Modeled Interannual Variability of Arctic Sea Ice Cover is Within Observational Uncertainty

Christopher Wyburn-Powell[a] , Alexandra Jahn[a] , Mark England[b]

[a] *Department of Atmospheric and Oceanic Sciences, and Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, Colorado*

[b] *Department of Earth and Planetary Science, University of California, Santa Cruz, California*

blished in edited form as:
yburn-Powell, C., A. Jahn, and M. R. England, 2022: Modeled Interannual Variability of Arctic Sea Ice Cover is within
bservational Uncertainty. J. Climate, 35, 6827–6842, https://doi.org/10.1175/JCLI-D-21-0958.1.

Corresponding author: C Wyburn-Powell, chwy8767@colorado.edu

ABSTRACT: Internal variability is the dominant cause of projection uncertainty of Arctic sea ice in the short and medium term. However, it is difficult to determine the realism of simulated internal variability in climate models, as observations only provide one possible realization while climate models can provide numerous different realizations. To enable a robust assessment of simulated internal variability of Arctic sea ice, we use a resampling technique to build synthetic ensembles for both observations and climate models, focusing on interannual variability which is the dominant timescale of Arctic sea ice internal variability. We assess the realism of the interannual variability of Arctic sea ice cover as simulated by six models from the Coupled Model Intercomparison Project 5 (CMIP5) that provide large ensembles compared to four observational datasets. We augment the standard definition of model and observational consistency by representing the full distribution of resamplings, analogous to the distribution of variability which could have randomly occurred. We find that modeled interannual variability typically lies within observational uncertainty. The three models with the smallest mean state biases are the only ones consistent in the pan-Arctic for all months, but no model is consistent for all regions and seasons. Hence, choosing the right model for a given task as well as using internal variability as additional metric to assess sea ice simulations is important. The fact that CMIP5 large ensembles broadly simulate interannual variability consistent within observational uncertainty gives confidence in the internal projection uncertainty for Arctic sea ice based on these models.

SIGNIFICANCE STATEMENT: The purpose of this study is to evaluate the historical simulated internal variability of Arctic sea ice in climate models. Determining model realism is important to have confidence in the projected sea ice evolution from these models, but so far only mean state and trends are commonly assessed metrics. Here we assess internal variability with a focus on the interannual variability, which is the dominant timescale for internal variability. We find that, in general, models agree well with observations, but as no model is within observational uncertainty for all months and locations, choosing the right model for a given task is crucial. Further refinement of internal variability realism assessments will require reduced observational uncertainty.

1. Introduction

Arctic sea ice has declined precipitously since 1979, at a faster rate than at any time over the last millennium (Brennan and Hakim 2022), with less than half the summer area and one quarter the summer volume remaining (Schweiger et al. 2011; Notz and Stroeve 2018). This observed decline is due to both anthropogenic climate change and internal variability, which can act to amplify or dampen the trend from external forcing alone (Kay et al. 2011; Notz and Marotzke 2012). The relative contribution of internal variability to the observed September sea ice area decline remains uncertain but has been estimated at 43-53% (Stroeve et al. 2007; Kay et al. 2011; Ding et al. 2019). Internal variability also influences future sea ice projections, leading to large internal variability uncertainty, especially for the next few decades (Kay et al. 2011; Jahn et al. 2016; Bonan et al. 2021). As internal variability is such a large contributor to the observed and projected changes in Arctic sea ice cover, but Global Climate Models (GCMs) differ in the magnitude of their simulated sea ice internal variability (Olonscheck and Notz 2017), it is imperative that we understand how realistically models simulate internal variability.

Internal variability of Arctic sea ice has been shown to be spatially heterogeneous (England et al. 2019), and act on multiple time scales from annual to multi-decadal (Zhang and Wallace 2015; Ding et al. 2017, 2019; Brennan et al. 2020). Over the historical period, internal variability has been the dominant cause of sea ice decline in many regions, most notably parts of the Kara Sea in summer and the Barents Sea in winter (Li et al. 2017; England et al. 2019; Dörr et al. 2021). Sea ice loss in recent decades has been most rapid and expansive in the summer, particularly in

the shelf seas which have transitioned from mainly ice-covered to ice-free for more of the summer, facilitating high internal variability (Onarheim et al. 2018; Mioduszewski et al. 2019). These areas of rapid and unpredictable change coincide with the most impactful areas for a range of stakeholders from shipping and oil interests to indigenous peoples and biodiversity (Kovacs et al. 2011; Petrick et al. 2017; Christensen and Nilsson 2017; Chen et al. 2020).

The established way to estimate internal variability in GCMs is to use multiple realizations of single model initial-condition large ensembles (SMILEs) or long constant forcing model runs to assess the ensemble spread or standard deviation (Olonscheck and Notz 2017; Lehner et al. 2020; Maher et al. 2020). SMILEs have successfully been used to study internal variability in the context of polar temperatures England (2021), precipitation trends (Dai and Bloecker 2019), and regional trends (McKinnon and Deser 2018; Hu et al. 2019). However, such analysis can not be done on observations, due to only one realization of reality and a limited length of the observational record. It is this single realization of reality over a relatively short period of time that has previously prevented direct assessment of internal variability of Arctic sea ice in models compared to observations. Hence, previous sea ice model assessments have been focused on the trends (e.g. Swart et al. 2015; Rosenblum and Eisenman 2017), sensitivity to warming (e.g. Winton 2011; Niederdrenk and Notz 2018), and mean state (e.g. Davy and Outten 2020). Furthermore, even if we were able to precisely disentangle internal variability from the forced response in observations, comparisons with GCMs are still challenging because we do not know where the one realization seen in the observations falls within the probability distribution obtained from a model ensemble (Notz 2015).

Here, we provide the first direct comparison of internal variability of Arctic sea ice from a suite of SMILEs from the Coupled Model Intercomparison Project Phase 5 (CMIP5) with observations, by using a statistical technique to construct a 'synthetic ensemble' of Arctic sea ice variability, following McKinnon et al. (2017). Synthetic ensembles have been used for several climate variability questions such as for sea surface temperature (Chan et al. 2020), climate extremes (Deser et al. 2020a), precipitation (McKinnon and Deser 2021), ocean chlorophyll concentration (Elsworth et al. 2021), and Antarctic sea ice trends (Chemke and Polvani 2020). Here we present the first use of a synthetic ensemble for studying Arctic sea ice, specifically to assess the realism of internal variability on interannual timescales. Using the synthetic ensemble method, we are able to

show that generally the simulated interannual variability fits within the observational uncertainty derived from different datasets, but that there are considerable seasonal and spatial differences, and that some models perform better than others for a given task. We also show that internannual variability makes up approximately three quarters of the total Arctic sea ice internal variability, and hence the majority of the sea ice internal variability over the past 42 years.

2. Data Sources

a. Observational data

We primarily use two observational datasets for sea ice concentrations (SIC), the National Snow and Ice Data Center (NSIDC) Climate Data Record (CDR) version 4 (Meier et al. 2021), and the Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST1) (Rayner et al. 2003). In order to further test the sensitivity of our results to the observational dataset used, we also utilize datasets derived from the satellite algorithums NASA Team (NT) (Cavalieri et al. 1984) and NASA Bootstrap (BT) (Comiso 1986). Together, these datasets are a representative sample of interpretations of past sea ice conditions, with both the mean state and variability differing between the datasets due to observational uncertainties (Comiso et al. 2017; Kern et al. 2019). Sea ice area (SIA) was chosen over sea ice extent (SIE) as the variability of SIA is more independent of satellite algorithms and is intrinsically more precise and thus better for comparing internal variability between models (Notz 2014). All analysis is performed using monthly data for 1979 to 2020. Missing data for NSIDC datasets, and discontinuities in the HadISST1 dataset, were filled using the same month's data in a different year, instead of interpolating to avoid unrealistic SIC values (see Table S1) for the specific replacements used).

b. Model data

Six models from the Climate Variability and Predictability Program (CLIVAR) Multi-Model Large Ensemble Archive (Deser et al. 2020b) are utilized in this analysis, as detailed in Table 1. EC-Earth was excluded from this analysis due to no available SIC output. All models are CMIP5-class and use historical and representative concentration pathway (RCP) 8.5 forcing, the

TABLE 1. Models used in this analysis from the CLIVAR Multi-Model Large Ensemble Archive (Deser et al. 2020b)

Modeling Center	Model	Members	Years	Reference
CCCma	CanESM2	50	1950-2100	Kirchmeier-Young et al. (2017)
NCAR	CESM1	40	1920-2100	Kay et al. (2015)
CSIRO	MK3.6	30	1850-2100	Jeffrey et al. (2013)
GFDL	CM3	20	1920-2100	Sun et al. (2018)
GFDL	ESM2M	30	1950-2100	Rodgers et al. (2015)
MPI	ESM1	100	1850-2100	Maher et al. (2019)

high emissions CMIP5 scenario. The models from the Multi-Model Large Ensemble Archive are diverse in their mean state and trends, spanning nearly the full range of CMIP5 sea ice projections (see Figure 1 in Bonan et al. 2021). In winter, model mean-state biases are typically smaller in absolute and relative terms than summer (see Table S2). The notable outliers in summer are CanESM2 with the largest negative mean-state bias of −54% in September and CSIRO MK3.6 being an extreme positive outlier for all seasons and +83% in September. Although GFDL CM3 is not as large an outlier in mean state in September, its SIA loss over the period 1979-2020 is by far the most rapid. The six models range in ensemble size between 20 and 100 (see Table 1). We present results for all members of the SMILEs to assess each GCM's ability to realistically simulate the observed interannual variability. We also provide subsampled results, scaled to 20 members, the size of the smallest large ensemble, for model inter-comparison with our consistency metric. Subsampling is discussed in more detail in section 3c.

3. Methods

a. Resampling technique

We estimate interannual variability in a single model member or observational time series by assuming the forced response is represented by an ordinary least squares regression linear trend. This assumption is deemed appropriate for 1979-2020, but may not be applicable for time periods extending further back (England et al. 2019; England 2021), and allows us to follow the methodology from McKinnon et al. (2017). Anomalies from this linear trend are therefore considered largely due to interannual variability alone. Typically the ensemble mean is a more

accurate measure of the forced response, but as discussed in section 3d and 4c, detrending using the individual member produces similar results and is chosen here for reasons detailed in 3d.

By using this technique we can calculate a consistent metric of interannual variability to directly compare model members and observations. We consider all months in the pan-Arctic and present spatial results for the minimum and maximum SIA months September and March respectively. We resample the anomalies from the linear trend 10,000 times for SIA and 1000 times for each SIC grid box, with replacement, and use a 2 year bootstrap block size. This can be considered analogous to shuffling independent anomalies to produce a range of alternative scenarios which would have been equally likely to occur, allowing us to calculate metrics of interannual variability for a representative sample of all possible scenarios (see Figure 1). As suggested by McKinnon et al. (2017) and McKinnon and Deser (2018), we retain spatial coherence by resampling in the time dimension for all grid boxes at once. 10,000 resamplings in the pan-Arctic were chosen for increased reliability of consistency classifications, whereas spatially 1000 was determined sufficient as each grid box has a lower impact on results if a classification were to change from rerunning the experiment. A 2 year block size is chosen because normalized autocorrelation frequently exceeds 0.4 for a lag of 1 year, and a marked drop-off in autocorrelation between a lag of 1 and 2 years occurs in comparison with years 2 and 3 (not shown), occurring both spatially and in the pan-Arctic time series. Resampling with a 1 year or 2 year block size leads to almost identical results (not shown).

We focus our analysis on the standard deviation of sea ice state over the 42-year period 1979-2020, not the trends, as we want to assess the realism of the models' simulated interannual variability, rather than the realism of the simulated trends (see Swart et al. (2015) for a discussion of simulated trends compared to observations). The standard deviation with respect to time is computed either for the 10,000 pan-Arctic SIA resampling or the 1000 SIC resamplings in each grid cell. To represent the distribution of these resamplings or ensemble members we use the standard deviation (σ) and mean (μ). Here, σ can be considered analogous to the range of interannual variability which could have occurred, given the underlying data; μ is analogous to the typical interannual variability represented in the resamplings.

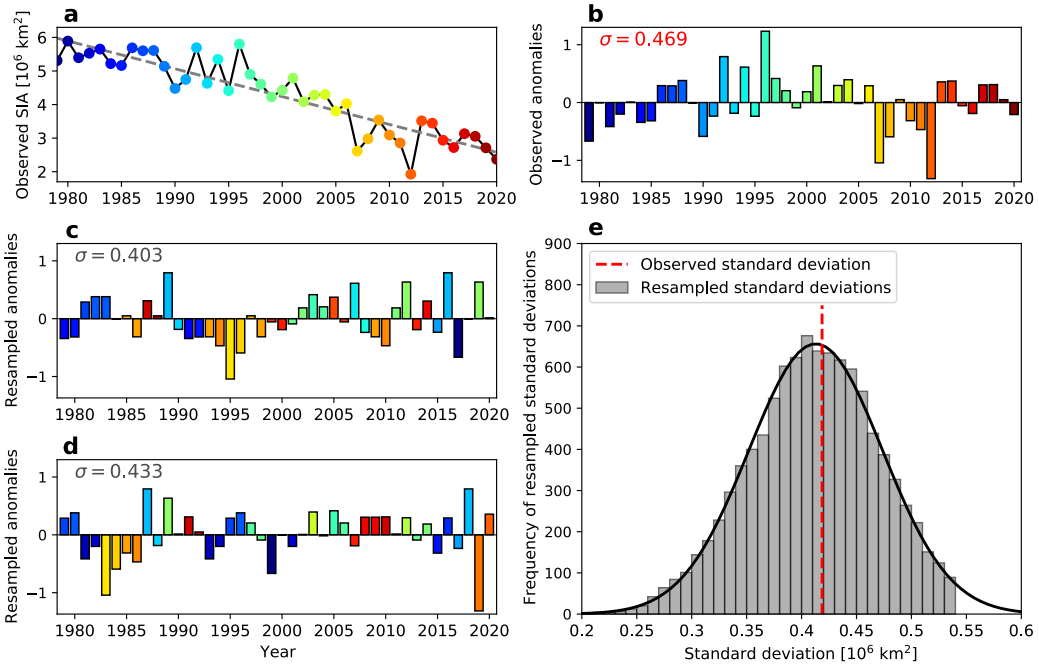


FIG. 1. **Resampling methodology, applied to the observed September SIA.** (a): Observed sea ice area from CDR (dots) with linear trend (grey dashed line). (b): Anomalies from the linear trend. (c,d) Two randomly different resamplings of the anomalies in (b), color coded to match the year of anomaly. (e): Distribution of the standard deviation with respect to time for all 10,000 resamplings. The printed statistics represent σ for standard deviation. In (e) the red vertical line represents the standard deviation of the original data and gray refers to the distribution of standard deviations for the 10,000 resamplings.

In order to directly compare interannual variability between models and observations we define three measures of variability as follows, where σ_{LE} is internal variability in SMILEs and both σ_{mem} and σ_{obs} are the interannual variability within a synthetic ensemble:

- σ_{LE} and μ_{LE} - Standard deviation and mean of standard deviations within a single large ensemble, without resampling, an established measure of the full range of internal variability.

- σ_{mem} and μ_{mem} - Standard deviation and mean of the standard deviations of all resamplings of a single model member. The resampling process for a given ensemble member is equivalent

to that of the observations in Figure 1. The median member's value across all members of the SMILE is denoted $\bar{\sigma}_{mem}$ and $\bar{\mu}_{mem}$.

- σ_{obs} and μ_{obs} - Standard deviations and mean of the standard deviation of all resamplings of the single realization of the observational dataset. These metrics relate to Figure 1 as the standard deviation and mean of the distribution in panel e.

b. Consistency

To assess the realism of simulated internal variability, we utilize a consistency metric to provide a binary classification as to whether the modeled variability is within or outside the range of observational uncertainty. For sea ice analysis in the past, consistency has typically been defined by at least one member of a large ensemble overlapping with observations (e.g. Notz 2015; Swart et al. 2015; Jahn 2018). However, this is a relatively low bar for models to reach. Other more elaborate consistency methods have been applied for other aspects of the climate system, e.g. Santer et al. (2008), and applied to Arctic sea ice by Stroeve et al. (2012). However the methodology of Santer et al. (2008) bases consistency assessments on trends rather than the internal variability independent of the trends, as is the goal here. Hence, we here use resampling and define consistency by comparing distributions, as it allows us to compare whether the resampled distributions overlap. Comparing distributions is a more stringent decision about consistency than comparing single values for each ensemble member or observational dataset that would be available without resampling. Further augmentation to this binary classification is achieved by comparing SMILEs with four diverse observational datasets independently, adding the category of 'consistent within observational uncertainty'. We only use this three-category consistency classification rather than a significance or probability value (e.g., from a student t-test), as both the resampled average variability (μ_{mem}) and standard deviation of variability (σ_{mem}) are positively skewed across members. Nonetheless, we find that a 95% confidence interval is in fact similar to our consistency classification, but classifies fewer instances of inconsistency in the pan-Arctic than our method.

Applying this consistency metric to Arctic SIA, each SMILE realization or observational dataset has a different value of interannual variability for each of the 10,000 resamplings. These 10,000

resamplings from a single member or observational time series are approximately normally distributed and as such can be thought of as probability distribution functions (PDFs) (see Figure 2). The width of the PDFs show the distribution of the 10,000 resamplings, indicating the range of possible interannual variabilities (proportional to σ_{mem} and σ_{obs}). The location on the horizontal-axis indicates the average interannual variability (μ_{mem} and μ_{obs}). For models and observations to be considered 'consistent' in the following, we require their means (their position on the horizontal-axis in Figure 2) and their standard deviations (height on the vertical-axis) to overlap such that at least one member is greater than the lowest observational dataset and one member is lower than the highest observations for each σ and μ metric independently. Average SIC differences does not preclude a consistent classification as variability may be equal between a SMILE and observational datasets but about different means. However, due to the zero-bound nature of SIC, if a mean state differs so much that SMILE members have at least some sea ice where there is no sea ice in the observational datasets, we exclude those regions from the analysis rather than classifying them as inconsistent. We do this as the focus of our analysis is on assessing the realism of actual sea ice variability, so we only compare regions where there is variability in both models and observations.

c. Ensemble size

We have included SMILEs with ensemble members as low as 20 in our analysis as the standard deviation between members (σ_{LE}), representing the full range of internal variability, increases only marginally beyond approximately 8-12 members, compared to the full range of 20-100 members (see Figure S2). This leads us to consider SMILEs of at least 12 members to generate enough diversity between realizations to capture most aspects of internal variability. The selection of a minimum number of members for SMILEs when assessing different time periods or other aspects of the climate system may require considerably more members (Milinski et al. 2020). With increasing ensemble size, the values of the minimum and maximum σ_{mem} diverge, making it easier for a SMILE to overlap with observations (see Figure S1). Our primary goal is to assess SMILEs' individual realism when compared with observations, using as much information from each model as is available. Hence, we present results without subsampling the members to a consistent ensemble size. However, as others may be interested in a direct comparison of the interannual variability in CMIP5 SMILEs, we provide subsampled results in the supplementary

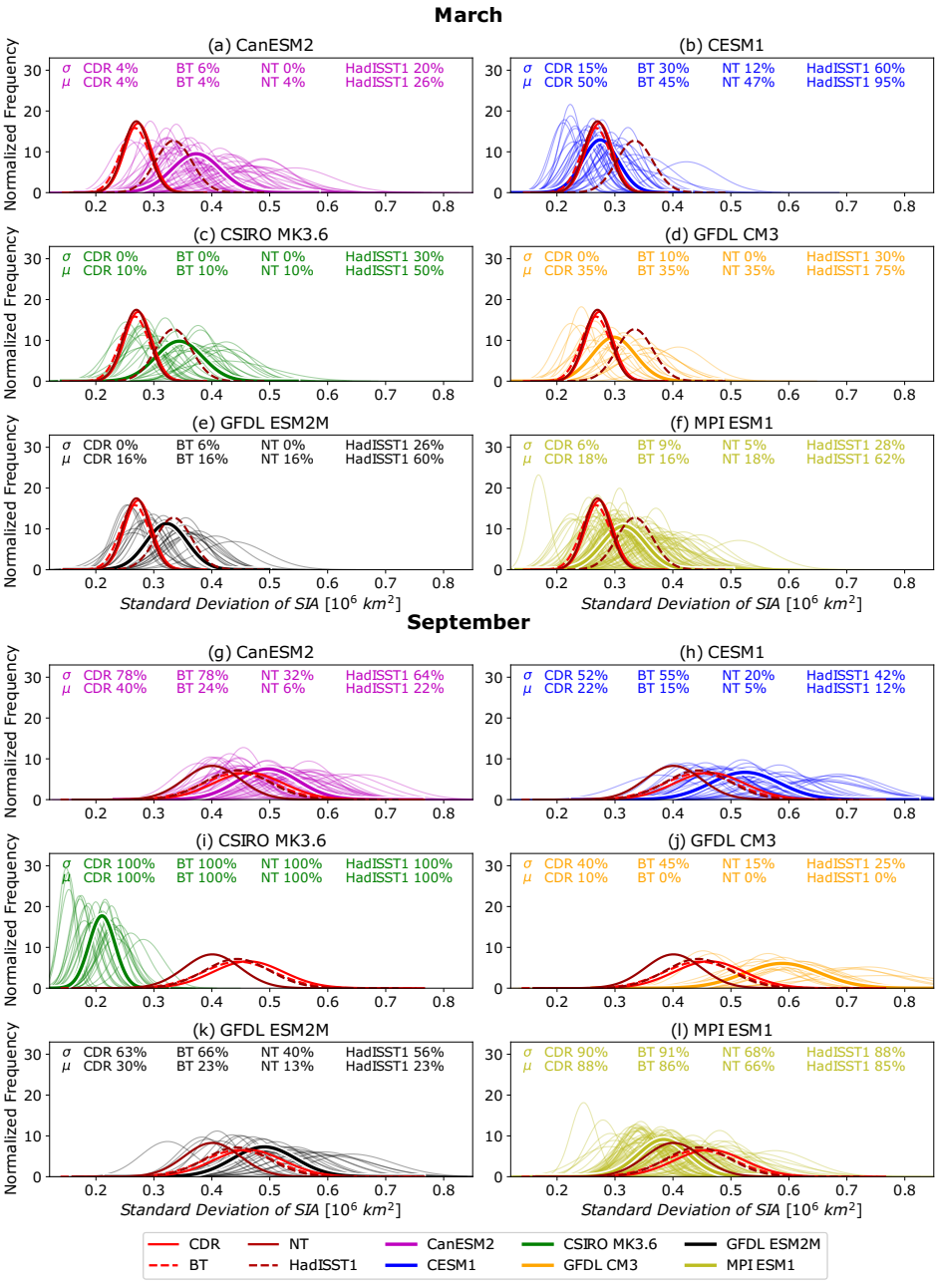


FIG. 2. **Distribution of pan-Arctic SIA standard deviations across members, resamplings, and observations.** Probability distribution functions (PDFs) for detrended standard deviation of pan-Arctic SIA, for March (a-f) and September (h-l). PDFs are produced from the mean (μ) and standard deviation (σ) across the 10000 resamplings. Each individual resampled member (σ_{mem}) is plotted with a thin lines colored according to the legend, the average resampled member ($\bar{\sigma}_{mem}$) is colored similarly with a thick line, and the resampled observations (σ_{obs}) are in red for the four datasets according to the legend. Percentiles noted on the figure are the single values of σ_{obs} or μ_{obs} for the observational datasets relative to the distribution of σ_{mem} and μ_{mem} across members.

section, where consistency is standardized to 20 members, the size of the smallest SMILE, in the pan-Arctic (Figure S5) and spatially (Figure S6).

d. Detrending

The ensemble mean of a SMILE is considered a good representation of 'forced response' of the model to the changing climate (Frankcombe et al. 2018). However, observations only have one realization, and hence the observed forced trend must be computed from that single realization. Hence, in our analysis we use the individual members' trends over the period 1979-2020 as representation of the forced response, to enable the same methodology to be applied to observations and models, for direct comparisons. The SMILEs provide the perfect place to test the impact of this method: we find that linear detrending rather than removing the ensemble mean, results in only a marginal decrease in variability (8% reduction for σ_{mem} and 11% for σ_{LE}) yielding a very similar ratio (see Figure S11).

Applying linear detrending largely removes low frequency variability. We reached this conclusion as detrending ensemble members and observations using a 2 year 5th order lowpass Butterworth filter (Roberts and Roberts 1978), which explicitly removes low-frequency variability, obtains almost identical consistency results as with a simple linear trend (see Figure S7 in comparison to Figure 8). This lowpass filter removes variability on frequencies in excess of 2 years, the time period beyond which autocorrelation in the sea ice is negligible. Good agreement between the linear detrending and the lowpass filtered data suggests that both anomaly calculation methods effectively isolate interannual variability. The variability in our resampled anomalies of individual SMILE members (σ_{mem}) capture approximately three quarters of internal variability across SMILE realizations without resampling (σ_{LE}), as discussed further in section 4c. This enables us to conclude that our detrending and resampling analysis primarily assesses interannual variability, and that this is the dominant timescale of internal variability for Arctic sea ice for the period 1979-2020.

In the spatial analysis we obtain a linear trend for each grid cell, using the same method of detrending as we did for the pan-Arctic. While we find some isolated incidences of grid cells where the linear SIC trend exceeds 100% or is lower than 0%, extremely small differences are

found in consistency if a different detrending method is used, such as a 2 year low pass filter (see Figure S7) or trends capped to physical bounds (not shown). Hence, the detrending method does not affect the conclusions drawn from the analysis.

e. Time periods

The time period considered is the observational period 1979-2020, focused on the seasonal extremes of March and September for the spatial analysis. 1979-2020 is chosen for observations due to high quality spatial data from 1979 onward, which is particularly important for assessing interannual sea ice variability. We found that shifting the time period used from the models to better match the observed mean sea ice state yielded negligible differences spatially and minimally affected pan-Arctic results for shifts of a few years to a decade. When matching the observed mean state required adjustments of many decades, the changes in the results were larger. However, in some instances a model did not have a time period where the mean state matched the observed mean state in the whole historical and future simulations. Furthermore, we want to assess the realism of the simulated interannual variability as simulated, to complement previous model assessments of trends and mean state which were done over the same periods in models and observations (e.g. Swart et al. 2015; SIMIP Community 2020). Hence, although internal variability has been shown to be sensitive to mean state (Goosse et al. 2009; Jahn et al. 2016; Olonscheck and Notz 2017; Massonnet et al. 2018), and some models have more linear SIA and SIC declines than others over 1979-2020, we find the choice of exact period analyzed did not materially impact our results.

As the use of a 42-year time period is out of necessity, this raises the question of whether a 42-year period is sufficient for our analyses. To assess this question different time period lengths were assessed within the models (see Figure 3). Time periods longer than approximately 20 years yield similar ratios in $\sigma_{mem}/\sigma_{obs}$ ratios, which gives confidence in our results for this metric being representative of a broad range of time periods. Similarly, the ratio σ_{mem}/σ_{LE} changes rapidly for short time periods but becomes relatively stable for time periods of at least a few decades. To confirm this, we conducted a similar time period analysis for the period 1953-2020 using lowpass filtered SIA. This more clearly indicates the stabilization of the ratio σ_{mem} to σ_{LE}, at approximately 75%, independent of the length of the time period in excess of approximately 30 years (see Figure

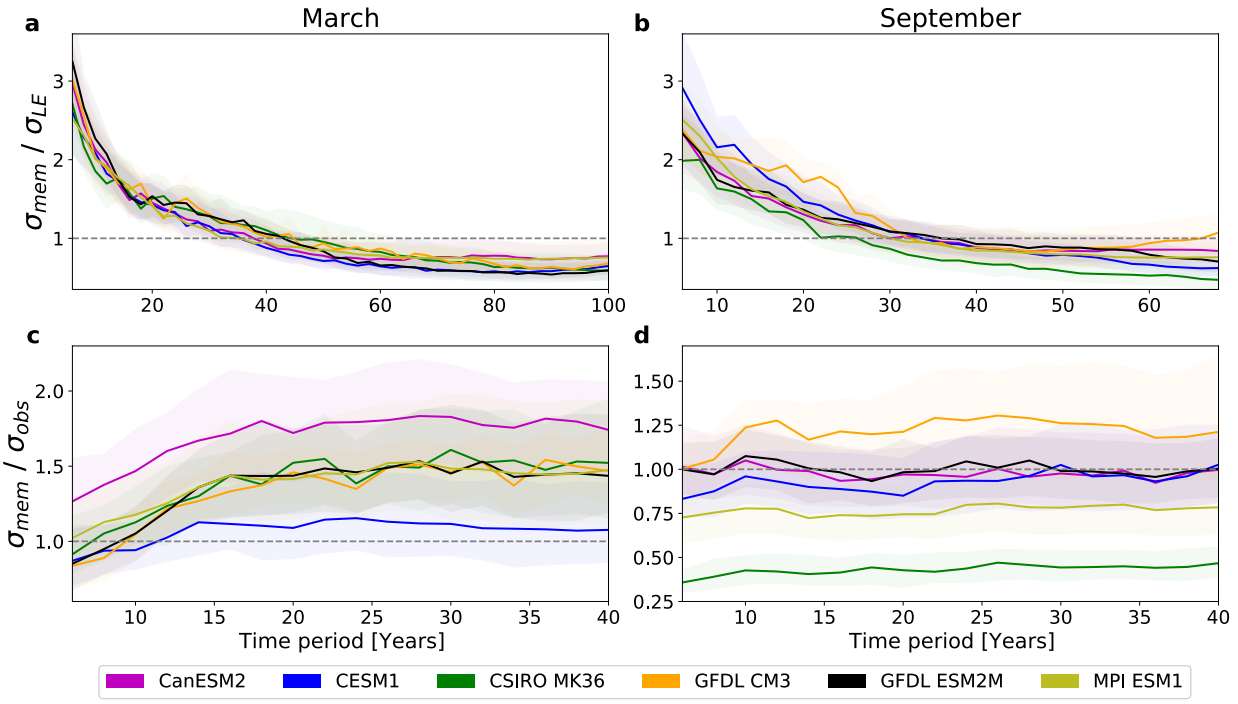


FIG. 3. **Influence of length of time period on standard deviation of pan-Arctic SIA**. Standard deviation with respect to time for time periods between 6 years and the maximum length of a linear trend in SIA, bootstrapped 1000 times. Thick lines show the median ensemble member, shading shows ± 1 standard deviation. (a, b): the ratio of standard deviation across resamplings (σ_{mem}) to standard deviation across members (σ_{LE}) over a subset of the time periods 1965-2066 for March (a) and 1970-2040 for September (b). (c, d): the ratio of standard deviation across resamplings (σ_{mem}) to standard deviation across resampled observations (σ_{obs}) in the HadISST1 dataset for the period 1979-2020 in March (c) and September (d).

S3). Spatially, when we compare the shorter 32-year time periods 1979-2010 and 1989-2020 to the full time period of 1979-2020, we find that there are small consistency differences between the time periods for some regions, but these differences are not substantial enough for our main conclusions to be altered (see Figure S4).

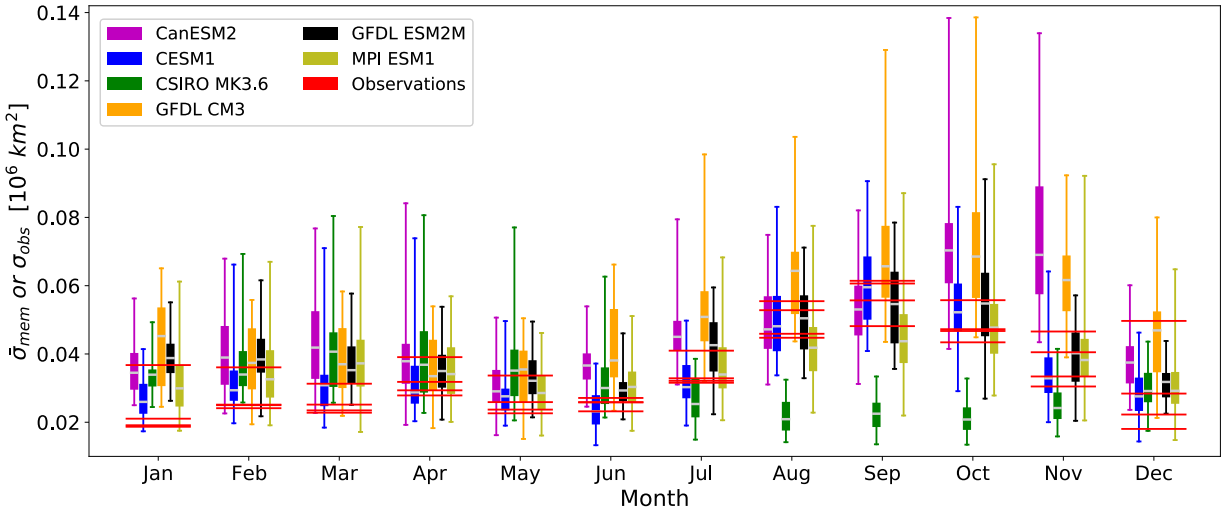


FIG. 4. **Seasonality of resampled variability in members and observations for pan-Arctic sea ice area.** The distribution of Standard deviations (σ_{mem}) across members is shown for each model and month as a box and whisker charts, where whiskers show the full range of members, boxes show the interquartile range, and gray bars indicate the median member. Values of resampled variability in observations (σ_{obs}) are shown as horizontal lines for each of the four datasets.

4. Results

a. Resampled variability in models and observations

Resampling the observations and SMILE models, we find that the variability of models is generally similar to observations, but with considerable seasonal and regional variability. Both the variability in models (σ_{mem}) and observations (σ_{obs}) show distinct seasonality in the pan-Arctic, peaking in the autumn with the exception of CSIRO MK3.6 (see Figure 4 and shown for average variability (μ) in Figure S8). In spring we find larger variation between different realizations of the same model than between model averages. This highlights the sensitivity of interannual variability to realization, and why we assess realism based on consistency rather than comparison between median SMILE member and observations (see section 3b). The results of this consistency assessment are discussed further in section 4b.

Observations have substantial uncertainties that impact the value of observational interannual variability (σ_{obs}). Hence, the choice of which dataset to use for comparison with models can affect

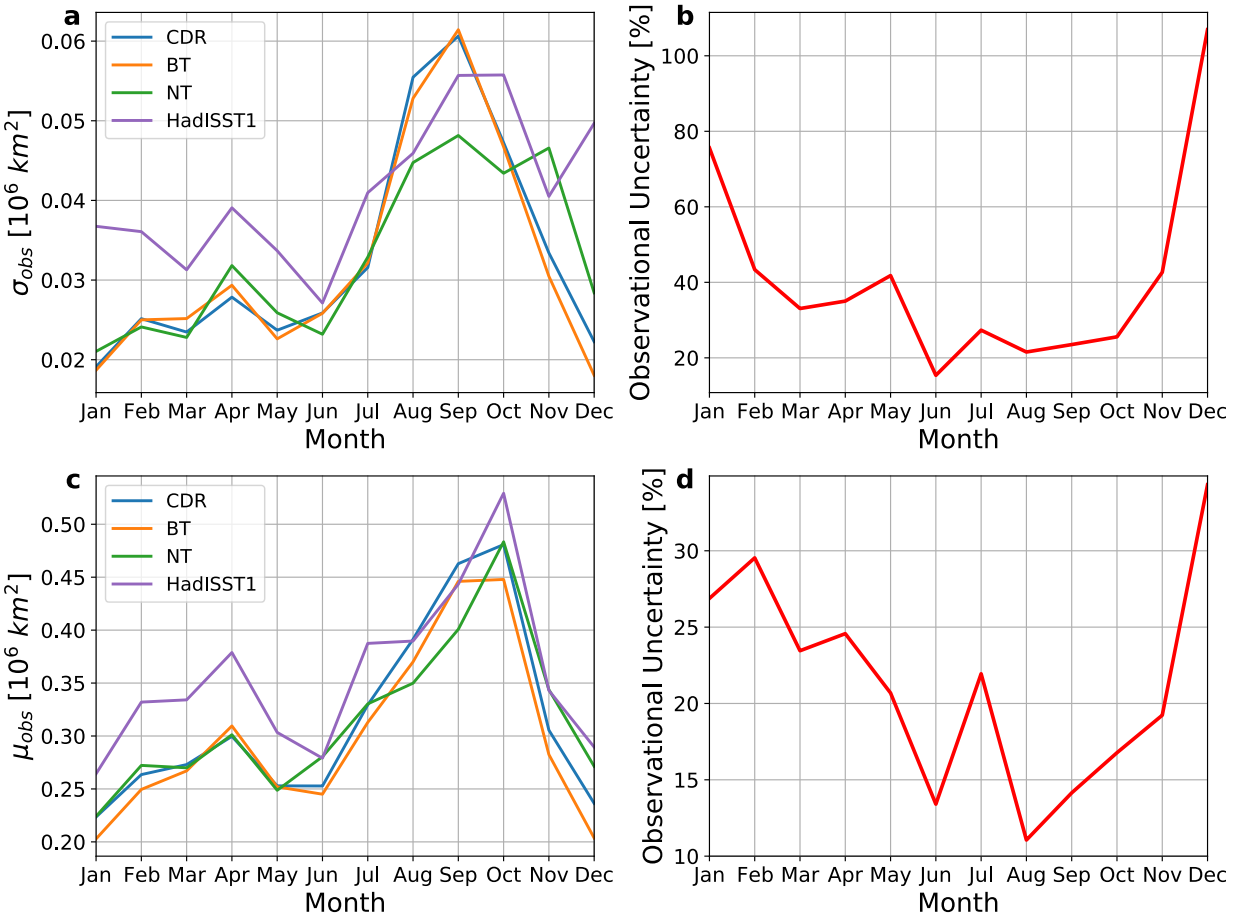


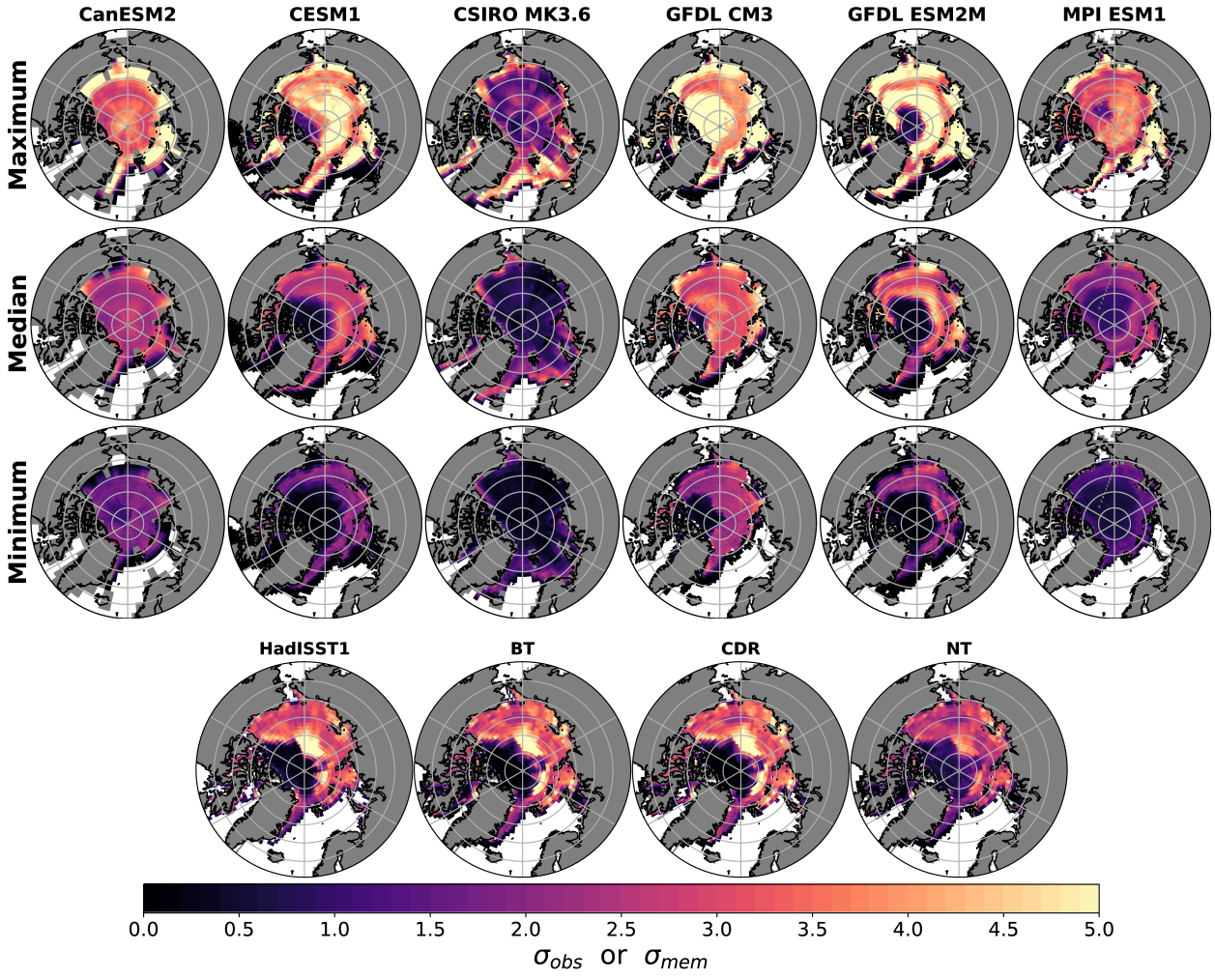
FIG. 5. **Resampled variability of pan-Arctic sea ice area for the four observational datasets.** Absolute values shown in a) and c), and percentage uncertainty shown in b) and d) as calculated from the range of σ_{obs} divided by the mean of σ_{obs}.

whether observations fall within the large ensemble range, both for the pan-Arctic and spatially (see Figures 4 and 6). Furthermore, the uncertainties vary seasonally, with the largest relative uncertainty of pan-Arctic observational variability in the winter and spring (see Figure 5). Hence, it is easier for models to fall within the observational uncertainty in the winter and spring than in the summer and autumn. For most months we find the majority of the ensemble median variability ($\bar{\sigma}_{mem}$, grey bars in Figure 4) are similar or higher than observations (σ_{obs}, in red). However, as we do not know how typical observations are, we cannot use these differences to diagnose model biases.

Spatially, there is considerable difference in the locations of maximum variability between models and the observational datasets in September (see Figure 6). We find large magnitude differences throughout the ice covered region between different models and when comparing models with observations. Despite these large differences in the ensemble medians between models, we find that the range between members for a given model is considerably larger in most instances. Again this draws attention to the difference of interannual variability between realizations. In comparison to September, the location and magnitude of highest variability in March is more similar between different models, with the range between members very large for the ice edge region (see Figure S9). Observational uncertainty is also highly variable between regions, for example NT exhibits much higher variability in the central Arctic in September than the other datasets (see Figure 6). When we combine both the spread of model simulations across realizations and the spread of interpretations of the observational record, we find broad agreement between models and observations. This is true both in the pan-Arctic and spatially in their representation of Arctic sea ice interannual variability.

b. Consistency of models and observations

When utilizing the range of observational datasets for the pan-Arctic, we find model consistency for a majority of the time (57%) across models and months (see Figure 7i). Models consistent within observational uncertainty accounts for 33% of months, far greater than the 10% of months identified as inconsistent. It is important to note that these proportions relate to the specific six models we analyzed, which capture the full spread of the CMIP5 sea ice simulations (Bonan et al. 2021). Nonetheless, the common pattern is for GCMs to be predominantly consistent within observational uncertainty. By our definition of consistency, all models except CSIRO MK3.6 and GFDL CM3 are consistent in September for all observational datasets. In the spring, when observational uncertainty is largest, we find all models are consistent within observational uncertainty and in April and May all models are consistent with all observational datasets. When looking across all months we find that only MPI ESM1 is unambiguously consistent with all observational datasets and CESM1 and GFDL ESM2M are consistent but not for all observational datasets. CanESM2, CSIRO MK3.6 and GFDL CM3 (the models with the largest mean-state bias) are the only models with inconsistent classifications beyond observational uncertainty. Our ability to more stringently



348 FIG. 6. **Resampled modeled and observed variability of September sea ice concentration.** Standard
349 deviation of resamplings for the six models (σ_{mem}) for the maximum, median and minimum member for each
350 grid cell in rows 1-3. Standard deviation of resamplings for the four observational datasets (σ_{obs}) along the
351 bottom row. The color bar applies to all subplots on this figure. The same analysis for March is shown in Figure
352 S9.

368 assess realism by using the two metrics is demonstrated by CanESM2 and GFDL CM3 being
369 considered consistent for all months for σ, but when also considering μ, we find both have two
370 months with inconsistencies.

380 When considering consistency spatially, each grid cell can be considered to have a distribution
381 of PDFs similar to Figure 2 and thus can be categorized in the same way. Consistency in σ and μ
382 are highly correlated but with some differences, indicating the benefit of using both metrics (areas

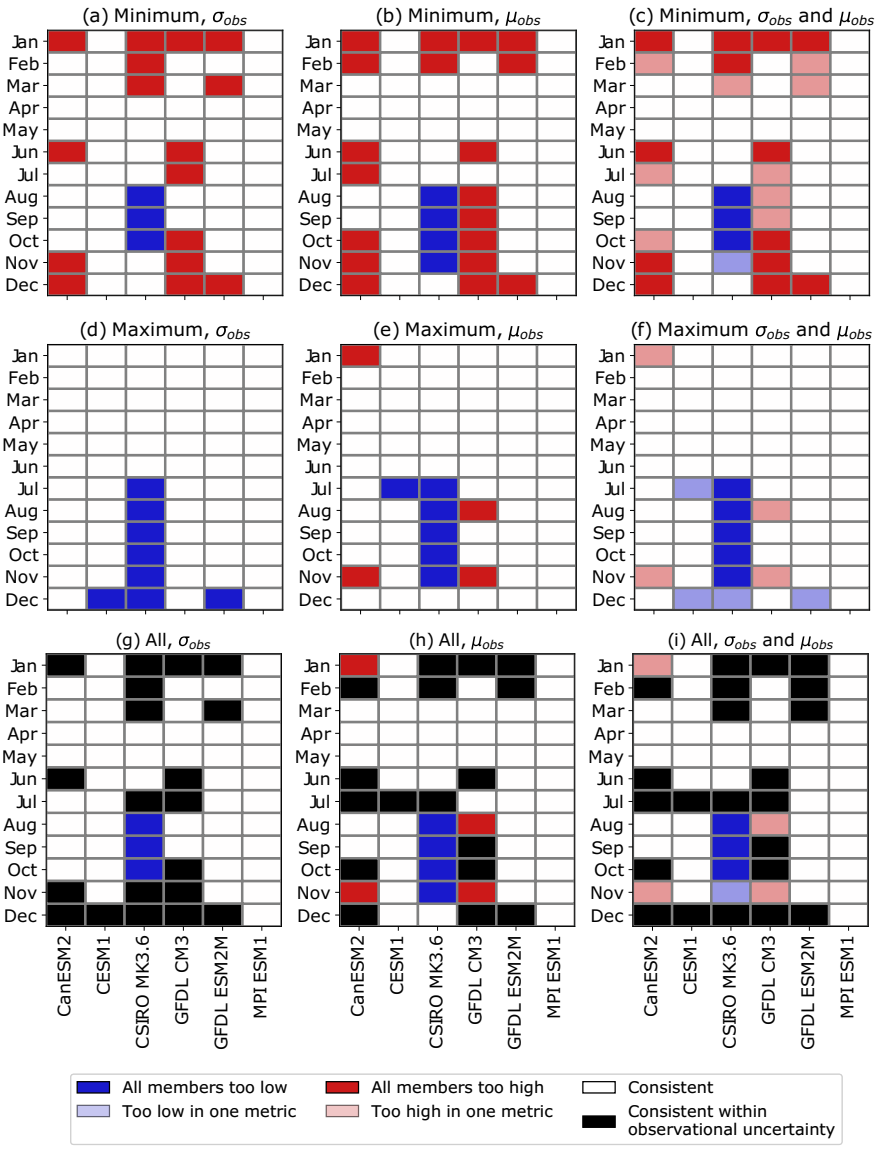


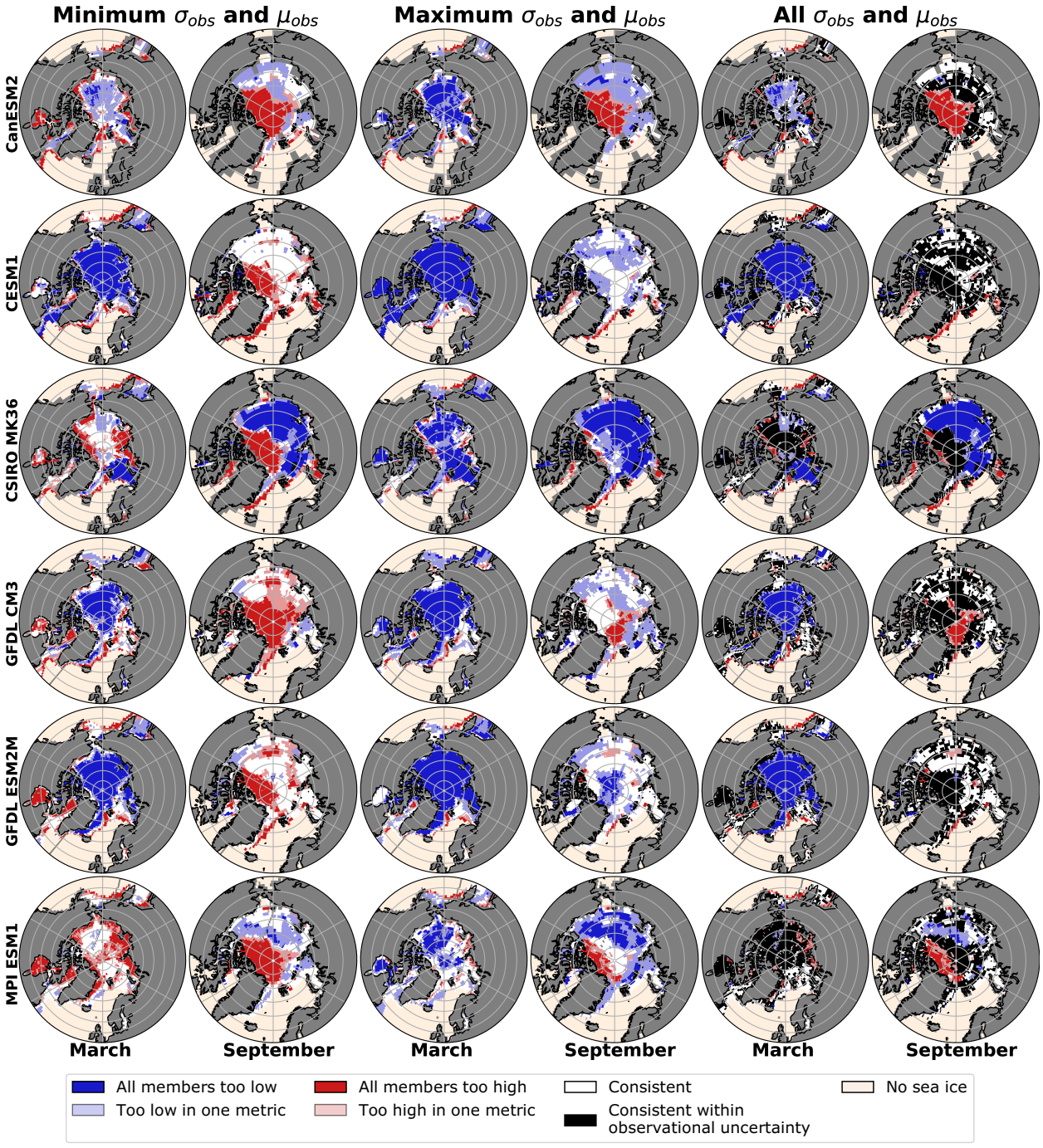
FIG. 7. **Consistency between models and observations in pan-Arctic SIA**. White indicates consistency between models and all observational datasets, while reds and blues indicate inconsistency in at least one metric. Specifically, dark blue indicates the model is inconsistent with observations as all members are too low while dark red indicates inconsistency due to all members being too high. In the third column (c, f, i), where two metrics are combined, light blue means one of the metrics classifies the model as too low while the other metric is consistent, light red indicates that the model is too high in one metric but consistent in the other metric. There are no instances of too high and too low classifications for a given month by the different metrics. Finally, in the bottom row (g-i), where all observational products are combined, black indicates disagreement in classification between the observational datasets, indicating consistency within observational uncertainty.

of light blue and light red in Figure 8). As noted earlier, we focus on the seasonal minimum and maximum sea ice area, September and March respectively and present a consistency classification only where both the model and observations exhibit non-zero sea ice.

Similarly to the pan-Arctic, we find no areas where the σ and μ metrics produce different signs of inconsistency. With the exception of CSIRO MK3.6, the shelf and marginal seas in September in all models are broadly consistent within observational uncertainty, with CESM1 and GFDL ESM2M performing the best. CSIRO MK3.6 shows the largest inconsistencies in March with underestimation of variability in the Barents Sea. All other models simulate consistent variability in the Barents Sea where atypically rapid SIC decline has occurred (Li et al. 2017). Both regions of too high variability and too low variability occur for MPI ESM1 in September, yet this model is consistent for September in the pan-Arctic, indicating these regions counteract each other for SIA. For March the models are more dissimilar than in September, with no regions of over or underestimation of interannual variability common to all models. Large portions of the central Arctic Ocean have very little observed and modeled variability in March, due to the 100% bounding of SIC. This means that small absolute biases in the modeled interannual variability can cause an inconsistent classification (see Figure S9). With our consistency classification we conclude that more models have greater realism of simulated interannual variability in September than in March. However, even well performing models in some regions in September or March generally do a poorer job in the other month, indicating that the skill of a certain model in simulating interannual variability is highly seasonally and regionally dependent.

c. Internal variability captured by resampling versus ensemble spread

Our best estimate of the full range of internal variability, on high and low frequency time scales, is through SMILEs, here we use the standard deviation between detrended members (σ_{LE}) to represent this. As we consider the resampled standard deviation of SMILE members and observations to be representative of interannual variability and not the full range of internal variability, we would expect the ratio σ_{mem}/σ_{LE} to be less than one. For all seasons, when looking at pan-Arctic SIA, interannual variability simulated by the median standard deviation across resamplings ($\bar{\sigma}_{mem}$) is less than the internal variability simulated by multiple realizations without resampling (σ_{LE}), an



403 FIG. 8. **Spatial consistency of interannual variability between large ensemble members and observations**.
404 Members of the large ensembles which have at least one member overlapping with the variability of resampled
405 observed SIC is shown in white, indicating consistency. Regions where the classification differs between the
406 maximum and minimum observational datasets are shaded black indicating consistency within observational
407 uncertainty. Areas without sea ice, either in the model or observations, are shaded beige. Shaded areas of red
408 and blue indicate inconsistency in at least one metric, using the same color scheme as in Figure 7.

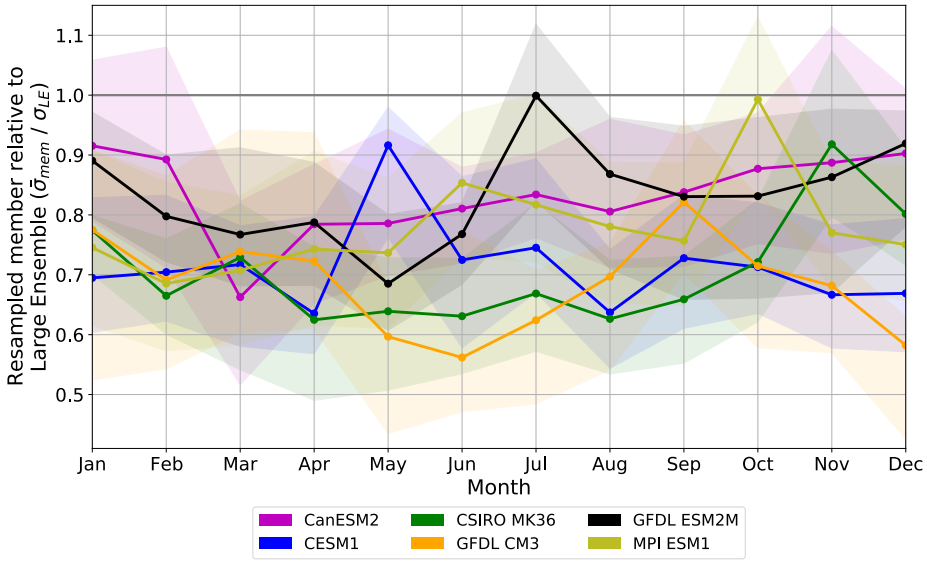


FIG. 9. **Seasonality of the ratio of internal variability across SMILEs and interannual variability of resampled members for pan-Arctic sea ice area.** Lines show the ratio of the standard deviation of the median resampled member to the standard deviation across members without resampling ($\bar{\sigma}_{mem}$ to σ_{LE}), shading shows the interquartile range of the ratios for all members.

annual average of 75.9% across models (Figure 9). This ratio is robust irrespective of detrending method with an average of 74.4% and 82.4% when the ensemble mean or a 2 year lowpass filter respectively is used for detrending (see Figure S11).

This ratio of three quarters interannual variability and one quarter lower frequency variability also holds for different time period lengths, as discussed in section 3e, and is relatively stable for a given 42 year time period sometime between 1950-1991 and 2050-2091 (see Figure **??**). Hence, we expect interannual variability to remain the dominant portion of internal variability for the near future. The general underestimation of the resampled variability, compared with the benchmark of large ensemble spread, is in agreement with previous uses of this methodology on surface temperature, precipitation and sea level pressure (McKinnon et al. 2017; McKinnon and Deser 2018). When considering the difference between σ_{LE} and σ_{mem} spatially, we find the largest underestimations along the ice edge but in general the signal in the pan-Arctic is replicated homogeneously across the Arctic (see Figure S10).

5. Discussion

Sea ice poses unique challenges in assessing internal variability: short time period of high quality observations, physical bounds of 0 to 100%, and changes in variability as mean state changes. Despite this, we were able to apply the synthetic ensemble method to Arctic sea ice as used in McKinnon et al. (2017) and McKinnon and Deser (2018) for temperature, precipitation and sea level pressure. Similarly to previous research, we found that resampling leads to an underestimation of the full range of internal variability captured by a large ensemble, both in the pan-Arctic (where $\bar{\sigma}_{mem} \approx 0.76\,\sigma_{LE}$, see Figure 9), and also locally across the Arctic Ocean (Figure S10). This agrees with the expectation that low frequency variability is not fully captured by the resampling (McKinnon and Deser 2018). Hence our analysis primarily assesses the interannual component of internal variability. Interestingly, this proportion of three quarters of the internal variability being due to interannual variability matches closely with the 75% contribution from atmospheric temperature fluctuation to Arctic sea ice variability found by Olonscheck et al. (2019) via a 'decoupling' methodology. Both of these independent analyses hence suggest that Arctic sea ice interannual variability is largely unpredictable.

Our analysis assumes that a given anomaly is equally likely to have occurred in 1979 or 2020. This is a dependable assumption, despite the fact that it has been shown that variability increases as sea ice extent decreases (Goosse et al. 2009; Jahn et al. 2016; Olonscheck and Notz 2017; Massonnet et al. 2018), as we showed that neither the length of the period considered (Figure 3) nor the period itself (Figure S12) substantially change the results. However, as the Arctic approaches seasonally ice-free conditions, an "equally likely" assumption will no longer be a valid approach. For example, it would not be appropriate to assume a September SIA negative anomaly of one million square km (as occurred in 2007) would be equally likely to occur when the mean state in September is practically zero in most models.

All of the SMILEs, except CSIRO MK3.6, capture the seasonal cycle of σ_{mem} and μ_{mem} with highest values in the summer. However, the magnitude of observational uncertainty also needs to be taken into account as it factors into how stringent consistent classifications are. Observational uncertainty is largest in the winter for the pan-Arctic (see Figure 5), therefore it is easier for models to be consistent during this part of the year. Spatially we find the largest differences in

variability between observational datasets in the central Arctic during September (see Figure 6). Nevertheless, we still find that most models simulate too high variability in this region in September, and it is only the extreme variability of NT compared with the other observational datasets that allow a 'consistent within observational uncertainty' classification for most models (see Figure 8). Consensus regarding which observational dataset is the most realistic for these areas would be required before determining which models have the better representation of variability in the high SIC regions.

As we have shown that almost all models can simulate consistent members across seasons, we can say most of the SMILE models are realistic in their simulation of historical interannual variability. Realism of internal variability is a complementary assessment to the analysis of mean state, sensitivity to warming, and trends (Swart et al. 2015; Rosenblum and Eisenman 2017; Winton 2011; Niederdrenk and Notz 2018; Davy and Outten 2020). Some of these metrics are inter-related but each provide part of the picture for a full model assessment for Arctic sea ice. We show the CMIP5 models with inconsistent months or large regions of inconsistency are those with the largest mean state biases, but even these models are consistent for several months of the year in the pan-Arctic and for most regions in March and September. This suggests that avoiding mean state biases is important for correctly simulating the evolution of the Arctic sea ice cover (see Massonnet et al. (2018)), but models can have moderately large mean-state biases and still simulate realistic sea ice interannual variability. Furthermore, as we find that most CMIP5 SMILE models agree with observations in terms of their interannual variability for the pan-Arctic in September, the internal variability prediction uncertainty of an ice-free Arctic of over two decades from climate models (Notz 2015; Jahn et al. 2016) is likely realistic. However, no SMILE model performs well in all months and regions. But if one wishes to only focus on one season or region, one can find a CMIP5 SMILE model where the interannual variability is consistent with observations. This is true even for hotspots of internal variability such as the Barents Sea in winter and the shelf seas in summer (England et al. 2019; Bonan et al. 2021), showing the robustness of the consistency classification.

6. Conclusions

In this study, we showed that simulated interannual variability of CMIP5 large ensemble models is typically within observational uncertainty, by generating a synthetic ensemble of Arctic sea ice variability and using a binary classification of consistency that considers the full distribution of resamplings to aid the assessment of model realism. This analysis method considers approximately three quarters of Arctic sea ice internal variability, on the dominant interannual timescale for the period 1979-2020. Sea ice variability is another metric that augments the realism assessment of GCMs in the context of Arctic sea ice beyond the typical mean state and trend consistency and the assessment of sea ice sensitivity (Swart et al. 2015; Rosenblum and Eisenman 2017; Winton 2011; Niederdrenk and Notz 2018; Davy and Outten 2020).

We showed that all models are able to simulate the seasonal cycle of interannual variability with peaks in the summer, except CSIRO MK3.6 which has by far the largest mean state biases (see Table 4), caused by aerosol issues (Uotila et al. 2013). We demonstrate that all modeled interannual variability is within observational uncertainty, except for CanESM2 in January and November, GFDL CM3 in August and November, and CSIRO MK3.6 in August-October for the pan-Arctic. Except for areas of low absolute variability in the central Arctic Ocean, there are no inconsistencies that are common across all six models we assessed. Spatially, we find the models underestimate interannual variability for most regions in March, and in September most models overestimate variability in the central Arctic. The marginal seas, which have high absolute variability, are generally realistically simulated, although our assessment is limited to where both models and observations have sea ice. No model simulated the spatial interannual variability in both March and September without inconsistencies, but most models simulated at least one of the two months realistically. CESM1 and GFDL ESM2M simulate September spatial variability very well, with very few areas of inconsistency, including the highly variable shelf seas. In March, MPI ESM1 performs best, with only the Siberian coast displaying too high variability.

In summary, in this first direct comparison of interannual variability between observations and models, we have shown that estimates of interannual variability from models are largely consistent with observations. However, model skill varies by month and region, highlighting that the best model to use for a study varies based on the context. To be able to assess the impact of the

full range of internal variability, including the low frequency variability (McKinnon and Deser 2018), first requires an improved understanding of the drivers of low-frequency variability on Arctic sea ice. Generally, the fact that the simulated interannual variability of most CMIP5 large ensembles agrees quite well with historical observations, especially in September, increases trust in the internal variability uncertainty of Arctic sea ice projections.

Acknowledgments. This work was supported by the National Science Foundation under grant number 1847398. We would also like to acknowledge high-performance computing on Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

Data availability statement. All code required to replicate this study has been made open-access via Zenodo at https://www.doi.org/10.5281/zenodo.6687725. All data used in the analysis is already freely available. The CLIVAR Large Ensemble Archive can be obtained from the NCAR Climate Data Gateway (https://www.earthsystemgrid.org/dataset/ucar.cgd.ccsm4.CLIVAR_LE.html). The NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration (Version 4) is available from https://doi.org/10.7265/efmz-2t65. The Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST) is available from https://www.metoffice.gov.uk/hadobs/hadisst/index.html.

Additionally, the summary statistics from the resampled synthetic ensemble (μ_{LE}, σ_{LE}, μ_{mem}, σ_{mem}, μ_{obs} and σ_{obs}), can be accessed from the Arctic Data Center at https://doi.org/10.18739/A2H98ZF3T.

References

Bonan, D. B., F. Lehner, and M. M. Holland, 2021: Partitioning uncertainty in projections of Arctic sea ice. *Environmental Research Letters*, **16 (4)**, https://doi.org/10.1088/1748-9326/abe0ec.

Brennan, M. K., and G. J. Hakim, 2022: Reconstructing Arctic Sea Ice over the Common Era Using Data Assimilation. *Journal of Climate*, **35 (4)**, 1231–1247, https://doi.org/10.1175/JCLI-D-21-0099.1.

Brennan, M. K., G. J. Hakim, and E. Blanchard-Wrigglesworth, 2020: Arctic Sea-Ice Variability During the Instrumental Era. *Geophysical Research Letters*, **47 (7)**, https://doi.org/10.1029/2019GL086843.

Cavalieri, D. J., P. Gloersen, and W. J. Campbell, 1984: Determination of sea ice parameters with the Nimbus 7 SMMR. *Journal of Geophysical Research*, **89 (D4)**, 5355–5369, https://doi.org/10.1029/JD089iD04p05355.

Chan, D., A. Cobb, L. R. Zeppetello, D. S. Battisti, and P. Huybers, 2020: Summertime Temperature Variability Increases With Local Warming in Midlatitude Regions. *Geophysical Research Letters*, **47 (13)**, 1–7, https://doi.org/10.1029/2020GL087624.

Chemke, R., and L. M. Polvani, 2020: Using Multiple Large Ensembles to Elucidate the Discrepancy Between the 1979–2019 Modeled and Observed Antarctic Sea Ice Trends. *Geophysical Research Letters*, **47 (15)**, https://doi.org/10.1029/2020GL088339.

Chen, J., and Coauthors, 2020: Changes in sea ice and future accessibility along the Arctic Northeast Passage. *Global and Planetary Change*, **195**, 103 319, https://doi.org/10.1016/j.gloplacha.2020.103319.

Christensen, M., and A. E. Nilsson, 2017: Arctic sea ice and the communication of climate change. *Popular Communication*, **15 (4)**, 249–268, https://doi.org/10.1080/15405702.2017.1376064.

Comiso, J. C., 1986: Characteristics of Arctic winter sea ice from satellite multispectral microwave observations. *Journal of Geophysical Research*, **91 (C1)**, 975–994, https://doi.org/10.1029/JC091iC01p00975.

Comiso, J. C., W. N. Meier, and R. Gersten, 2017: Variability and trends in the Arctic Sea ice cover: Results from different techniques. *Journal of Geophysical Research: Oceans*, **122 (8)**, 6883–6900, https://doi.org/10.1002/2017JC012768.

Dai, A., and C. E. Bloecker, 2019: Impacts of internal variability on temperature and precipitation trends in large ensemble simulations by two climate models. *Climate Dynamics*, **52**, 289–306, https://doi.org/10.1007/s00382-018-4132-4.

Davy, R., and S. Outten, 2020: The arctic surface climate in CMIP6: Status and developments since CMIP5. *Journal of Climate*, **33 (18)**, 8047–8068, https://doi.org/10.1175/JCLI-D-19-0990.1, 1912.11654.

Deser, C., and Coauthors, 2020a: Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, **10 (4)**, 277–286, https://doi.org/10.1038/s41558-020-0731-2.

Deser, C., and Coauthors, 2020b: Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, **10 (4)**, 277–286, https://doi.org/10.1038/s41558-020-0731-2.

Ding, Q., and Coauthors, 2017: Influence of high-latitude atmospheric circulation changes on summertime Arctic sea ice. *Nature Climate Change*, **7 (4)**, 289–295, https://doi.org/10.1038/nclimate3241.

Ding, Q., and Coauthors, 2019: Fingerprints of internal drivers of Arctic sea ice loss in observations and model simulations. *Nature Geoscience*, **12 (1)**, 28–33, https://doi.org/10.1038/s41561-018-0256-8.

Dörr, J., M. Årthun, T. Eldevik, and E. Madonna, 2021: Mechanisms of regional winter sea-ice variability in a warming arctic. *Journal of Climate*, **34 (21)**, 8635–8653, https://doi.org/10.1175/JCLI-D-21-0149.1.

Elsworth, G. W., N. S. Lovenduski, and K. A. McKinnon, 2021: Alternate History: A Synthetic Ensemble of Ocean Chlorophyll Concentrations. *Global Biogeochemical Cycles*, **35 (9)**, 1–13, https://doi.org/10.1029/2020GB006924.

England, M., A. Jahn, and L. Polvani, 2019: Nonuniform contribution of internal variability to recent Arctic sea ice loss. *Journal of Climate*, **32 (13)**, 4039–4053, https://doi.org/10.1175/JCLI-D-18-0864.1.

England, M. R., 2021: Are Multi-Decadal Fluctuations in Arctic and Antarctic Surface Temperatures a Forced Response to Anthropogenic Emissions or Part of Internal Climate Variability? *Geophysical Research Letters*, **48 (6)**, e2020GL090 631, https://doi.org/10.1029/2020GL090631.

Frankcombe, L. M., M. H. England, J. B. Kajtar, M. E. Mann, and B. A. Steinman, 2018: On the choice of ensemble mean for estimating the forced signal in the presence of internal variability. *Journal of Climate*, **31 (14)**, 5681–5693, https://doi.org/10.1175/JCLI-D-17-0662.1.

Goosse, H., O. Arzel, C. M. Bitz, A. De Montety, and M. Vancoppenolle, 2009: Increased variability of the Arctic summer ice extent in a warmer climate. *Geophysical Research Letters*, **36 (23)**, 1–5, https://doi.org/10.1029/2009GL040546.

Hu, K., G. Huang, and S. P. Xie, 2019: Assessing the internal variability in multi-decadal trends of summer surface air temperature over East Asia with a large ensemble of GCM simulations. *Climate Dynamics*, **52 (9-10)**, 6229–6242, https://doi.org/10.1007/s00382-018-4503-x.

Jahn, A., 2018: Reduced probability of ice-free summers for 1.5 °c compared to 2 °c warming. *Nature Climate Change*, **8 (5)**, 409–413, https://doi.org/10.1038/s41558-018-0127-8.

Jahn, A., J. E. Kay, M. M. Holland, and D. M. Hall, 2016: How predictable is the timing of a summer ice-free Arctic? *Geophysical Research Letters*, **43 (17)**, 9113–9120, https://doi.org/10.1002/2016GL070067.

Jeffrey, S., L. Rotstayn, M. Collier, S. Dravitzki, C. Hamalainen, C. Moeseneder, K. Wong, and J. Syktus, 2013: Australia's CMIP5 submission using the CSIRO-Mk3.6 model. *Australian Meteorological and Oceanographic Journal*, **63 (1)**, 1–13, https://doi.org/10.22499/2.6301.001.

Kay, J. E., M. M. Holland, and A. Jahn, 2011: Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world. *Geophysical Research Letters*, **38 (15)**, 2–7, https://doi.org/10.1029/2011GL048008.

Kay, J. E., and Coauthors, 2015: The community earth system model (CESM) large ensemble project : A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, **96 (8)**, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1.

Kern, S., T. Lavergne, D. Notz, L. Toudal Pedersen, R. Tage Tonboe, R. Saldo, and A. Macdonald Sørensen, 2019: Satellite passive microwave sea-ice concentration data set intercomparison: Closed ice and ship-based observations. *Cryosphere*, **13 (12)**, 3261–3307, https://doi.org/10.5194/TC-13-3261-2019.

Kirchmeier-Young, M. C., F. W. Zwiers, and N. P. Gillett, 2017: Attribution of extreme events in Arctic Sea ice extent. *Journal of Climate*, **30 (2)**, 553–571, https://doi.org/10.1175/JCLI-D-16-0412.1.

Kovacs, K. M., C. Lydersen, J. E. Overland, and S. E. Moore, 2011: Impacts of changing sea-ice conditions on Arctic marine mammals. *Marine Biodiversity*, **41 (1)**, 181–194, https://doi.org/10.1007/s12526-010-0061-0.

Lehner, F., C. Deser, N. Maher, J. Marotzke, E. M. Fischer, L. Brunner, R. Knutti, and E. Hawkins, 2020: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth Syst. Dynam*, **11**, 491–508, https://doi.org/10.5194/esd-11-491-2020.

Li, D., R. Zhang, and T. R. Knutson, 2017: On the discrepancy between observed and CMIP5 multi-model simulated Barents Sea winter sea ice decline. *Nature Communications*, https://doi.org/10.1038/ncomms14991.

Maher, N., F. Lehner, and J. Marotzke, 2020: Quantifying the role of internal variability in the temperature we expect to observe in the coming decades. *Environmental Research Letters*, **15 (5)**, 54 014, https://doi.org/10.1088/1748-9326/ab7d02.

Maher, N., and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability. *Journal of Advances in Modeling Earth Systems*, **11 (7)**, 2050–2069, https://doi.org/10.1029/2019MS001639.

Massonnet, F., M. Vancoppenolle, H. Goosse, D. Docquier, T. Fichefet, and E. Blanchard-Wrigglesworth, 2018: Arctic sea-ice change tied to its mean state through thermodynamic processes. *Nature Climate Change*, **8 (7)**, 599–603, https://doi.org/10.1038/s41558-018-0204-z.

McKinnon, K. A., and C. Deser, 2018: Internal variability and regional climate trends in an observational large ensemble. *Journal of Climate*, **31 (17)**, 6783–6802, https://doi.org/10.1175/JCLI-D-17-0901.1.

McKinnon, K. A., and C. Deser, 2021: The Inherent Uncertainty of Precipitation Variability, Trends, and Extremes due to Internal Variability, with Implications for Western U.S. Water Resources. *Journal of Climate*, **34 (24)**, 9605–9622, https://doi.org/10.1175/JCLI-D-21-0251.1.

McKinnon, K. A., A. Poppick, E. Dunn-Sigouin, and C. Deser, 2017: An "observational large ensemble" to compare observed and modeled temperature trend uncertainty due to internal variability. *Journal of Climate*, **30 (19)**, 7585–7598, https://doi.org/10.1175/JCLI-D-16-0905.1.

Meier, W. N., F. Fetterer, A. Windnagel, and J. Stewart, 2021: NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, Version 4. Tech. rep., NSIDC: National Snow and Ice Data Center, Boulder, Colorado USA. https://doi.org/10.7265/efmz-2t65.

Milinski, S., N. Maher, and D. Olonscheck, 2020: How large does a large ensemble need to be? *Earth Syst. Dynam*, **11**, 885–901, https://doi.org/10.5194/esd-11-885-2020.

Mioduszewski, J. R., S. Vavrus, M. Wang, M. Holland, and L. Landrum, 2019: Past and future interannual variability in Arctic sea ice in coupled climate models. *Cryosphere*, **13 (1)**, 113–124, https://doi.org/10.5194/tc-13-113-2019.

Niederdrenk, A. L., and D. Notz, 2018: Arctic Sea Ice in a 1.5°C Warmer World. *Geophysical Research Letters*, **45 (4)**, 1963–1971, https://doi.org/10.1002/2017GL076159.

Notz, D., 2014: Sea-ice extent and its trend provide limited metrics of model performance. *Cryosphere*, **8 (1)**, 229–243, https://doi.org/10.5194/tc-8-229-2014.

Notz, D., 2015: How well must climate models agree with observations? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **373 (2052)**, https://doi.org/10.1098/rsta.2014.0164.

672 Notz, D., and J. Marotzke, 2012: Observations reveal external driver for Arctic sea-ice retreat.
673 *Geophysical Research Letters*, **39 (8)**, https://doi.org/10.1029/2012GL051094.

674 Notz, D., and J. Stroeve, 2018: The Trajectory Towards a Seasonally Ice-Free Arctic Ocean.
675 Springer, 407–416 pp., https://doi.org/10.1007/s40641-018-0113-2.

676 Olonscheck, D., T. Mauritsen, and D. Notz, 2019: Arctic sea-ice variability is primarily driven
677 by atmospheric temperature fluctuations. *Nature Geoscience*, **12 (6)**, 430–434, https://doi.org/
678 10.1038/s41561-019-0363-1.

679 Olonscheck, D., and D. Notz, 2017: Consistently estimating internal climate variability from
680 climate model simulations. *Journal of Climate*, **30 (23)**, 9555–9573, https://doi.org/10.1175/
681 JCLI-D-16-0428.1.

682 Onarheim, I. H., T. Eldevik, L. H. Smedsrud, and J. C. Stroeve, 2018: Seasonal and regional
683 manifestation of Arctic sea ice loss. *Journal of Climate*, **31 (12)**, 4917–4932, https://doi.org/
684 10.1175/JCLI-D-17-0427.1.

685 Petrick, S., K. Riemann-Campe, S. Hoog, C. Growitsch, H. Schwind, R. Gerdes, and K. Rehdanz,
686 2017: Climate change, future Arctic Sea ice, and the competitiveness of European Arctic
687 offshore oil and gas production on world markets. *Ambio*, **46**, 410–422, https://doi.org/10.1007/
688 s13280-017-0957-z.

689 Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent,
690 and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine
691 air temperature since the late nineteenth century. *Journal of Geophysical Research*, **108 (D14)**,
692 4407, https://doi.org/10.1029/2002JD002670.

693 Roberts, J., and T. D. Roberts, 1978: Use of the Butterworth low-pass filter for oceanographic
694 data. *Journal of Geophysical Research: Oceans*, **83 (C11)**, 5510–5514, https://doi.org/https:
695 //doi.org/10.1029/JC083iC11p05510.

696 Rodgers, K. B., J. Lin, and T. L. Frölicher, 2015: Emergence of multiple ocean ecosystem drivers
697 in a large ensemble suite with an Earth system model. *Biogeosciences*, **12 (11)**, 3301–3320,
698 https://doi.org/10.5194/bg-12-3301-2015.

Rosenblum, E., and I. Eisenman, 2017: Sea ice trends in climate models only accurate in runs with biased global warming. *Journal of Climate*, **30 (16)**, 6265–6278, https://doi.org/10.1175/JCLI-D-16-0455.1, 1606.08519.

Santer, B. D., and Coauthors, 2008: Consistency of modelled and observed temperature trends in the tropical troposphere. *International Journal of Climatology*, **28 (13)**, 1703–1722, https://doi.org/10.1002/JOC.1756.

Schweiger, A., and Coauthors, 2011: Uncertainty in modeled Arctic sea ice volume. *J. Geophys. Res*, **116**, 0–06, https://doi.org/10.1029/2011JC007084.

SIMIP Community, 2020: Arctic Sea Ice in CMIP6. *Geophysical Research Letters*, **47 (10)**, https://doi.org/10.1029/2019GL086749.

Stroeve, J., M. M. Holland, W. Meier, T. Scambos, and M. Serreze, 2007: Arctic sea ice decline: Faster than forecast. *Geophysical Research Letters*, **34 (9)**, 9501, https://doi.org/10.1029/2007GL029703.

Stroeve, J. C., V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. N. Meier, 2012: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophysical Research Letters*, **39 (16)**, n/a–n/a, https://doi.org/10.1029/2012GL052676.

Sun, L., M. Alexander, and C. Deser, 2018: Evolution of the Global Coupled Climate Response to Arctic Sea Ice Loss during 1990–2090 and Its Contribution to Climate Change. *Journal of Climate*, **31 (19)**, 7823–7843, https://doi.org/10.1175/JCLI-D-18-0134.1.

Swart, N. C., J. C. Fyfe, E. Hawkins, J. E. Kay, and A. Jahn, 2015: Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change*, **5 (2)**, 86–89, https://doi.org/10.1038/nclimate2483.

Uotila, P., S. O'farrell, S. J. Marsland, and D. Bi, 2013: The sea-ice performance of the Australian climate models participating in the CMIP5. *Australian Meteorological and Oceanographic Journal*, **63**, 121–143, https://doi.org/10.22499/2.6301.008.

Winton, M., 2011: Do climate models underestimate the sensitivity of Northern Hemisphere sea ice cover? *Journal of Climate*, **24 (15)**, 3924–3934, https://doi.org/10.1175/2011JCLI4146.1.

Zhang, R., and J. M. Wallace, 2015: Mechanisms for low-frequency variability of summer Arctic sea ice extent. **112 (15)**, https://doi.org/10.1073/pnas.1422296112.

Supporting Information for "Modeled Interannual Variability of Arctic Sea Ice Cover is Within Observational Uncertainty"

Christopher Wyburn-Powell[a] , Alexandra Jahn[a] , Mark England[b]

[a] *Department of Atmospheric and Oceanic Sciences, and Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, Colorado*

[b] *Department of Earth and Planetary Science, University of California, Santa Cruz, California*

Corresponding author: C Wyburn-Powell, chwy8767@colorado.edu

TABLE S1. **Replacement months for datasets with missing data** for NSIDC or with discontinuities as for HadISST1.

Dataset	Discontinuity Month	Replacement Month
NSIDC CDR,BT,NT	1984-07	1985-07
NSIDC CDR,BT,NT	1987-12	1988-12
NSIDC CDR,BT,NT	1988-01	1989-01
HadISST1	2009-03	2007-03
HadISST1	2009-04	2008-04

Month	CanESM2		CESM1		CSIRO MK36		GFDL CM3		GFDL ESM2M		MPI ESM1		Model Mean	Model Mean Bias	NSIDC CDR
Jan	11.97	-1.42	13.53	0.13	14.88	1.49	11.90	-1.49	13.28	-0.11	12.01	-1.39	**12.93**	-0.46	**13.39**
Feb	13.36	-0.95	14.25	-0.06	15.69	1.38	12.91	-1.41	14.73	0.42	12.89	-1.42	**13.97**	-0.34	**14.31**
Mar	13.78	-0.65	14.39	-0.03	15.93	1.50	13.21	-1.22	15.06	0.64	13.11	-1.31	**14.25**	-0.18	**14.43**
Apr	12.99	-0.57	13.83	0.27	15.61	2.05	12.71	-0.85	14.25	0.69	12.56	-1.00	**13.66**	0.10	**13.56**
May	11.35	-0.64	12.30	0.30	14.88	2.88	11.41	-0.58	12.09	0.09	11.18	-0.81	**12.20**	0.21	**11.99**
Jun	8.69	-1.39	10.32	0.24	13.69	3.61	9.37	-0.71	9.53	-0.55	8.71	-1.37	**10.05**	-0.03	**10.08**
Jul	5.10	-2.62	7.93	0.21	11.50	3.78	6.57	-1.15	6.91	-0.81	5.84	-1.88	**7.31**	-0.41	**7.72**
Aug	2.75	-2.96	5.68	-0.03	10.11	4.41	4.44	-1.26	4.83	-0.88	4.01	-1.70	**5.30**	-0.40	**5.70**
Sep	2.44	-2.84	5.42	0.14	9.64	4.37	3.87	-1.40	4.63	-0.65	3.73	-1.55	**4.96**	-0.32	**5.28**
Oct	4.10	-3.03	7.25	0.12	9.77	2.64	5.37	-1.76	6.72	-0.41	5.44	-1.69	**6.44**	-0.69	**7.13**
Nov	7.04	-2.55	9.64	0.05	11.21	1.62	7.96	-1.63	9.47	-0.12	8.13	-1.45	**8.91**	-0.68	**9.59**
Dec	9.91	-1.87	11.82	0.04	13.32	1.54	10.18	-1.60	11.32	-0.46	10.26	-1.52	**11.13**	-0.65	**11.78**

TABLE S2. **Mean SIA for the period 1979-2020 for all models and observations.** SIA in 10^6 km^2, biases relative to CDR observations.

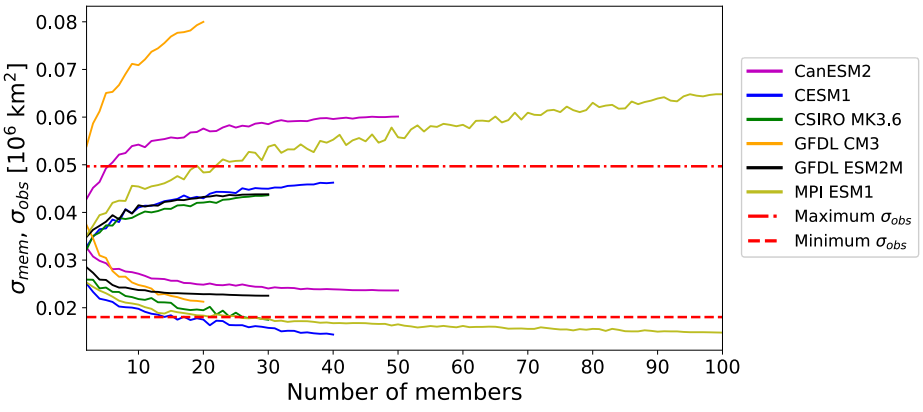


FIG. S1. **Effect of ensemble size on maximum and minimum resampled standard deviations.** The maximum and minimum ensemble member for σ_{mem} when subsampled from 2-100 members 100 times is compared with the minimum and maximum values of σ_{obs} for December.

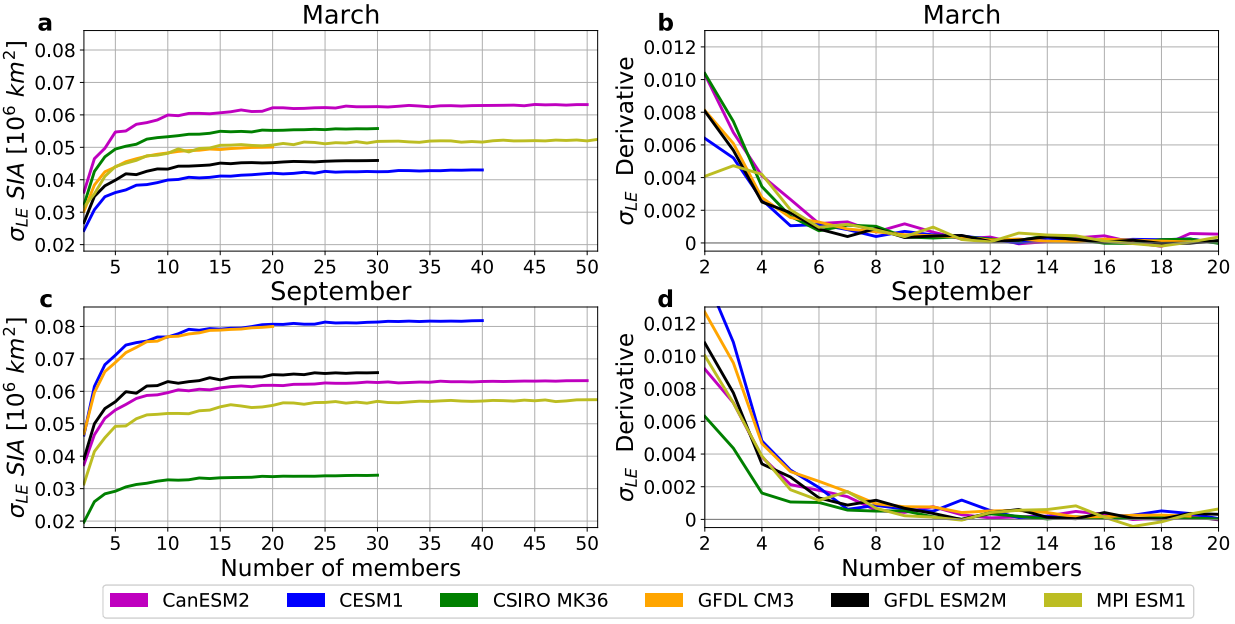


FIG. S2. **Influence of ensemble size on standard deviation of pan-Arctic SIA**. Standard deviation with respect to time for a number of subsampled members between 2 and the lesser of 51 and the full ensemble for the time period 1979-2020. (a, c): average standard deviation across members (σ_{LE}) for 1000 bootstrapped members of a given number, for March (a) and September (c). (b, d): the derivative of the average standard deviation per member shown in (a) and (c).

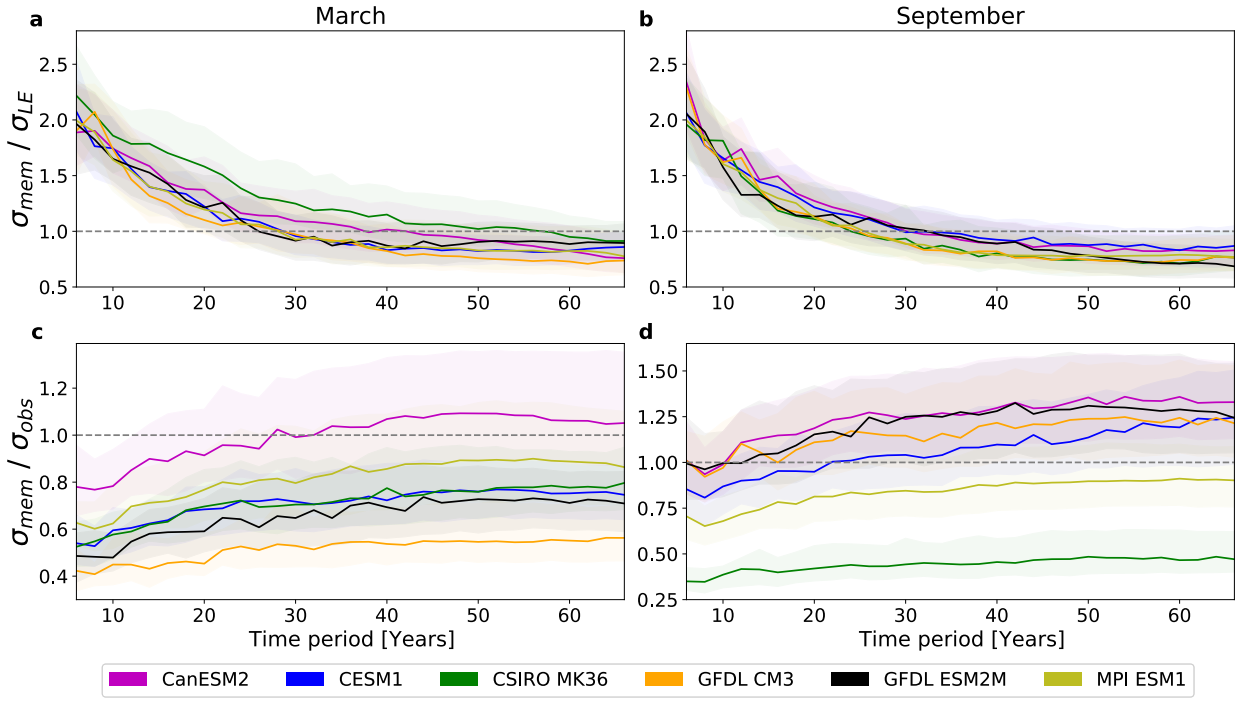


FIG. S3. **Influence of length of time period on standard deviation of pan-Arctic SIA**. Standard deviation with respect to time for time periods between 6 and 66 years in SIA derived from anomalies relative to a 2 year lowpass Butterworth filter with 10 random start times for σ_{mem} and 1000 for σ_{LE} and σ_{obs}. Thick lines show the median ensemble member, shading shows ± 1 standard deviation. (a, b): the ratio of standard deviation across resamplings (σ_{mem}) to standard deviation across members (σ_{LE}) over a subset of the time periods for March (a) and September (b). (c, d): the ratio of standard deviation across resamplings (σ_{mem}) to standard deviation across resampled observations (σ_{obs}) in the HadISST1 dataset in March (c) and September (d).

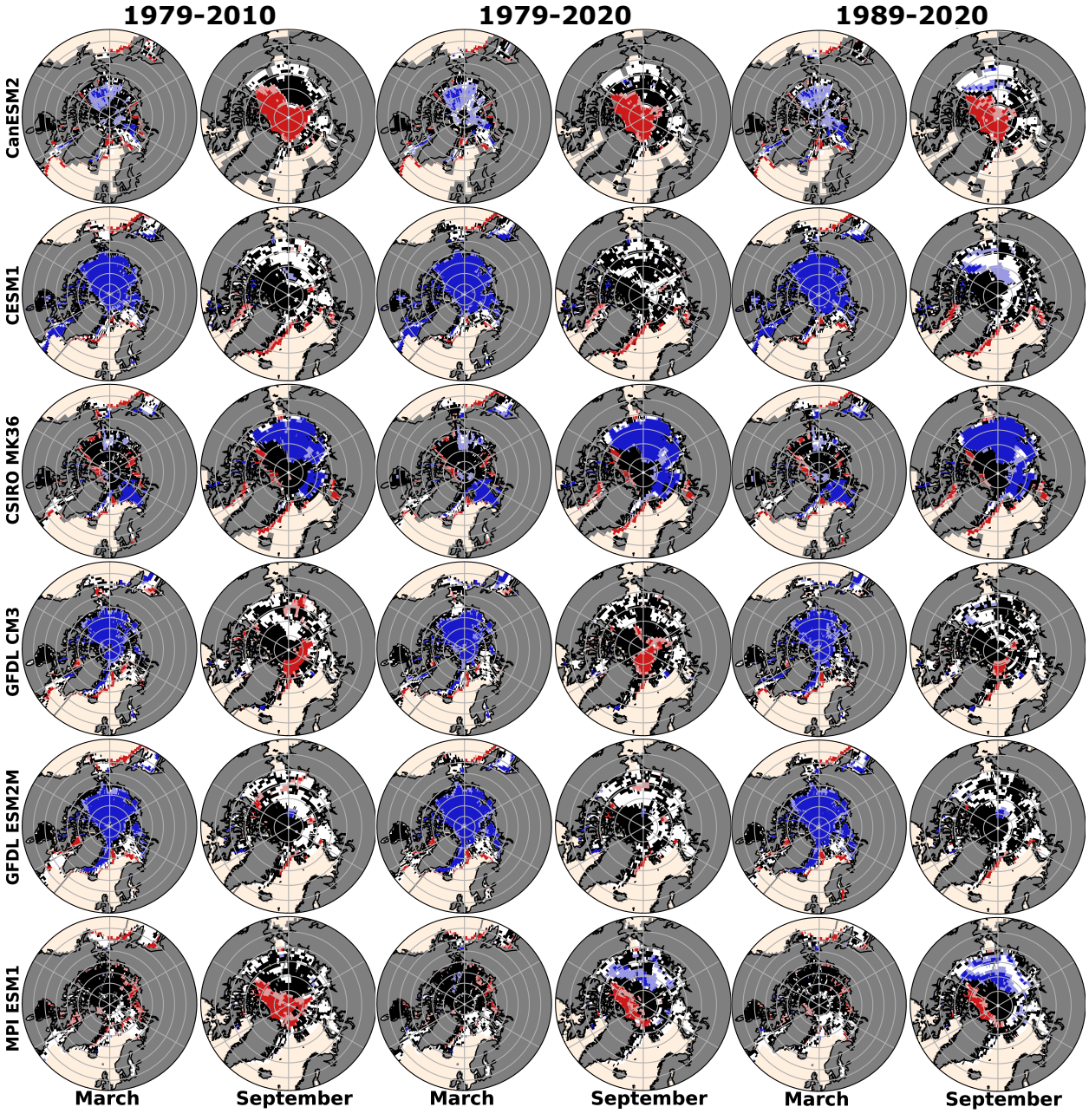


FIG. S4. **Influence of time period on spatial consistency**. SIC is detrended using a linear trend. Same as 'All σ_{obs} and μ_{obs}' columns in Figure 8 for '1979-2020' columns. Model and observational data for 1979-2010 and 1989-2020 are shown in the columns labeled accordingly.

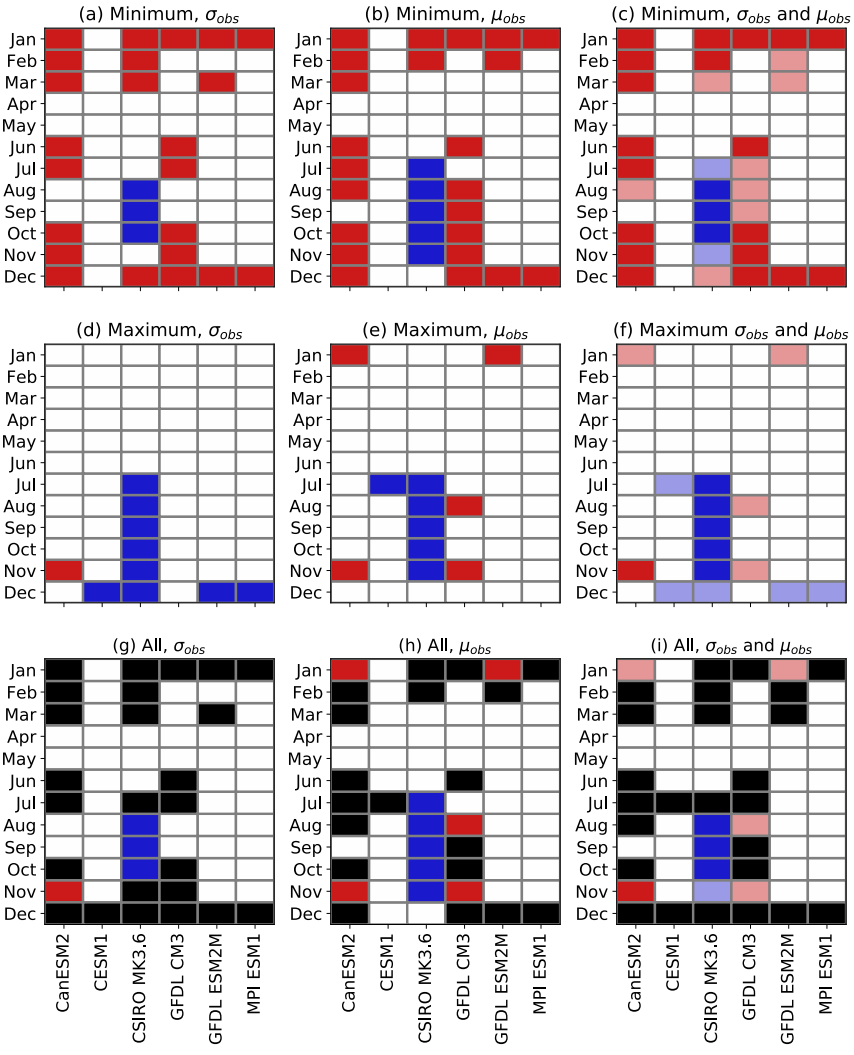


FIG. S5. **Consistency between models and observations in pan-Arctic SIA, subsampled to 20 members**. Same as Figure 7, except now subsampled 1000 times to 20 members.

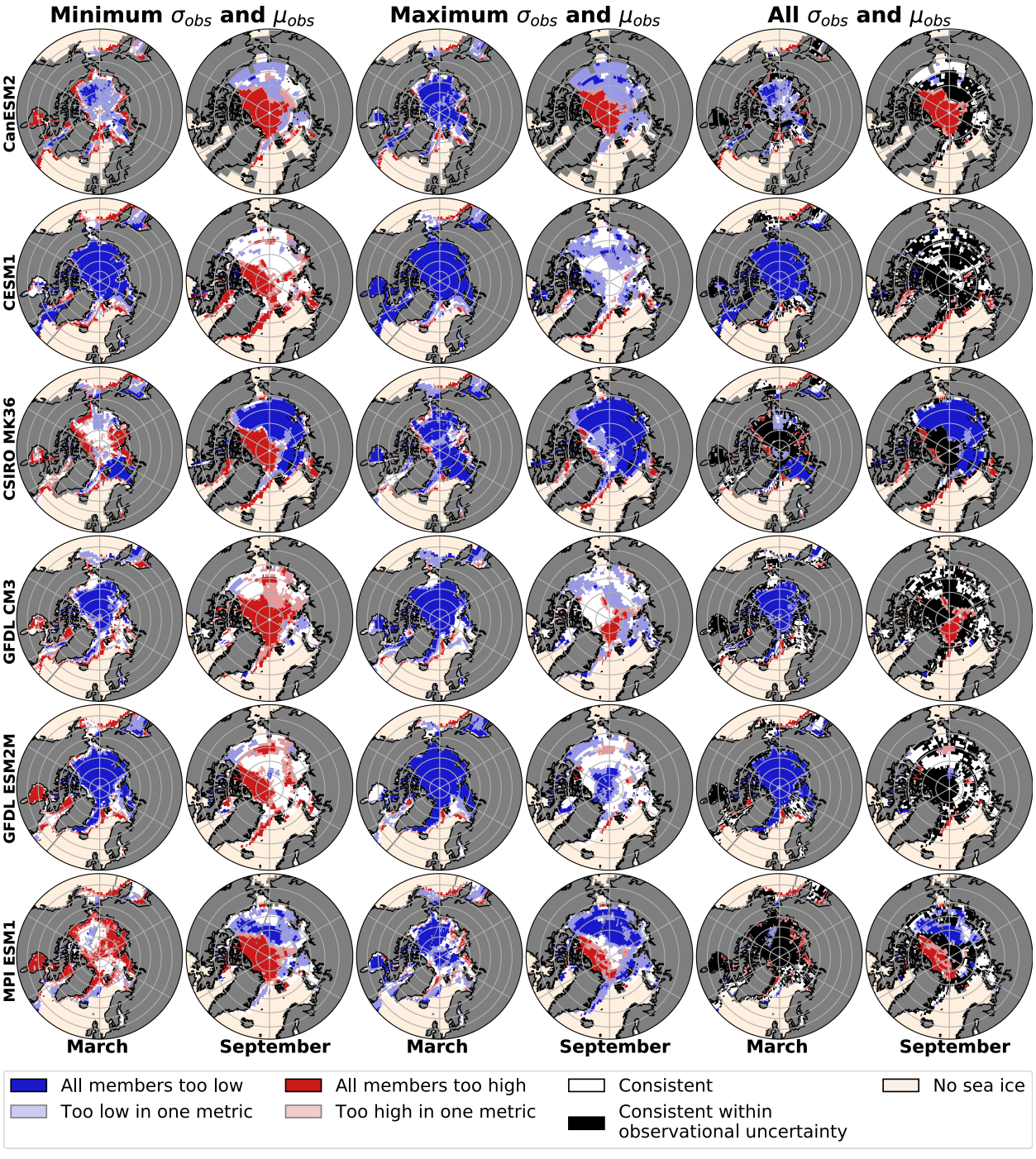


FIG. S6. **Spatial consistency of internal variability between large ensemble members, subsampled to 20 members**. Same as Figure 8 except now members are subsampled 1000 times to a size of 20 members.

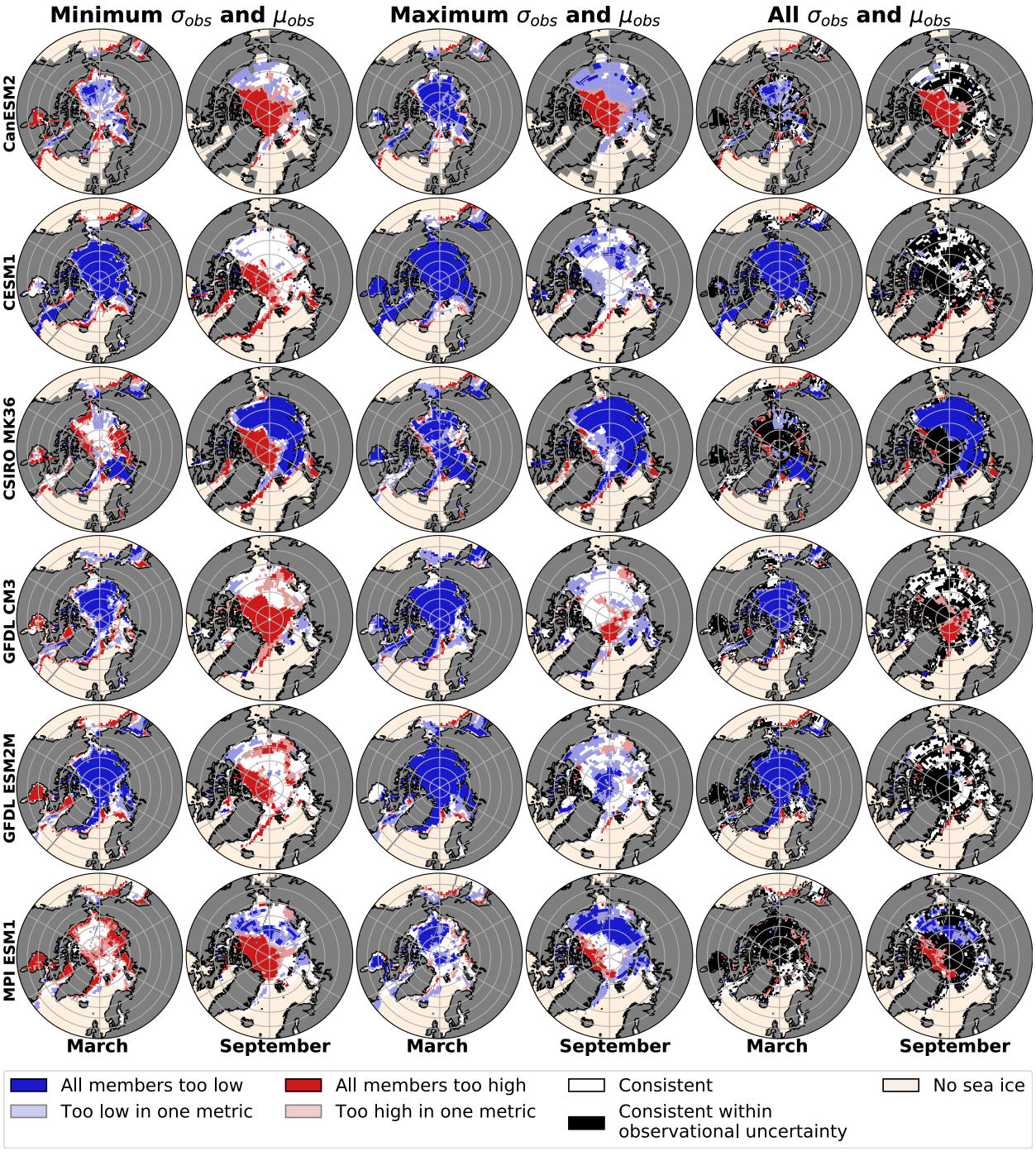


FIG. S7. **Spatial consistency of internal variability between large ensemble members and observations using lowpass filtered data**. Same as Figure 8 except now SICs are detrended using a 2 year lowpass filter before resampling.

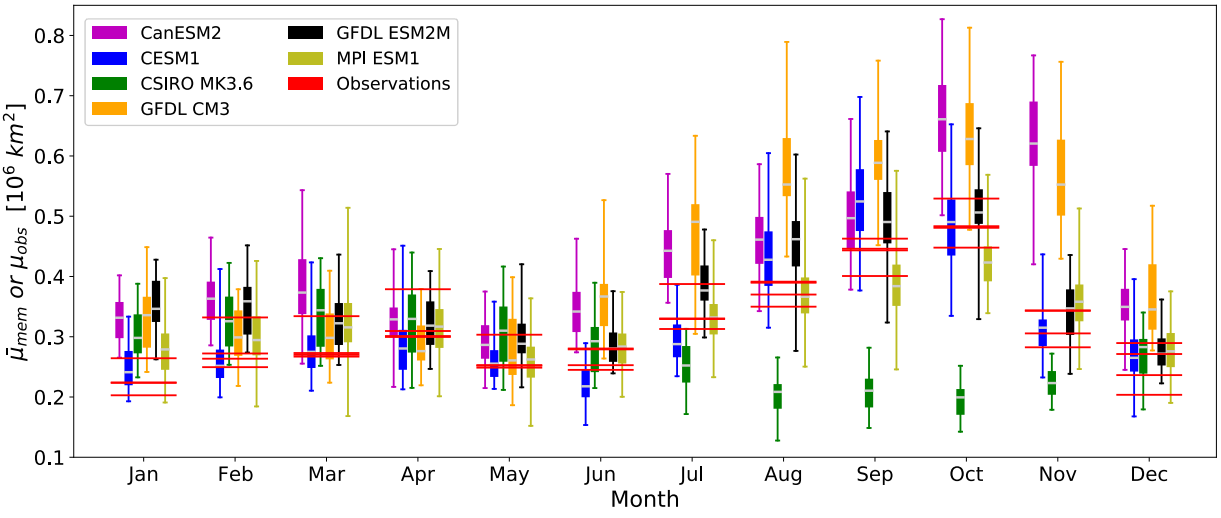


FIG. S8. **Seasonality of average resampled variability in members and observations for pan-Arctic sea ice area.** Same as Figure 4, except now for μ instead of σ.

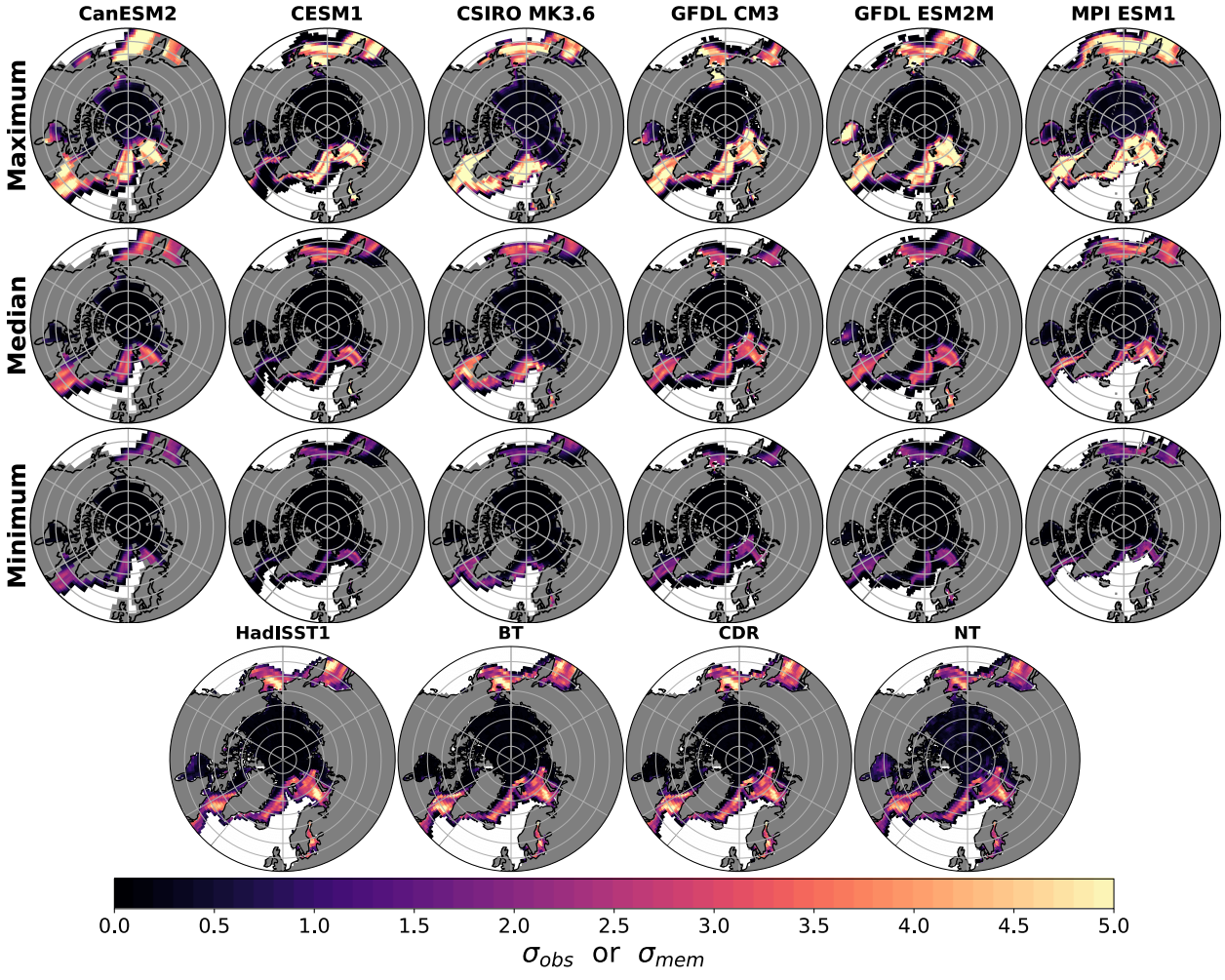


FIG. S9. **Resampled modeled and observed variability of March SIC.** Same as Figure 6, but now for March.

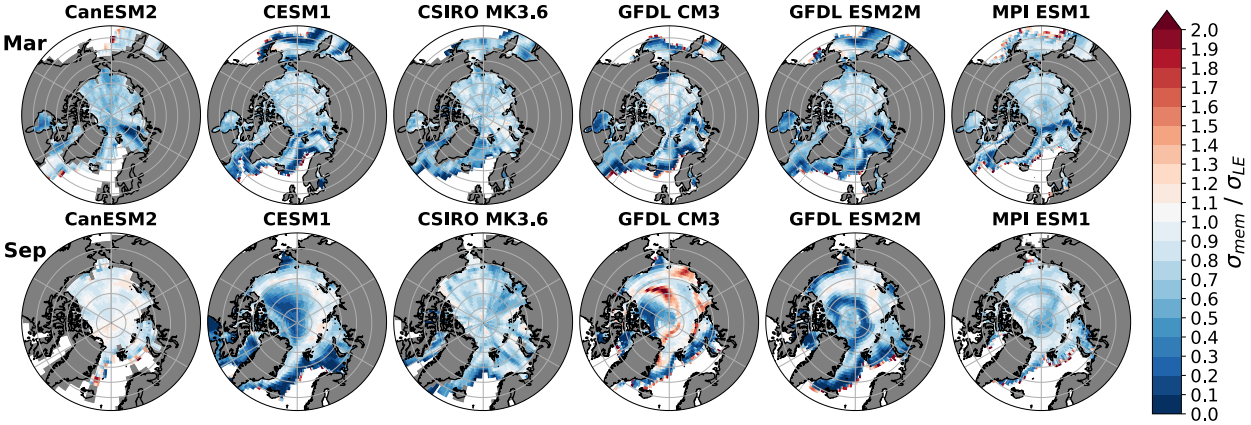


FIG. S10. **Spatial distribution of the ratio of internal variability in large ensembles and resampled members.** The ratio $\bar{\sigma}_{mem}$ to σ_{LE} for March and September in all models represents the proportion of large ensemble variability captured via the resampling technique.

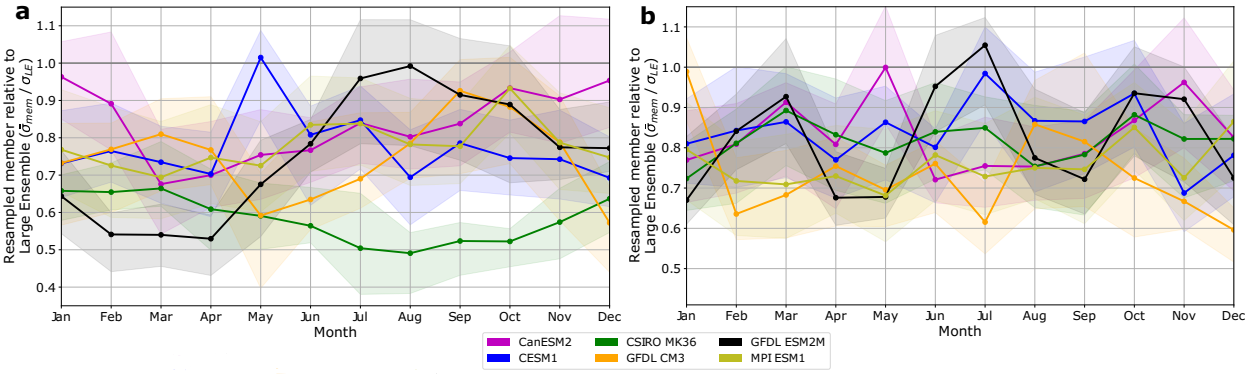


FIG. S11. **Seasonality of the ratio of internal variability across SMILEs and interannual variability of resampled members for pan-Arctic sea ice area, using ensemble mean detrended data and lowpass filtered datta.** Same as Figure 9, except now SIA anomalies were detrended relative to the linear trend of the ensemble mean for (a) and relative to a 2 year lowpass filter for (b).

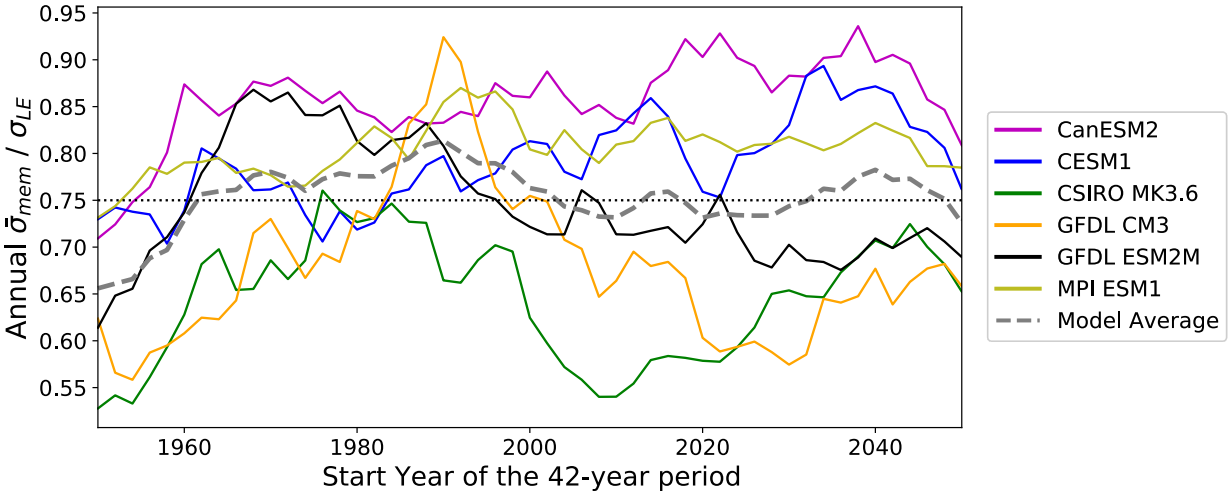


FIG. S12. **Effect of starting year on $\bar{\sigma}_{mem}$ / σ_{LE} ratio for a 42-year time period.** The ratio of $\bar{\sigma}_{mem}$ to σ_{LE} is calculated for 42-year time periods between 1950-1991 and 2050-2091 for the annual mean. $\bar{\sigma}_{mem}$ is calculated from 1000 resamplings.