res. Oceans*, **124**, 6286–6322, https://doi.org/10.1029/2019JC015101.

Newton, R., P. Schlosser, D. G. Martinson, and W. Maslowski, 2008: Freshwater distribution in the Arctic Ocean: Simulation with a high-resolution model and model-data comparison. *J. Geophys. Res.*, **113**, C05024, https://doi.org/10.1029/2007JC004111.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110**, 699–706, https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2.

Petty, A. A., J. K. Hutchings, J. A. Richter-Menge, and M. A. Tschudi, 2016: Sea ice circulation around the Beaufort Gyre: The changing role of wind forcing and the sea ice state. *J. Geophys. Res. Oceans*, **121**, 3278–3296, https://doi.org/10.1002/2015JC010903.

Pickart, R. S., and M. A. Spall, 2007: Impact of Labrador Sea convection on the North Atlantic meridional overturning circulation. *J. Phys. Oceanogr.*, **37**, 2207–2227, https://doi.org/10.1175/JPO3178.1.

Poli, P., and Coauthors, 2016: ERA-20C: An atmospheric reanalysis of the twentieth century. *J. Climate*, **29**, 4083–4097, https://doi.org/10.1175/JCLI-D-15-0556.1.

Polyakov, I. V., and Coauthors, 2008: Arctic Ocean freshwater changes over the past 100 years and their causes. *J. Climate*, **21**, 364–384, https://doi.org/10.1175/2007JCLI1748.1.

——, U. S. Bhatt, J. E. Walsh, E. P. Abrahamsen, A. V. Pnyushkov, and P. F. Wassmann, 2013: Recent oceanic changes in the Arctic in the context of long-term observations. *Ecol. Appl.*, **23**, 1745–1764, https://doi.org/10.1890/11-0902.1.

Proshutinsky, A. Y., and M. A. Johnson, 1997: Two circulation regimes of the wind-driven Arctic Ocean. *J. Geophys. Res.*, **102**, 12 493–12 514, https://doi.org/10.1029/97JC00738.

——, and Coauthors, 2009: Beaufort Gyre freshwater reservoir: State and variability from observations. *J. Geophys. Res.*, **114**, C00A10, https://doi.org/10.1029/2008JC005104.

Rabe, B., and Coauthors, 2014: Arctic Ocean basin liquid freshwater storage trend 1992–2012. *Geophys. Res. Lett.*, **41**, 961–968, https://doi.org/10.1002/2013GL058121.

Rigor, I. G., J. M. Wallace, and R. L. Colony, 2002: Response of sea ice to the Arctic Oscillation. *J. Climate*, **15**, 2648–2663, https://doi.org/10.1175/1520-0442(2002)015<2648:ROSITT>2.0.CO;2.

Schauer, U., and M. Losch, 2019: "Freshwater" in the ocean is not a useful parameter in climate research. *J. Phys. Oceanogr.*, **49**, 2309–2321, https://doi.org/10.1175/JPO-D-19-0102.1.

Schreiber, T., and A. Schmitz, 2000: Surrogate time series. *Physica D*, **142**, 346–382, https://doi.org/10.1016/S0167-2789(00)00043-9.

Serreze, M. C., and Coauthors, 2006: The large-scale freshwater cycle of the Arctic. *J. Geophys. Res.*, **111**, C11010, https://doi.org/10.1029/2005JC003424.

Shaffrey, L. C., and Coauthors, 2009: U.K. HiGEM: The new U.K. high-resolution global environment model—Model description and basic evaluation. *J. Climate*, **22**, 1861–1896, https://doi.org/10.1175/2008JCLI2508.1.

Shu, Q., F. Qiao, Z. Song, J. Zhao, and X. Li, 2018: Projected freshening of the Arctic Ocean in the 21st century. *J. Geophys. Res. Oceans*, **123**, 9232–9244, https://doi.org/10.1029/2018JC014036.

Steele, M., R. Morley, and W. Ermold, 2001: PHC: A global ocean hydrography with a high-quality Arctic Ocean. *J. Climate*, **14**, 2079–2087, https://doi.org/10.1175/1520-0442(2001)014<2079:PAGOHW>2.0.CO;2.

——, J. Morison, W. Ermold, I. Rigor, M. Ortmeyer, and K. Shimada, 2004: Circulation of summer Pacific halocline water in the Arctic Ocean. *J. Geophys. Res.*, **109**, C02027, https://doi.org/10.1029/2003JC002009.

Thompson, D. W., and J. M. Wallace, 1998: The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.*, **25**, 1297–1300, https://doi.org/10.1029/98GL00950.

——, and ——, 2001: Regional climate impacts of the Northern Hemisphere annular mode. *Science*, **293**, 85–89, https://doi.org/10.1126/science.1058958.

Timmermans, M. L., A. Proshutinsky, R. A. Krishfield, D. K. Perovich, J. A. Richter-Menge, T. P. Stanton, and J. M. Toole, 2011: Surface freshening in the Arctic Ocean's Eurasian Basin: An apparent consequence of recent change in the wind-driven circulation. *J. Geophys. Res.*, **116**, C00D03, https://doi.org/10.1029/2011JC006975.

Tsamados, M., D. L. Feltham, D. Schroeder, D. Flocco, S. L. Farrell, N. Kurtz, S. W. Laxon, and S. Bacon, 2014: Impact of variable atmospheric and oceanic form drag on simulations of Arctic sea ice. *J. Phys. Oceanogr.*, **44**, 1329–1353, https://doi.org/10.1175/JPO-D-13-0215.1.

Voldoire, A., and Coauthors, 2013: The CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dyn.*, **40**, 2091–2121, https://doi.org/10.1007/s00382-011-1259-y.

Wang, H., S. Legg, and R. Hallberg, 2018: The effect of Arctic freshwater pathways on North Atlantic convection and the Atlantic meridional overturning circulation. *J. Climate*, **31**, 5165–5188, https://doi.org/10.1175/JCLI-D-17-0629.1.

Wang, Q., and Coauthors, 2016: An assessment of the Arctic Ocean in a suite of interannual CORE-II simulations. Part II: Liquid freshwater. *Ocean Modell.*, **99**, 86–109, https://doi.org/10.1016/j.ocemod.2015.12.009.

——, C. Wekerle, S. Danilov, D. Sidorenko, N. Koldunov, D. Sein, B. Rabe, and T. Jung, 2019: Recent sea ice decline did not significantly increase the total liquid freshwater content of the Arctic Ocean. *J. Climate*, **32**, 15–32, https://doi.org/10.1175/JCLI-D-18-0237.1.

Watanabe, M., and Coauthors, 2010: Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *J. Climate*, **23**, 6312–6335, https://doi.org/10.1175/2010JCLI3679.1.

Wernli, H., and L. Papritz, 2018: Role of polar anticyclones and mid-latitude cyclones for Arctic summertime sea-ice melting. *Nat. Geosci.*, **11**, 108–113, https://doi.org/10.1038/s41561-017-0041-0.

Yang, J., A. Proshutinsky, and X. Lin, 2016: Dynamics of an idealized Beaufort Gyre: 1. The effect of a small beta and lack of western boundaries. *J. Geophys. Res. Oceans*, **121**, 1249–1261, https://doi.org/10.1002/2015JC011296.

Zhang, J., and M. Steele, 2007: Effect of vertical mixing on the Atlantic Water layer circulation in the Arctic Ocean. *J. Geophys. Res.*, **112**, C04S04, https://doi.org/10.1029/2006JC003732.

Zhang, X., M. Ikeda, and J. E. Walsh, 2003: Arctic sea ice and freshwater changes driven by the atmospheric leading mode in a coupled sea ice–ocean model. *J. Climate*, **16**, 2159–2177, https://doi.org/10.1175/2758.1.