# The Value of Initialization on Decadal Timescales: State-Dependent Predictability in the CESM Decadal Prediction Large Ensemble

H. M. CHRISTENSEN

*Atmospheric Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom*

J. BERNER AND S. YEAGER

*National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 26 July 2019, in final form 28 May 2020)

ABSTRACT

Information in decadal climate prediction arises from a well-initialized ocean state and from the predicted response to an external forcing. The length of time over which the initial conditions benefit the decadal forecast depends on the start date of the forecast. We characterize this state-dependent predictability for decadal forecasts of upper ocean heat content in the Community Earth System Model. We find regionally dependent initial condition predictability, with extended predictability generally observed in the extratropics. We also detect state-dependent predictability, with the year of loss of information from the initialization varying between start dates. The decadal forecasts in the North Atlantic show substantial information from the initial conditions beyond the 10-yr forecast window, and a high degree of state-dependent predictability. We find some evidence for state-dependent predictability in the ensemble spread in this region, similar to that seen in weather and subseasonal-to-seasonal forecasts. For some start dates, an increase of information with lead time is observed, for which the initialized forecasts predict a growing phase of the Atlantic multidecadal oscillation. Finally we consider the information in the forecast from the initial conditions relative to the forced response, and quantify the crossover time scale after which the forcing provides more information. We demonstrate that the climate change signal projects onto different patterns than the signal from the initial conditions. This means that even after the crossover time scale has been reached in a basin-averaged sense, the benefits of initialization can be felt locally on longer time scales.

## 1. Introduction

Information in long-term climate projections arises from a good estimate of future greenhouse gas emissions, coupled with a climate model that responds well to such emissions (IPCC 2014). However on shorter climate time scales, information can also arise from the climate model's initial conditions, in particular from a well-initialized ocean state. Decadal climate forecasts capitalize on this potential, predicting the climate system on multiannual to decadal time scales when information from the initialization is understood to benefit the prediction (Yeager and Robson 2017; Smith et al. 2019; Kushnir et al. 2019).

Recent years have seen significant progress in the field of decadal climate forecasting (Merryfield et al. 2020), with skillful multimodel forecasts possible for a range of climate phenomena (Smith et al. 2019). In particular, multimodel ensembles can produce skillful forecasts of North Atlantic variability (Smith et al. 2020, manuscript submitted to *Nature*), including atmospheric modes of variability such as blocking and the North Atlantic Oscillation (Athanasiadis et al. 2020). The World Climate Research Programme has highlighted the importance of moving toward operational decadal forecasting through its grand challenge on near-term climate prediction, although scientific and technical challenges remain (Kushnir et al. 2019). The coordinated multimodel Decadal Climate Prediction Project (DCPP), part of phase 6 of the Coupled Model Intercomparison Project (CMIP6), provides data to enable the community to work toward

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-19-0571.s1.

*Corresponding author*: Hannah M. Christensen, hannah.christensen@physics.ox.ac.uk

DOI: 10.1175/JCLI-D-19-0571.1

this grand challenge over the coming years (Boer et al. 2016).

Before assessing the predictive skill of a model compared to observations, it is helpful to assess a model's *potential predictability*. This can be defined as the ability of a model to predict itself (Boer et al. 2019), or the rate at which an initialized forecast becomes indistinguishable from an uninitialized forecast (Branstator and Teng 2010). Since modern climate models represent the climate system with considerable fidelity (Flato et al. 2013), potential predictability estimates can provide some indication of the true predictability of the climate system. In particular, potential predictability studies can offer insights as to the source of skill on different time scales. Decadal predictions are hybrid predictions of the first and the second kind (Lorenz 1975), concerning both the initial value problem and the response of the climate system to an external forcing. At longer lead times, the influence of the initial condition decays, while the forced response grows. The relative importance of these two sources of predictability changes over the course of the prediction, and a key concern is assessing for how long the initial conditions benefit the prediction (Collins et al. 2006), and at what point in time the forced response begins to dominate (Branstator and Teng 2010).

A number of studies have assessed these time scales. For example, Branstator and Teng (2010) use a long unforced control simulation and a pair of ensemble experiments, one with and one without anthropogenic forcing, to estimate these time scales for upper ocean heat content in different ocean basins, and find that the initial conditions dominate for approximately seven years. Branstator et al. (2012) develop statistical methods to estimate the initial value predictability from long control runs. Branstator and Teng (2012) apply these techniques to upper ocean heat content in the CMIP5 ensemble, and find the forced predictability begins to dominate initial value predictability at time scales of 6.5 and 8 years in the North Pacific and North Atlantic basins, respectively. Corti et al. (2015) compare four different climate models and find that the forced response dominates sea surface temperature (SST) predictions for lead times longer than one year, although in some regions extended initial condition predictability is present.

A limitation of past studies is the reliance on statistical techniques, or the small number of initial ocean states considered. However on weather and seasonal forecasting time scales, it is known that the predictability of the coupled atmosphere–ocean system is state dependent (Palmer 2006). It is therefore expected that the time scales over which the ocean initial conditions benefit decadal forecasts will also be state dependent. Furthermore, the rate of ensemble dispersion in shorter-range forecasts is commonly used as an indicator of the predictability of the state of the system (e.g., Leutbecher 2010; Weisheimer and Palmer 2014; Christensen et al. 2015). For certain start dates, the ensemble spread increases slowly indicating extended predictability, while the converse is true for other start dates (Slingo and Palmer 2011). The presence of state-dependent predictability has not been assessed in decadal forecasts, where often only the ensemble mean forecast is evaluated.

To robustly address the question of state-dependent predictability in decadal forecasts, we require a large set of initialized decadal ensemble predictions combined with an uninitialized control ensemble. In this study, we make use of the Community Earth System Model (CESM) Decadal Prediction Large Ensemble (DPLE; Yeager et al. 2018), combined with the uninitialized CESM Large Ensemble (LENS; Kay et al. 2015). With 62 start dates, the DPLE samples a wide range of initial states. LENS shares a codebase with DPLE, so it can be used as a control dataset. This pair of community datasets provide an unprecedented opportunity to revisit the question of initial condition versus forced response in decadal forecasts.

Many previous studies have focused on SST predictions, because oceanic influence on the atmosphere is almost entirely mediated through SST. However, subsurface fields can provide a source of predictability for SST (Yeager and Robson 2017) and ultimately for atmospheric fields, but are less affected by weather noise. For this reason, we focus our analysis on upper ocean heat content. Our study focuses on the potential predictability of decadal forecasts, as a first step toward assessing forecast skill compared to observations: if the decadal forecast is indistinguishable from a reference forecast, there can be no additional skill in the decadal forecast when compared to observations. Following Branstator and Teng (2010), we will use relative entropy to assess potential predictability in terms of the information content in the decadal forecast over a reference forecast. Relative entropy is sensitive to both information in the ensemble mean and spread, and so allows for a direct comparison of the two potential sources of information in the forecast. In section 2 we outline our methodology, and in section 3 we consider the limits of initial value predictability. In sections 4 and 5 we consider the balance between information from the initial conditions and from the forced response. In section 6 we discuss our results and highlight the main conclusions.

## 2. Methods

### a. Model data

This study will analyze a pair of existing model datasets produced using the NCAR CESM. First, the CESM

decadal prediction large ensemble (DPLE; Yeager et al. 2018), which consists of 40 ensemble member hindcasts, each of 122-month duration, initialized every 1 November between 1954 and 2015 (62 start dates in total). The DPLE will be compared to the CESM Large Ensemble (LENS; Kay et al. 2015), which is a 40-member uninitialized set of hindcasts.

The DPLE and LENS are produced using the same model and configuration, CESM 1.1, allowing for a clean comparison between the two datasets. The model includes fully coupled atmosphere, ocean, sea ice, and land components, all at a nominal resolution of 1°. The historical forcings (pre-2005) and projected forcings (2006 onward) are consistent between DPLE and LENS, and include greenhouse gases, short-lived gases, aerosols, and volcanic forcings. The projected forcings are those specified for CMIP5 representative concentration pathway (RCP) 8.5. The DPLE dataset is a contribution to the CMIP6 DCPP (Boer et al. 2016).

The DPLE ocean and sea ice components were initialized from CESM simulations in which the ocean and sea ice components of CESM were forced using an estimate of the observed atmospheric state and fluxes. Despite the lack of assimilation of sea ice or ocean observations, this approach has been shown to reproduce key aspects of the ocean and ice initial conditions well (Yeager and Danabasoglu 2014; Yeager et al. 2015; Danabasoglu et al. 2016). In contrast, the atmosphere and land components of CESM were initialized by selecting initial conditions from a single LENS ensemble member. These initial conditions contain the externally forced response, but no further information on atmospheric or land state. The DPLE ensemble is generated through the introduction of round-off sized perturbations to the atmospheric initial conditions. All model components were initialized using a full-field approach.

We will focus on forecasts of annually averaged ocean heat content of the upper 295 m (T295) since this is a leading reservoir of ocean memory on decadal time scales. Annual averages are taken from January to December, such that "lead year 1" corresponds to lead months 2–14. Prior to analysis, we remove any ocean data point that is covered by sea ice at any point in any of the LENS or DPLE ensemble members.

The full-field initialization puts the DPLE coupled model ocean in a state far from model climatology. The ocean adjustment toward the model attractor (e.g., to reach stable Atlantic meridional overturning circulation) can take hundreds of simulation years. This means that a drift is present in the DPLE, which is removed prior to any analysis. The raw DPLE data are corrected for drift by subtracting the lead time–dependent difference between the DPLE ensemble mean and the LENS ensemble mean for each verification year, averaged across all start years. This is the approach recommended by the CMIP6 DCPP (Boer et al. 2016). The end result is that, for each lead time, DPLE and LENS share the same climatology over the verification window. To illustrate the method, Fig. 1 shows the lead time–dependent drift between the DPLE and LENS. In most regions, the DPLE tends toward the LENS as lead time increases (i.e., a drift away from observations toward the model attractor). However, this adjustment need not manifest as a monotonic approach toward the LENS solution. In fact, in some regions in the North Atlantic (such as around 60°N), the DPLE is observed to drift *away* from the LENS.

### b. Initial value and forced datasets

Following Branstator and Teng (2010), we consider the information in the DPLE as arising from two sources, the initial conditions and the forced response, while the information in LENS arises purely from the forced response. To unpick the relative contributions from each source of information, we process the LENS and DPLE datasets to produce three new datasets.

To separate internal variability from the forced response, we must first remove the forced signal. There are many possible approaches to removing this signal. These include modeling the forced signal as a linear trend at each grid point (e.g., Sutton and Hodson 2005; Delworth et al. 2017), representing the forced climate response using the global mean (Trenberth and Shea 2006), or regressing local changes on the global mean to estimate a local forced response: see Frankignoul et al. (2017) for a succinct review of different methods. A more sophisticated approach that has become widely used in recent years (e.g., Qasmi et al. 2020) is the signal-to-noise maximizing EOF method proposed by Ting et al. (2009). These approaches are appropriate for removing the forced signal from an observed record, or from a single model simulation. However, since we have a large ensemble of forced simulations available, the natural approach is to take the mean of that ensemble as the forced signal.

Therefore we first calculate the mean over all LENS ensemble members as a function of year independently for every spatial location, and define this to be the *local climate change signal*. We subtract this local climate change signal from each of the LENS ensemble members to leave the climate anomalies, LENS*. This provides an estimate of CESM climate variability in a hypothetical unforced climate. For comparison, we also calculate the evolving unforced DPLE anomalies, DPLE*, which are calculated by subtracting the local climate change signal (LENS mean) from DPLE. Comparing the evolving DPLE* forecast with the (assumed stationary) LENS* distribution gives an estimate of the information in the DPLE forecast arising from the initialization. This information is expected to decay with time, as the chaotic nature of the Earth-system results in the DPLE* forecast

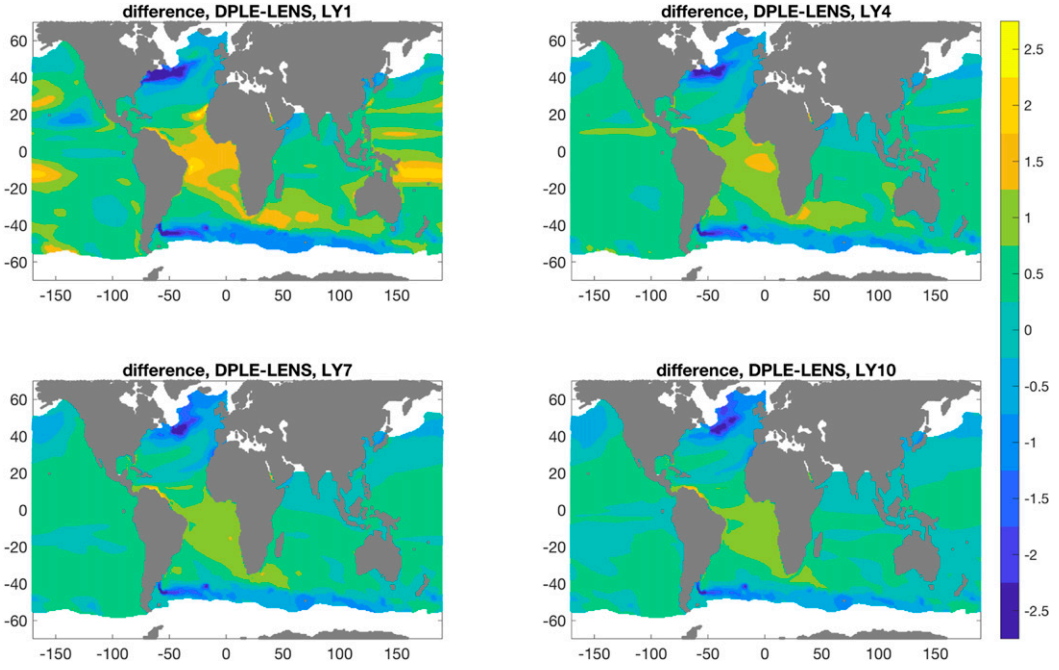


FIG. 1. The lead time–dependent drift in T295 (°C), defined as the lead time–dependent difference between the DPLE ensemble mean and the LENS ensemble mean, averaged over all start dates. This is subtracted from DPLE raw data during the drift correction step.

tending toward the background probability distribution function (pdf). We can characterize this decay of information by defining the time scale at which information from the initial conditions is lost ($\tau_{IC}$). This is expected to change regionally and as a function of the initial state. It gives us an estimate of the time over which the initialization of the climate model brings added value to the climate forecast.

However, the climate that we live in is changing. We expect to also see information in decadal forecasts arising from correctly predicting external climate forcings (Branstator and Teng 2010). While information from initial conditions decays with time, information from the forced response is expected to increase with time. A key question is: At what lead time can we expect to see the emergence of this forced response? Furthermore, for how long can we expect information from initial conditions to exceed that from correct specification of the forcing? The associated time scales, the *forcing time scale* ($\tau_F$) and *crossover time* ($\tau_X$), respectively, are also likely to change as a function of region and start date.

The forcing and crossover time scales will change as a function of the rate of climatic change. We wish to determine the crossover time most relevant for today, to assess the benefit of initializing forecasts to predict the coming decades. We therefore construct a new twenty-first-century DPLE (DPLE$^{21stC}$) dataset, where an estimate of the LENS *climate change trend signal* from 2016 to 2025 is added to the DPLE* anomalies. This climate change trend signal $C$ at a given lead time $l$ is calculated as

$$C_{y,l} = \overline{x_{j,y+l}} - \overline{x_{j,y}}, \quad (1)$$

where the average is calculated across LENS ensemble members $x_j$ and $y$ indicates the reference year for the signal. The signal is zero at lead time zero by construction[1] and is smoothed by averaging together the signals for the nine reference years 2012–20. We emphasize that this same trend signal is added to DPLE* anomalies for all 62 start dates. The resultant DPLE$^{21stC}$ dataset samples internal variability between 1954 and 2015, but with a modern climate change signal superimposed for all start dates. This allows us to estimate the crossover time scale for today's climate while retaining the large sample of start dates. This is possible because our study focuses on potential predictability, without verifying against observations.

Figure 2 shows the estimated twenty-first-century climate change trend signal as a function of lead year. By comparing the evolution of DPLE$^{21stC}$ forecasts with DPLE* forecasts, we can quantify the information in the forecast arising from the external forcing. By comparing DPLE$^{21stC}$ forecasts to

[1] Note that we are not using the LENS ensemble mean to introduce a climate change signal because the reference datasets, DPLE* and LENS*, are expressed as anomalies (i.e., have zero mean). The DPLE$^{21stC}$ dataset also has an initial mean of zero, where the mean is computed across ensemble members and start years. Adding a constant offset to all datasets to convert from a trend signal to absolute signals does not impact the diagnostics.

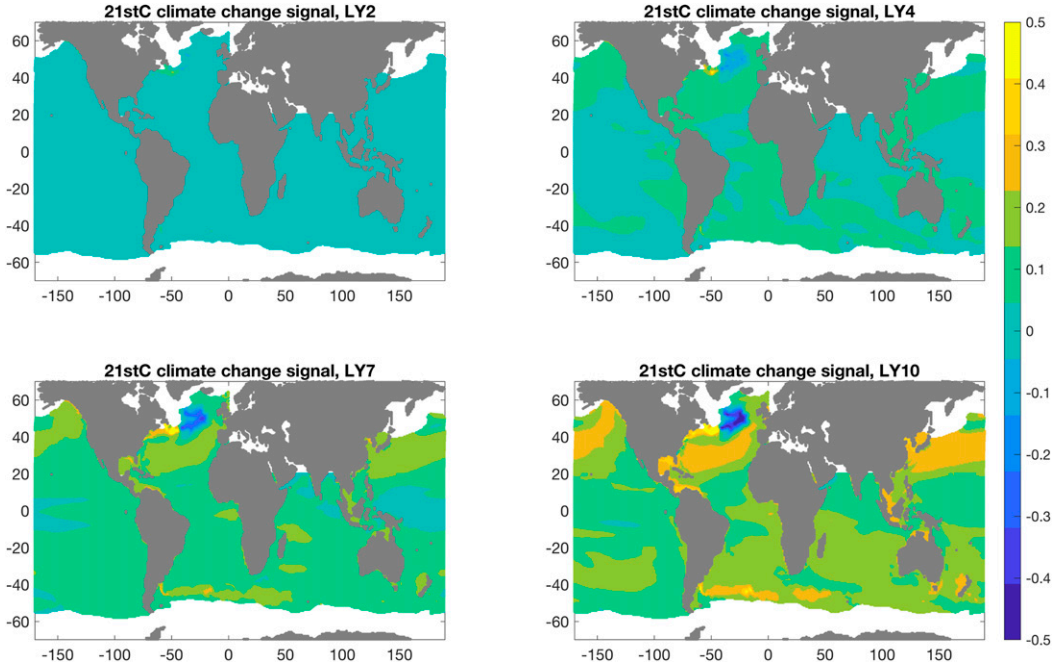


FIG. 2. The lead time–dependent climate change trend signal in T295 (°C) between 2016 and 2025, added on to the DPLE* anomalies to generate the $\text{DPLE}^{21\text{stC}}$.

the LENS* climatology, we can assess the total potential skill of decadal forecasts, including both information from the initial conditions and the forced response. The different processed DPLE and LENS datasets and the intercomparisons of interest are summarized in Fig. 3.

### c. *Measures of potential predictability*

We will assess information in the forecasts following Branstator and Teng (2010). Relative entropy $R$ is a measure of the difference between a forecast pdf $P_f$ and a baseline pdf $P_b$:

$$R = \int_S P_f(s) \ln\left[\frac{P_f(s)}{P_b(s)}\right] ds, \tag{2}$$

where $s$ represents the system state (e.g., the upper ocean heat content) and $S$ represents the system state space (e.g., the possible values taken by the upper ocean heat content). Relative entropy is often interpreted as the information contained in the forecast pdf (as measured in binary bits) compared to the information one already had in knowing the background probability of the system. In other words, relative entropy is the number of bits needed to communicate the forecast pdf to a friend, if he or she already had knowledge of the background pdf. The larger the value of $R$, the more different the forecast pdf is from the baseline pdf.

Let us assume that the DPLE- and LENS-based ensemble forecasts represent Gaussian distributions.[2] Making this assumption, we can expand Eq. (2):

$$R = \frac{1}{2}\left( \underbrace{(\boldsymbol{\mu}_f - \boldsymbol{\mu}_b)^{\mathrm{T}} (\boldsymbol{\sigma}_b^2)^{-1} (\boldsymbol{\mu}_f - \boldsymbol{\mu}_b)}_{R(\mu)} + \underbrace{\ln\left[\frac{\det(\boldsymbol{\sigma}_b^2)}{\det(\boldsymbol{\sigma}_f^2)}\right] + \operatorname{trace}(\boldsymbol{\sigma}_f^2/\boldsymbol{\sigma}_b^2) - n}_{R(\sigma)} \right), \tag{3}$$

where $\boldsymbol{\mu}_f$ and $\boldsymbol{\mu}_b$ are the forecast and baseline ensemble means respectively, and $\boldsymbol{\sigma}_f^2$ and $\boldsymbol{\sigma}_b^2$ are the forecast and baseline homogeneous covariance matrices respectively. The matrix transpose (T), determinant (det), and trace

[2] While the true pdfs underlying the DPLE and LENS predictions are likely non-Gaussian, a very large number of ensemble members would be required to detect this robustly.

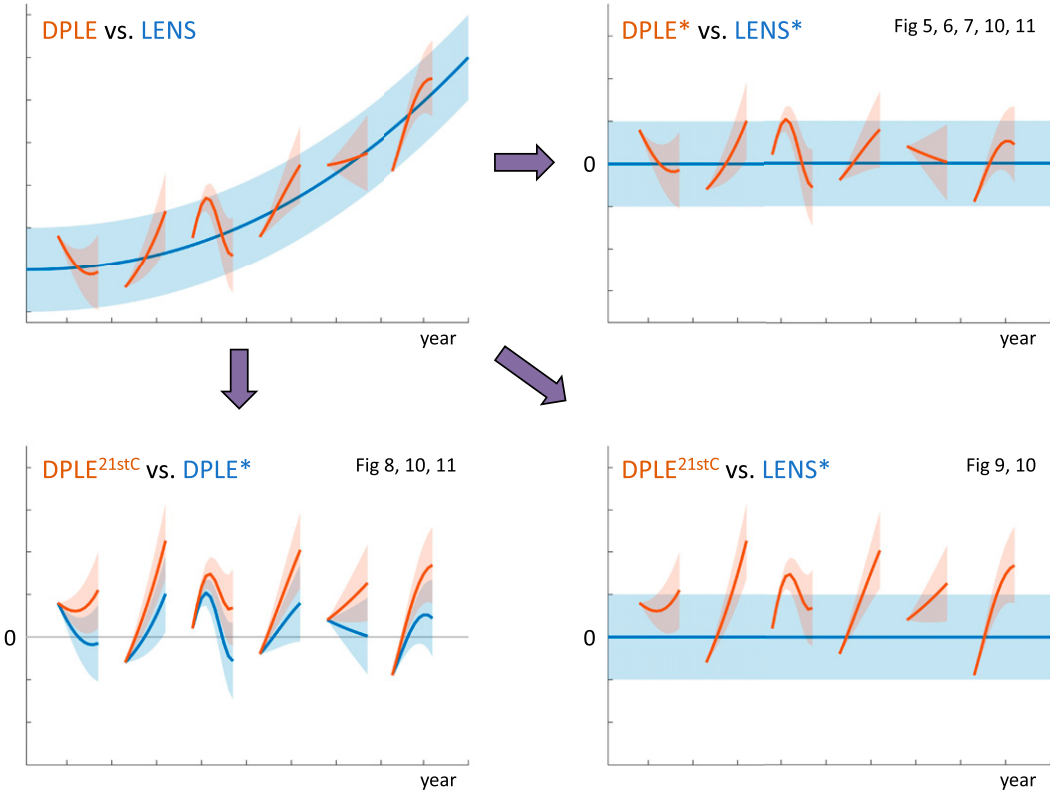


FIG. 3. Schematic summarizing the three DPLE- or LENS-based datasets used to assess the information in the decadal forecast. The arrows indicate the transformations applied to the raw LENS and DPLE data to assess information arising from (top right) the initial conditions, (bottom left) the forcing, and (bottom right) the relative contributions of both initial conditions and forced response. The orange ensemble in each panel is treated as the "forecast" whereas the blue ensemble is treated as the "baseline." The figure numbers included in each subpanel indicate where the results from each comparison can be found in this paper.

follow the usual definitions. Note that this equation is written generally for forecasts of a vector quantity of length $n$, but the equation simplifies in the 1D case by recalling that $\det(a) = \text{trace}(a) = a$ for scalar $a$. The information from the ensemble mean is represented entirely by the first term, whereas the latter terms depend only on the ensemble covariances. These are referred to as the signal and dispersion components (Kleeman 2002), $R(\mu)$ and $R(\sigma)$, and correspond to the information in the forecast arising from the ensemble mean and from the ensemble spread, respectively. In other words, the larger the values of $R(\mu)$ or $R(\sigma)$, the greater the difference between the forecast and baseline ensemble means or covariances, respectively. When comparing to the LENS* distribution, which has zero mean and climatological variance, a large $R(\mu)$ indicates a large anomaly signal in the forecast ensemble mean. In contrast, a large $R(\sigma)$ indicates a forecast distribution that has significantly smaller covariance (i.e., is significantly sharper) than the baseline climatological distribution. A key benefit of the use of relative entropy is immediately apparent. It allows us to assess the importance of the ensemble mean and spread on the same scale, such that the relative benefits of these two aspects of a probabilistic forecast may be compared. We collectively call $R$, $R(\mu)$, and $R(\sigma)$ "information measures."

### d. *Significance testing*

We assume that the LENS* ensemble forecast anomalies are stationary. In other words, we assume the ensemble covariance does not change with climate change, only the ensemble mean. Removing the ocean data points covered by sea ice ensures this is a good assumption. Ocean points covered by sea ice in the early part of the LENS* simulations but open to the atmosphere in later years due to climate change would be expected to show enhanced spread at the end of the simulation.

Since we assume stationarity, we no longer distinguish between ensemble member and year for the LENS* anomalies when assessing significance. To assess the

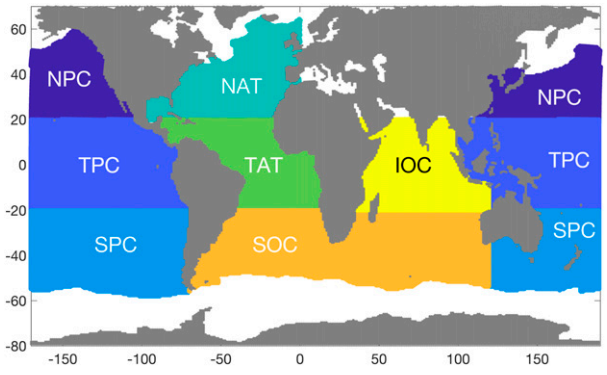


FIG. 4. The seven ocean basins considered and their short names. White regions are covered by sea ice at some point in either the LENS or DPLE datasets, and so are discarded.

TABLE 1. The seven ocean basins considered. Variance explained (ev; %) in T295 for the LENS* dataset using 15 EOFs for each ocean basin, for data concurrent with the DPLE (1955–2025). We also show the projected explained variance (pev) for the DPLE* dataset, i.e., the variance explained in DPLE* using the leading 15 LENS* EOFs.

| Basin | Short name | ev LENS* | pev DPLE* |
|---|---|---|---|
| North Pacific | NPC | 87.4 | 86.6 |
| Tropical Pacific | TPC | 94.3 | 93.7 |
| South Pacific | SPC | 79.6 | 76.5 |
| North Atlantic | NAT | 86.2 | 85.1 |
| Tropical Atlantic | TAT | 91.6 | 91.0 |
| Southern Ocean | SOC | 77.9 | 69.7 |
| Indian Ocean | IOC | 95.8 | 94.9 |

significance of the different information measures, we calculate each information measure for two 40-member ensembles, one for the forecast pdf and one for the baseline pdf, randomly sampled with replacement from across all LENS* anomalies. These two ensembles are not significantly different, being drawn from the same baseline pdf. This is repeated 62 000 times to give a distribution for each information measure due to sampling variability. The forecast pdf *for a given start year* is said to be significantly different from the baseline pdf if the information measure is greater than the 95th percentile of this distribution.

Averaging the information measures across many start years reduces the noise in the measure, thereby allowing the detection of smaller predictable signals. To calculate the significance of the difference between the forecast and baseline ensembles *averaged across 62 start years*, we calculate the distribution of information measures for the equivalent LENS* sample, that is, taking the information measures previously calculated and averaging across 62 independent pairs of LENS* samples to give 1000 estimates of this averaged metric. As before, the forecast ensembles are said to be significantly different from the baseline ensembles if the information measure is greater than the 95th percentile of this averaged distribution.

### e. *Summary statistics using EOF decomposition*

We will consider both spatial maps and statistics summarized by basin. When summarizing by basin, we will follow the approach taken by Branstator and Teng (2010). Seven ocean basins are defined in Fig. 4, such that spatial locations with similar time scales of variability are grouped into the same basin (Branstator and Teng 2010). We introduce short names for each basin in Table 1. Within each basin, we represent both DPLE- and LENS-based T295 fields on an empirical orthogonal function (EOF) basis, where the EOFs are determined from the LENS* dataset. While it is not possible to calculate Eq. (3) for the full state vector, as a 40-member ensemble is insufficient to estimate the covariance matrix between every pair of grid points, using EOFs reduces the dimensionality of the problem, and enables a single information measure to be calculated for each basin. This has benefits over simply considering the average of information metrics over a basin. EOFs are suitable for representing the propagating signals that provide predictability on decadal time scales (Teng and Branstator 2011; Yeager et al. 2015). Furthermore, if the leading EOFs are skillfully predicted for an ocean basin, then it follows that patterns or gradients of temperature over that basin have been skillfully predicted. There is evidence that it is specific, local patterns or gradients in ocean temperature that lead to predictability over land, as opposed to basin-averaged quantities (e.g., Årthun et al. 2017; Sheen et al. 2017; Ossó et al. 2018; Simpson et al. 2019).

All forecast and baseline ensembles are projected onto the leading LENS* EOFs. We choose to retain the leading 15 LENS* EOFs for each basin. This reduces the size of dataset while capturing 75%–95% of the variance in LENS* and 70%–95% of the variance in DPLE* (see Table 1). The projected explained variance of the DPLE* data using the LENS* EOFs was calculated following Bayr and Dommenget (2014). In general, the leading few EOFs explain substantially more variance than the others. For example, in the North Atlantic region, the first two EOFs explain almost half of the variance, with the following four EOFs accounting for a further 25%.

## 3. Value of initialization

As indicated in section 2b, the value of initialization is assessed by comparing the evolving DPLE*

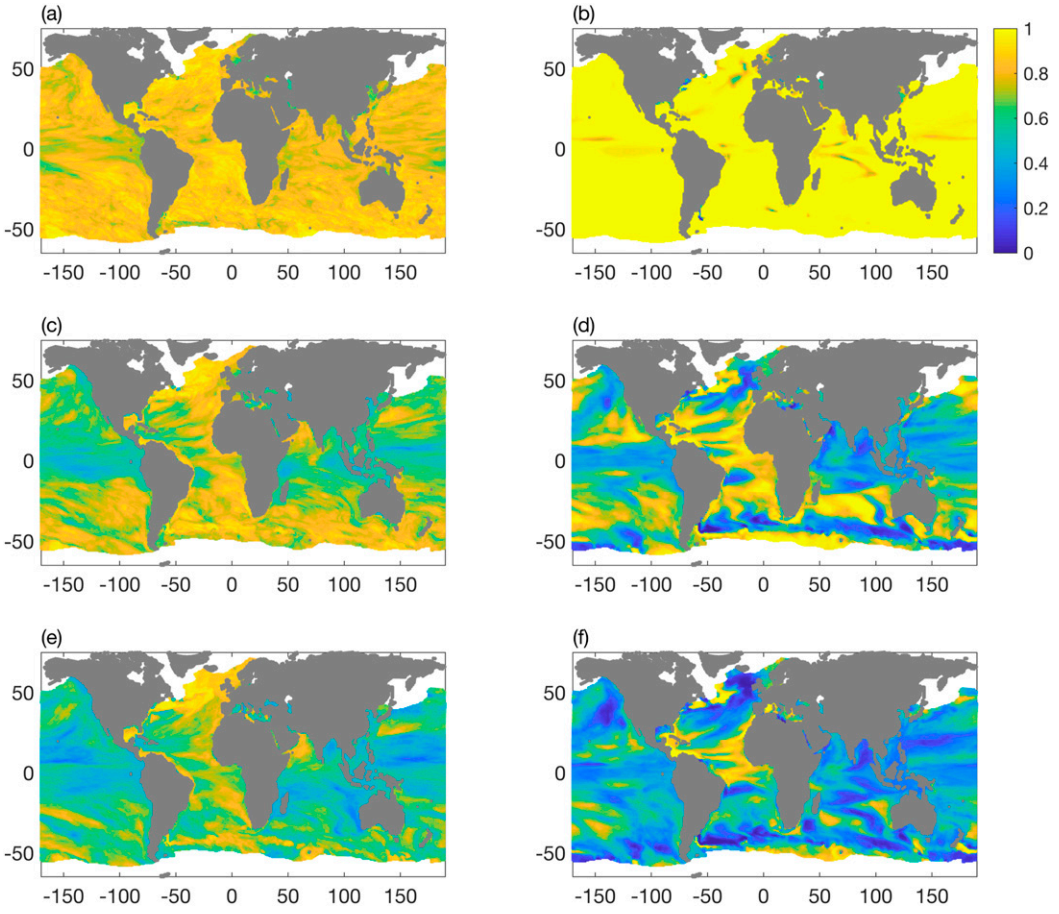


FIG. 5. The number of start years (shown as a fraction out of 62) for which (left) the DPLE* ensemble mean and (right) the DPLE* ensemble spread show a significant signal compared to LENS*, evaluated for different lead times: (a),(b) lead year 1, (c),(d) lead year 5, and (e),(f) lead year 10. The color bar in (b) is used for all panels.

ensemble forecasts with the stationary LENS* ensemble, as a function of start date, spatial location, and lead time.

### *a. Grid point verification of ensemble mean and spread*

Before using the relative entropy information metrics to assess the DPLE forecasts, we consider two more familiar verification measures in grid point space (i.e., no EOF decomposition was performed). First, we compute the ensemble mean anomaly (EM), as an indication of the strength of the mean signal in DPLE*. This is computed for every spatial location, start date, and lead time. In general, EM will decrease in time, as the ensemble mean signal in the starting conditions decays toward the LENS* climatology, which has an ensemble mean of zero by construction. Second, we compute the ensemble standard deviation (STD) for DPLE* for every spatial location, start date, and lead time. The STD will increase in time, as the ensemble members diverge from one another to span the reference climate attractor.

To assess the significance of EM and STD, we calculate EM and STD for 500 forty-member random samples drawn from the LENS* ensemble across all DPLE start dates. Once the DPLE* EM drops below the 95th percentile of the LENS* EM distribution, we say there is no longer any significant signal in the ensemble mean at the 95% level. Once the DPLE* STD exceeds the 5th percentile of the LENS* STD distribution, we say there is no longer any significant signal in the ensemble spread at the 95% level.

Figure 5 shows the fraction of start dates for which there is a significant signal in the ensemble mean (left column) and spread (right column) at a lead time of 1, 5, and 10 years. There is clearly significant information in the ensemble mean across the whole Atlantic basin. This is particularly the case in the North Atlantic north of 40°N, where large regions show a significant signal in the ensemble mean for 80%–100% of start dates even at a

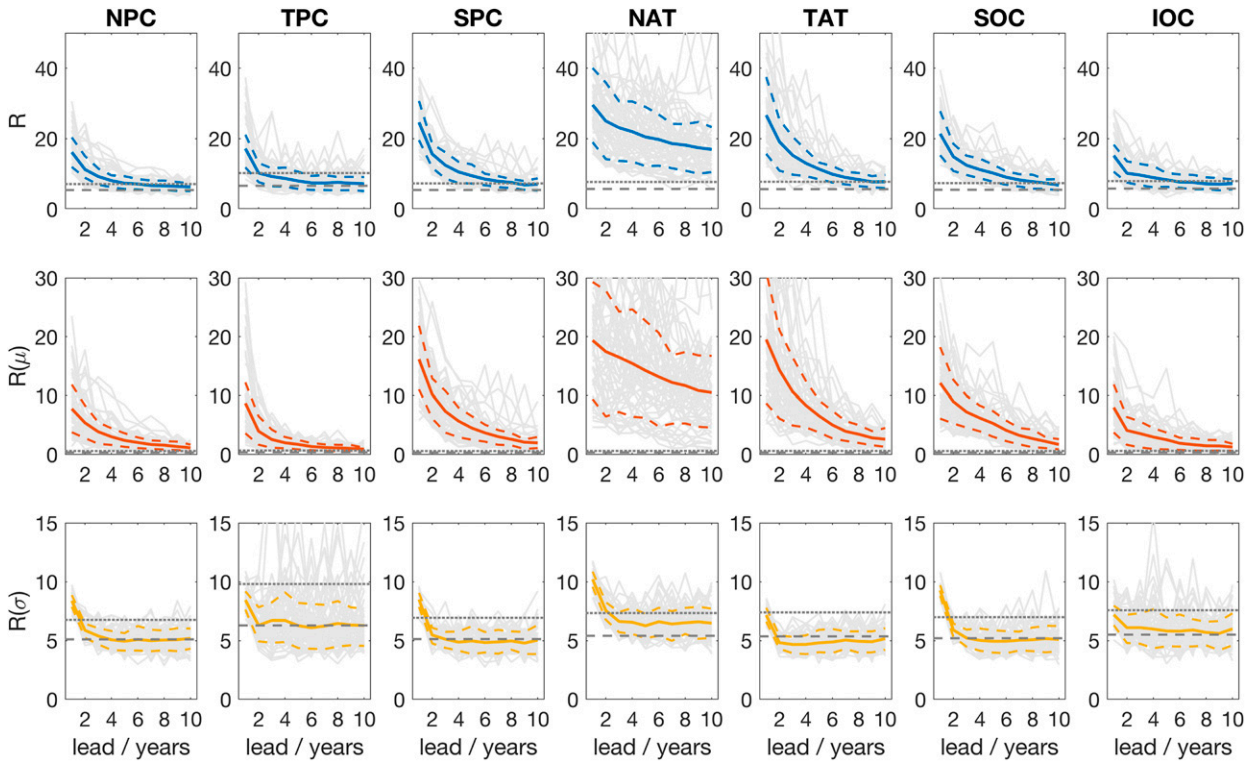


FIG. 6. Measures of information content [relative entropy $R$ and its signal $R(\mu)$ and dispersion $R(\sigma)$ components] as a function of lead time for the unforced DPLE* anomalies relative to LENS*. The individual pale gray lines indicate the information measure for each of the 62 individual start dates as a function of lead time. The solid colored line shows the average information measure as a function of lead time, with the dashed lines indicating the 16th and 84th percentiles of the distribution across the different start dates. The dark gray lines indicate the 95% significance levels: the dotted line is the significance level for a single start date, whereas the dashed line shows the significance level for the average over 62 start dates.

lead time of 10 years. Other regions also show enhanced potential predictability, particularly the Southern Ocean. On the other hand, the Pacific and Indian Oceans show relatively lower potential predictability. The signal in the ensemble spread is initially considerably higher than the signal in the ensemble mean. This is to be expected, as the DPLE ensemble is initialized using round-off errors. However, this decays more rapidly such that little signal remains at lead year 10, except in very localized regions.

### b. Basinwide information measures

The analysis in section 3a is limited, as it gives no indication of the relative importance of the signal in the ensemble mean and in the standard deviation. To address this we turn to the information measures $R$, $R(\mu)$, and $R(\sigma)$. While section 3a showed substantial spatial variability in the signal in the ensemble mean and standard deviation, there was a degree of consistency within the ocean basins outlined in Fig. 4. To efficiently summarize the information, we choose to use the EOF decomposition within each basin, as outlined in section 2e.

Figure 6 shows the information measures $R$, $R(\mu)$, and $R(\sigma)$ between DPLE* and LENS* as a function of lead time for each ocean basin. The thin gray lines show the decay of information with lead time for each of the 62 start dates. The solid colored line shows the mean information across all start dates, while the dashed colored lines show the 16th and 84th percentiles,[3] calculated across the different start dates, to summarize the distribution.

We observe that the potential predictability due to the initialization is regionally dependent. The extratropics generally show higher potential predictability than the tropics on multiannual to decadal time scales, with forecasts containing information for longer on average. We also observe regionally dependent *state-dependent predictability*. This is diagnosed by comparing the information drop-off between different start dates for a given region for both $R(\mu)$ and $R(\sigma)$. For example,

[3] For normally distributed data, these percentiles correspond to plus and minus one standard deviation from the mean.

compared to other regions, the TPC shows larger year-on-year variations in the initial information in the mean, and in the information in the spread at longer lead times.

By comparing the second and third rows we can assess the relative contributions of the ensemble mean and spread to the total information content, $R$. Some regions show comparable levels of initial information in the spread and the mean (e.g., NPC, TPC). However, the information in the ensemble spread decreases faster than in the mean. In addition the 95% significance level is substantially higher for $R(\sigma)$, because accurately calculating the covariance requires more ensemble members than accurately calculating the mean, such that the metric is more susceptible to sampling variability. At mid- to long lead times, all the information from initialization can be attributed to the ensemble mean.

In general, we observe more variation between start dates for $R(\mu)$ than for $R(\sigma)$. This indicates there is substantial flow-dependent information in the ensemble mean but that there is little flow-dependent spread in the DPLE ensemble (i.e., variation in the rate of ensemble dispersion due to the predictability of the initial state). This is unlike what is observed in ensemble forecasts on weather and seasonal time scales (Leutbecher 2010; Christensen et al. 2015; MacLeod et al. 2018). In part, this could be due to the ensemble initialization methodology used in the DPLE: while the different ensemble members of weather and seasonal forecasts receive initial condition perturbations consistent with an estimate of the state-dependent uncertainty in the initial conditions (Palmer and Zanna 2013), the DPLE ensemble members are perturbed using state-independent round-off error. Nevertheless, it is harder to interpret the dispersion of the ensemble on decadal time scales as providing information on state-dependent predictability.

To quantify the value of initialization and the degree of state-dependent predictability, and to compare these results between regions, we define the time scale at which information from the initial conditions is lost, $\tau_{IC}$. This is the first year for which the information measure falls below the 95% significance level defined using the reference LENS* climatology. These time scales can be identified in Fig. 6 as the lead time at which each gray line crosses the dotted line. Each region shows a distribution of $\tau_{IC}$, dependent on the start date. Table 2 summarizes this information and shows the mean and standard deviation of $\tau_{IC}$ as a function of region for the total information content $R$. For some regions and start dates, significant information is present at a lead time of 10 years. For the purposes of calculating the statistics in this summary table, linear extrapolation is used to estimate the crossover time for these start dates. The equivalent diagnostics for $R(\mu)$ and $R(\sigma)$ are shown in Tables S1 and S2 of the online supplemental material.

TABLE 2. Three time scales of interest for the relative entropy $R$. The mean time scale is calculated across all 62 start dates, with the standard deviation shown in parentheses; $\tau_{IC}$: the year of loss of information from the initial conditions; $\tau_F$: the year of emergence of information from the forcing; $\tau_X$: the crossover time, after which external forcing provides more information than the initialization. Linear extrapolation is used to estimate time scales longer than 10 years.

| Basin | $\tau_{IC}$ | $\tau_F$ | $\tau_X$ |
|---|---|---|---|
| NPC | 5.0 (1.8) | 11.3 (2.3) | 10.0 (1.0) |
| TPC | 2.4 (0.8) | 14.4 (2.7) | 11.9 (4.7) |
| SPC | 7.0 (3.5) | 8.0 (1.1) | 9.1 (1.3) |
| NAT | 15.5 (8.1) | 7.2 (0.9) | 8.5 (2.1) |
| TAT | 7.7 (2.4) | 8.0 (0.9) | 8.4 (0.9) |
| SOC | 7.3 (1.9) | 8.9 (1.1) | 8.4 (1.1) |
| IOC | 4.4 (2.6) | 14.7 (2.9) | 12.1 (2.5) |

The mean value of $\tau_{IC}$ shown in Table 2 varies substantially between regions. Several regions have $\tau_{IC}$ of approximately 7 years on average, consistent with earlier studies (Branstator and Teng 2010, 2012), although TPC shows initial value predictability for only 2.4 years. The standard deviation of $\tau_{IC}$ characterizes the degree of dependence of $\tau_{IC}$ on the initial state of the system. For most regions, this is substantial, with a standard deviation of two to three years. The TPC region shows less variability, and rapidly loses information from the initial conditions for all start dates. In the case of the North Atlantic (NAT) region, forecasts initialized from all but three start dates still show significant $R$ at a lead time of 10 years. Linear extrapolation suggests the average value for $\tau_{IC}$ in this region is over 15 years, with a high standard deviation of 8 years.

Previous studies have compared the year of loss of information for different regions, to order them from most to least predictable (e.g., Branstator and Teng 2010). However, such studies use only a single start date. For conclusions from such studies to be general, one region must have consistently higher $\tau_{IC}$ than another. To assess this, we consider the joint distribution of $\tau_{IC}$ for pairs of regions (Fig. S1). We find that there is generally little correlation between the year of loss of information diagnosed for a given start year for different basins (Fig. S1). Since some pairs of regions have similar $\tau_{IC}$ on average, this means that the ordering of regions from most to least predictable changes depending on the start year, with implications for predictability studies which consider only one start date.

The NAT region stands out as a key region of interest. First, it shows information beyond the 10-yr window in both $R$ and $R(\mu)$. It also shows a very high degree of state-dependent predictability, particularly for the ensemble mean, but also to some extent for the ensemble

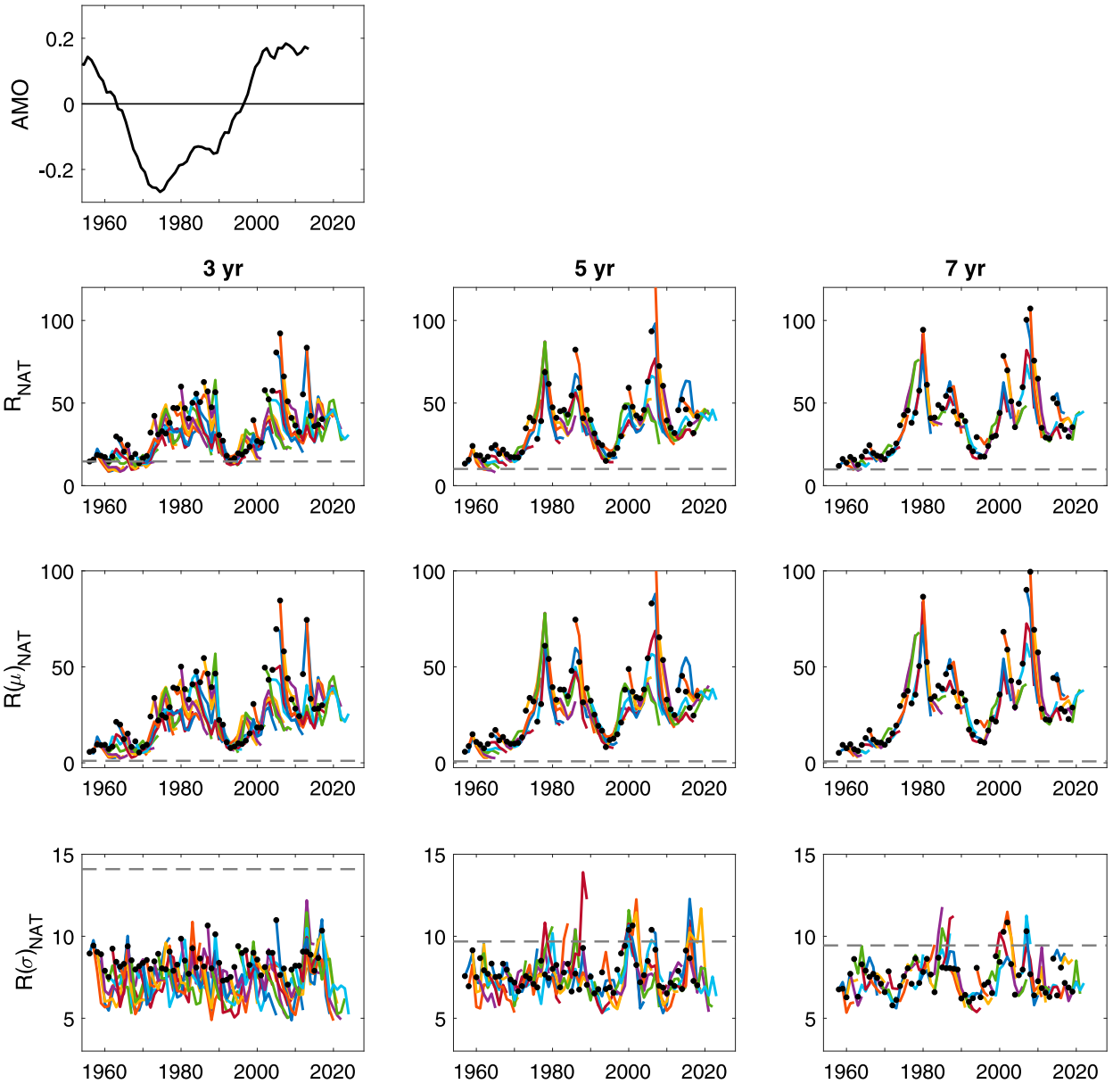


FIG. 7. Measures of information for the NAT region as a function of year of verification. The 62 DPLE forecasts are shown as different colored lines, with the forecast year one indicated by a black circle in each case. The forecasts have been averaged in time, with the three columns showing the results from predicting 3-, 5- and 7-yr running averages, respectively. The top left panel shows the AMO signal for reference, as provided by the National Oceanic and Atmospheric Administration's Physical Sciences Laboratory (Enfield et al. 2001).

spread. In the North Atlantic, the dominant oceanic mode of variability is the Atlantic multidecadal oscillation (AMO). This low-frequency variability imprints on the forecasts in this region, with different start dates including different phases of the AMO in their initial conditions.

Figure 7 shows information measures for the NAT region as a function of the year of verification. It is evident that the information content in DPLE is maximum when AMO is at peak amplitude (e.g., Ting et al. 2009) and that (as already seen) almost all of the state dependency in the signal comes from the mean forecasts. It is interesting to note that in this region we do not always see a *decay* of information after initialization: the DPLE is able to predict a growing AMO phase, even if the T295 initial conditions appear neutral. For example, consider start dates around 1970 and 1995, for which $R$ is initially small but then increases with lead time. This is visible when annual averages are considered (not shown), but taking further temporal averages as in Fig. 7

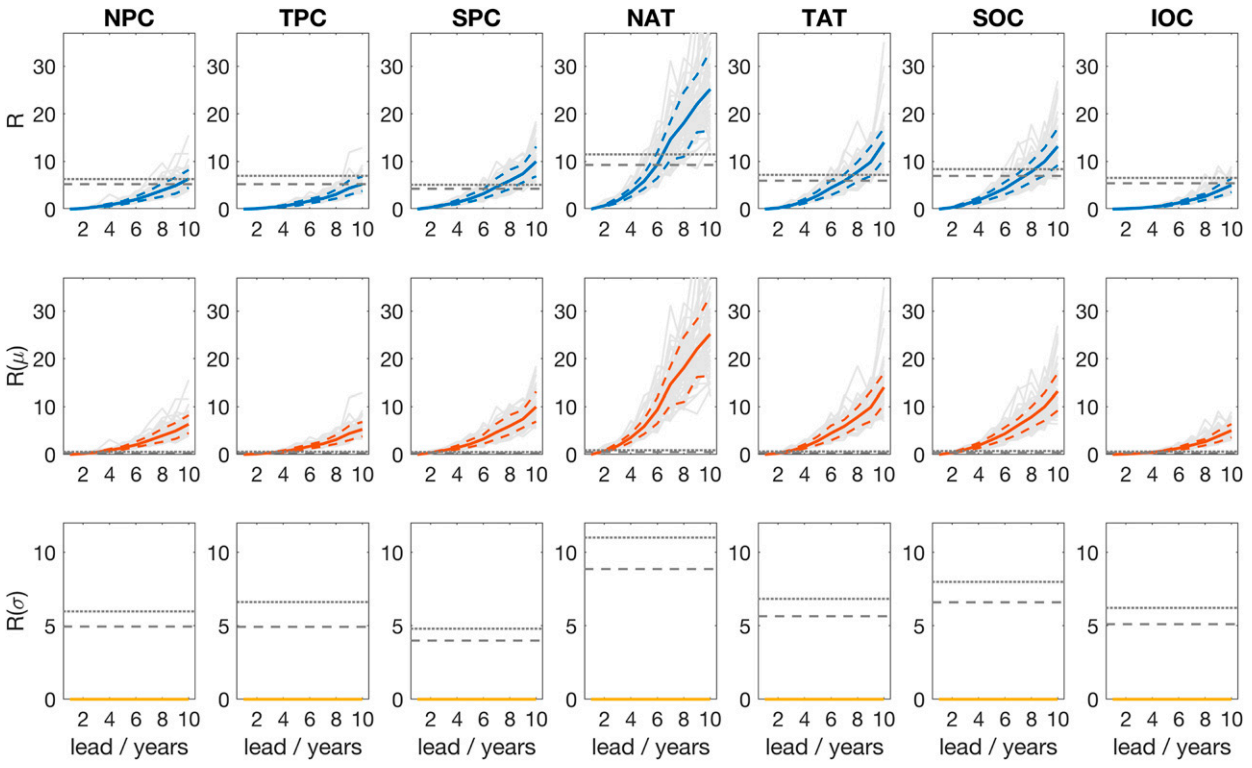


FIG. 8. As in Fig. 6, but for measures of information content as a function of lead time for the forced DPLE anomalies, DPLE$^{21stC}$, compared to unforced DPLE* anomalies. Compared to the baseline, the forecast pdf contains additional information from the forced response. The dark gray dotted/dashed lines indicate the 95% significance levels as before.

makes the signal clearer. When longer temporal averages are taken, the information in the forecast is dependent on verification year not on lead time. In other words, the potential predictability of the low-frequency AMO is dependent only on the state of the AMO at verification. For short temporal averages, state-dependent variability is not readily detectable in the ensemble spread. However, for the 7-yr running average, there is an indication of increased information in the ensemble spread concurrent with increased information in the ensemble mean. Increased information in the ensemble spread indicates that the ensemble has smaller standard deviation than the climatological pdf (i.e., enhanced sharpness) (Gneiting and Raftery 2007). For a well-calibrated forecast, increased sharpness indicates enhanced predictability. *The DPLE therefore indicates higher predictability for strong positive or negative phases of the AMO than for the neutral phase.* This can be intuitively understood: a system far from equilibrium will tend to be drawn back toward its mean state.

## 4. Value of initialization in a changing world

Having assessed the value of initialization in a stationary climate in section 3, we turn our attention to decadal climate prediction in a changing climate, where information in the forecasts also arises due to predicting the forced climate response.

To assess the relative importance of information from the initial conditions and from the forced response, we consider two further pairs of experiments. We first assess potential predictability in decadal forecasts arising from just the forced response. Figure 8 shows the information measures for the DPLE$^{21stC}$ ensemble compared to DPLE*. As expected, the information from the forced response increases with time. The information is entirely contained in the ensemble mean, as each ensemble member receives the same large-scale forcing. Even though every DPLE$^{21stC}$ forecast has the same forced signal, there is substantial variability in the rate of growth of information between different start dates. In other words, for certain states of the climate system we see a stronger impact of the climate change signal than for others.

The forcing time scale, $\tau_F$, is the time scale at which information due to the forced response emerges. This is the first year for which the information measure for the difference between the DPLE$^{21stC}$ and DPLE* ensembles increases above the 95% significance level defined

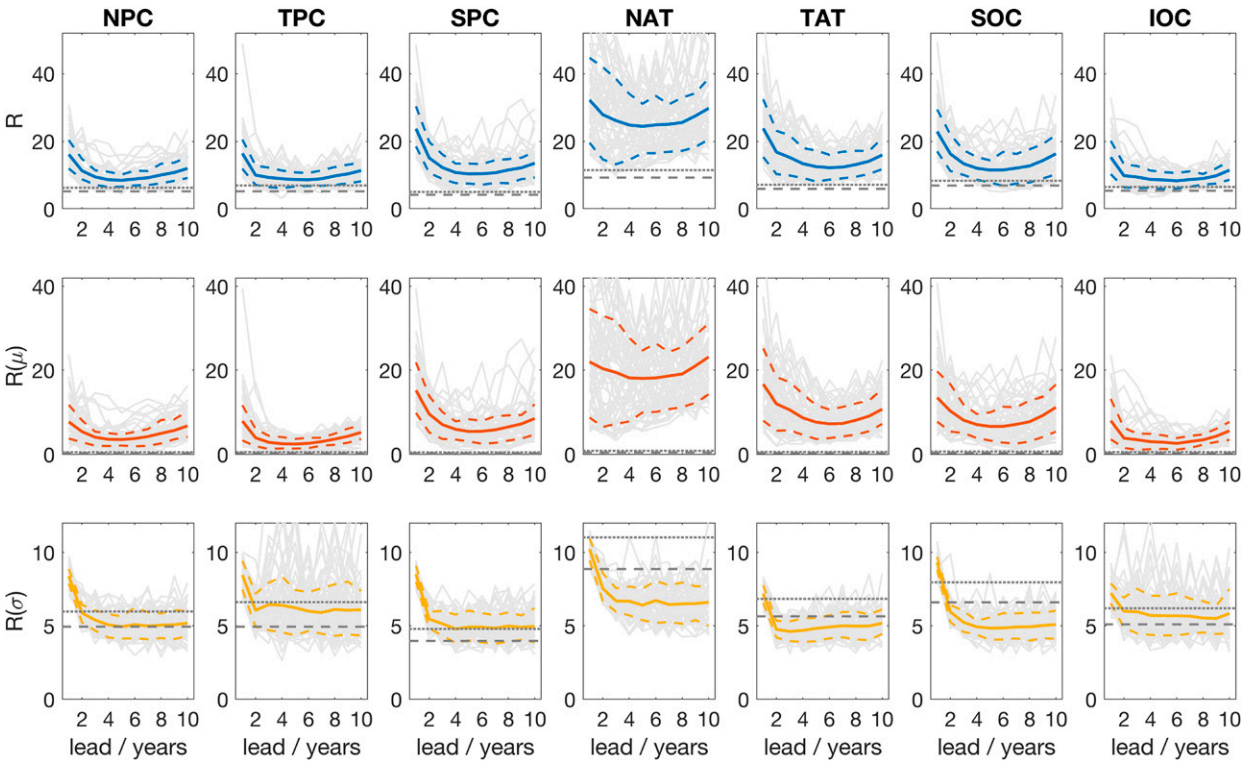


FIG. 9. As in Fig. 6, but for measures of information content as a function of lead time for the forced DPLE anomalies, DPLE$^{21stC}$, compared to unforced LENS*. Compared to the baseline, the forecast pdf contains additional information from both the initial conditions and forced response. The dark gray dotted/dashed lines indicate the 95% significance levels as before.

using the reference LENS* climatology. These time scales can be identified in Fig. 8 as the lead time at which each gray line crosses the dotted line. This is shown for $R$ in Table 2, and for $R(\mu)$ in Table S1. The emergence of information is regionally dependent, and occurs no earlier than a lead time of seven years. For the TPC and IOC regions, the forecasts from many start dates do not show significant information from the forced response even at lead year 10. From Table S1 and Fig. 8, we can see that the emergence of information in the ensemble mean occurs earlier, between lead years 3 and 5, with some variability between regions and start dates.

To assess potential predictability arising from both the initial conditions and the forced response, we compute the information measures for DPLE$^{21stC}$ forecasts compared to LENS*. These are shown in Fig. 9. As expected, information in the DPLE$^{21stC}$ initially decays, before the external forcing provides a source of information to the forecast. Combining both sources of information results in forecasts with significant potential predictability for most start dates.

For each start date, we compare the decay of information from the initial conditions (DPLE* vs LENS*) with the increase of information due to the forced response (DPLE$^{21stC}$ vs DPLE*). Figure 10 shows the fraction of information arising from the initialization compared to that arising from the forcing for $R$ and $R(\mu)$ for each region. It demonstrates that the initial conditions provide the dominant source of information over the duration of the 10-yr forecasts, though with some regional dependency.

The year for which the forcing first gives more information than the initial conditions is the crossover time scale $\tau_X$. This is the first year for which the relative entropy calculated between DPLE$^{21stC}$ and DPLE* exceeds the relative entropy calculated between DPLE* and LENS*. The average value for $\tau_X$ is indicated in Fig. 10 as the year for which the ensemble mean line crosses the dotted line. The years for which the 16th and 84th percentile lines cross the dotted line indicate the variability in $\tau_X$. The numerical values for the crossover time scale are summarized in Table 2 for $R$, and in Table S1 for $R(\mu)$. On average, the crossover time scale for $R$ is between 8 and 12 years depending on region. The crossover time is relatively late because the initialization also gives information via the ensemble spread, which the forcing does not impact.

The crossover time scale for information in the ensemble mean is earlier than for the full information measure. In fact, we observe this to be remarkably consistent across different regions, at approximately 6–7 years. This indicates that, when considering ocean basins as a whole, ocean basins that show more potential predictability also show a larger response to climate change.

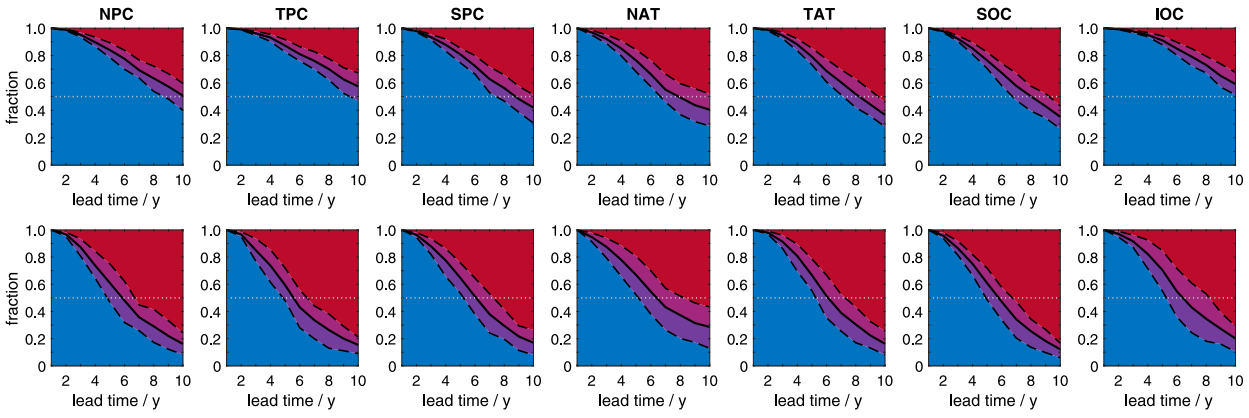


FIG. 10. The fraction of information in the forecast arising from the initial conditions (blue) or from the external forcing (red) for each region. The top row shows the results for the total information content, while the bottom row shows the results for ensemble mean only. The fraction was calculated independently for each start date. The solid black line indicates the mean partitioning; the dashed black lines are the 16th and 84th percentiles to indicate the variability in the partitioning. The gray dashed line can be used to estimate the "crossover time" shown in Table 2. All information in the ensemble spread arises from the initialization, so the results for spread are not shown.

## 5. Local information measures

To further investigate the balance between decay of information from the initialization and growth of information from the external forcing, we return to considering grid point diagnostics. We compute each information measure $R$, $R(\mu)$, and $R(\sigma)$ at each grid point as a function of lead time. The calculation uses the scalar version of Eq. (3). The information measures are calculated for DPLE* versus LENS*, to assess the decay of information from the initial conditions, as well as for DPLE$^{21stC}$ versus DPLE*, to assess the growth of information from the forced response. For each case, we summarize the measures by averaging over start dates.

Figure 11 shows the information at lead years 1 and 10 due to the initialization, compared to the information at lead year 10 due to the forced response for the NAT region. It is clear that the information from the forcing projects onto a different pattern than the information due to the initial conditions. While in the area average the information from the initial conditions is overtaken by that from the forcing by around lead year 8, locally it can be substantially higher. In fact, in the NAT the region with the greatest potential predictability from the initial conditions is also the region that is slowest to show a forced response. The difference in the spatial patterns impacted by information from the initial conditions and forced response is observed in other regions—for global maps, see Fig. S2 in the online supplemental material.

## 6. Discussion and conclusions

We have analyzed the potential predictability of upper ocean heat content in the CESM Decadal Prediction Large Ensemble (DPLE) due to the ocean initialization and the forced response. By using relative entropy as a metric for the information contained in the DPLE, we attribute the source of the potential predictability in the decadal forecast to either the ensemble mean or the ensemble spread. A key consideration was the extent to which the DPLE shows *state-dependent predictability*, that is, predictability arising in either the ensemble mean or spread that is dependent on the initial state of the forecast. This is of particular interest, as state-dependent predictability is observed in forecasts on weather and subseasonal-to-seasonal time scales, but has not yet been considered on decadal time scales. The large number of start dates in the DPLE (62) makes this analysis possible.

Assessing the potential predictability of upper ocean heat content, as opposed to the realizable predictability from comparison with observations, allows us to separate predictability due to the initialization from predictability due to the forced climate response. We created a hypothetical dataset of decadal forecasts with different initial conditions, but where all start dates experience the forcing relevant for today's greenhouse gas emissions. This allowed us to quantify the time scale at which the forced climate response emerges and the time scale at which it provides more information than the initial conditions, where both time scales are relevant for decadal forecasts made over the coming years.

Predictability due to the initialization was found to be regionally dependent, with the tropics generally showing faster rates of decay of information than the extratropics. The entire Atlantic basin showed extended predictability compared to other basins, with forecasts from a large fraction of start years containing significant

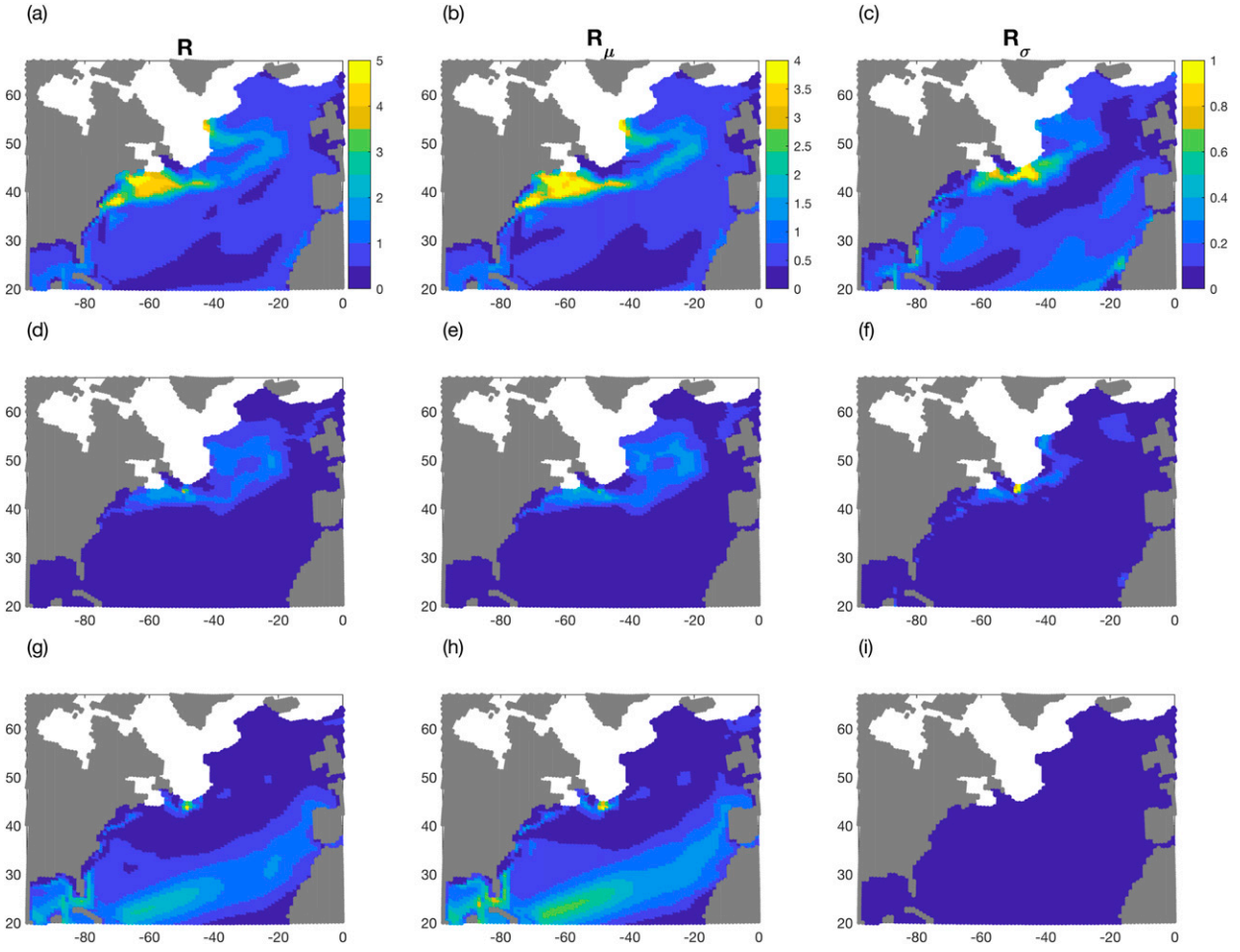


FIG. 11. The local information content as measured by the relative entropy $R$ and its signal $R(\mu)$ and dispersion $R(\sigma)$ components, focusing on the North Atlantic region. Information from the initial conditions diagnosed by comparing DPLE* and LENS* at (a)–(c) lead year 1 and (d)–(f) lead year 10. (g)–(i) Information from the forcing diagnosed by comparing $\text{DPLE}^{21\text{stC}}$ and DPLE*, at lead year 10. The color bar shown for each panel in the top row corresponds to all panels in that column.

information due to initialization even at lead times of 10 years. We also detected substantial state-dependent predictability in the DPLE. The year of loss of information from the initial conditions showed a standard deviation across start dates of two to three years, with some regional dependencies. This highlights the importance of considering many start years when analyzing decadal predictability, as one start year will not necessarily give a representative answer. Considering the joint distribution of the year of loss of information calculated for different basins confirmed that the choice of start year can lead to differences in the ordering of regions from most to least predictable.

In general, the year of loss of information for the ensemble mean varies more across different start years than the year of loss of information calculated for the ensemble spread. This indicates the ensemble mean shows much higher state-dependent predictability than the ensemble spread. This means it is difficult to interpret the rate of spreading out of the ensemble forecasts as an indicator of the predictability of forecasts from that initial state, as is commonplace in weather and seasonal forecasts.

Our analysis highlights the North Atlantic (NAT) as a clear region of interest. In the NAT, initialization is a source of information in the forecasts beyond a 10-yr lead time for 95% of the start dates considered. This information is due to the signal in the ensemble mean. The NAT also shows a very high degree of state-dependent predictability, largely due to the ensemble mean. It has long been known that initialized forecasts for the North Atlantic show high potential predictability compared to uninitialized forecasts (Griffies and Bryan 1997), and several authors have documented skilful decadal predictions in this region (e.g., Collins et al. 2006; García-Serrano et al. 2012; Robson et al. 2014). The

dominant mode of ocean variability in this region is the Atlantic multidecadal oscillation (AMO), with a period of approximately 70 years. Different start dates therefore include different phases of the AMO in their initial conditions.

The DPLE forecasts in the NAT region do not always show a decay of information with increasing lead time. For certain start years, an increase of information with lead time indicates that the DPLE can predict a growing (positive or negative) phase of the AMO. Yeager et al. (2015) demonstrate that initialized forecasts with CESM are able to predict realistic ocean heat transport anomalies. During initialization, the correctly specified atmospheric fluxes associated with a positive North Atlantic Oscillation lead to the formation of North Atlantic deep water in the initial ocean state. This results in predictable ocean thermohaline circulation changes in the subpolar Atlantic that drive ocean heat content changes consistent with observations (Yeager et al. 2015; Yeager 2020). Furthermore, Yeager et al. (2015) show that CESM is able to propagate such preformed deep water anomalies from the Labrador Sea to the central North Atlantic. Here the initialized signal re-emerges and provides predictability to the upper ocean, explaining the observed increase of information with lead time for T295. While the information in the decadal forecasts is largely due to the ensemble mean, there is some evidence for state-dependent predictability in the ensemble spread in this region, with enhanced information in the ensemble spread concurrent with enhanced information in the mean. This would indicate the potential for the ensemble spread to be used as a measure of predictability in this region, and that forecasts for strong positive or negative phases of the AMO would be identified as ''more predictable'' using this measure.

We quantified the fraction of information in the decadal forecasts arising from the initialization and from the forcing, respectively. We find that the initial conditions dominate the total information in the forecasts out to lead times of 8–10 years, depending on the region. This is in line with estimates from previous studies (e.g., Branstator and Teng 2010, 2012). However, we also find substantial variability in this crossover time, with the crossover time varying by up to four years for different start dates for some regions. Combining information from the initialization and the forced response leads to potential predictability for the entire 10-yr window for all regions, for all but a small number of start dates.

The crossover time scale is a useful summary statistic, indicating when the forced response provides more information than the initial conditions for a given region. However, it can mask the fact that initialization provides valuable information to the forecast beyond the crossover time. We demonstrate that the information from the initial conditions and the forced response project onto very different patterns in the ocean. Even at a lead time of 10 years, the initialization provides information to the forecasts in regions that are different to those showing a forced response. Multidecadal simulations would need to be performed to characterize the true time scale over which the initialization benefits the forecast. Nevertheless, our results provide strong motivation for the initialization of climate models used for climate projections on longer than decadal time scales.

The relationship between potential predictability and actual realizable skill is an area of ongoing research (e.g., Kumar et al. 2014; Eade et al. 2014; Scaife and Smith 2018; Boer et al. 2019). The presence of potential predictability (or potential skill) in a model is not an indicator of realizable skill, and potential skill is not always an upper bound on actual skill (e.g., Scaife et al. 2014). Nevertheless, if there is no potential predictability in a model, then there can be no actual skill because, in that region on that time scale, initialized forecasts do not differ from uninitialized forecasts. Potential predictability studies such as the one presented can therefore indicate the limit of useful predictability possible for a given model, and highlight regions to focus on when assessing actual skill. While a key strength of our work is the consideration of potential predictability for many start dates, a key limitation of is the use of a single model to draw conclusions. The details of our results are likely to be model dependent. Future work will seek to address this limitation.

*Acknowledgments.* HMC was funded by a National Center for Atmospheric Research Advanced Study Program Postdoctoral Fellowship, and by the Natural Environment Research Council Grant NE/P018238/1. SY acknowledges the support of National Science Foundation (NSF) Grant OCE-1243015. We acknowledge the use of data produced by the CESM-DPLE community project. CESM-DPLE made use of computer resources provided by the National Energy Research Scientific Computing Center, supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC02-05CH11231, as well as resources provided by an Accelerated Scientific Discovery grant for Cheyenne (https://doi.org/10.5065/D6RX99HX) that was awarded by NCAR's Computational and Information Systems Laboratory. We also acknowledge the CESM Large Ensemble (LENS) Community Project, which made use of supercomputing resources awarded by NCAR's Computational and Information Systems Laboratory on the Yellowstone computer. DPLE and LENS data are available for download from the NCAR Climate Data

Gateway (https://www.earthsystemgrid.org). NCAR is a major facility sponsored by NSF under Cooperative Agreement 1852977. All analysis code is available on request from HMC.

## REFERENCES


Årthun, M., T. Eldevik, E. Viste, H. Drange, T. Furevik, H. L. Johnson, and N. S. Keenlyside, 2017: Skillful prediction of northern climate provided by the ocean. *Nat. Commun.*, **8**, 15875, https://doi.org/10.1038/NCOMMS15875.

Athanasiadis, P. J., S. Yeager, Y.-O. Kwon, A. Bellucci, D. W. Smith, and S. Tibaldi, 2020: Decadal predictability of North Atlantic blocking and the NAO. *npj Climate Atmos. Sci.*, **3**, 20, https://doi.org/10.1038/S41612-020-0120-6.

Bayr, T., and D. Dommenget, 2014: Comparing the spatial structure of variability in two datasets against each other on the basis of EOF-modes. *Climate Dyn.*, **42**, 1631–1648, https://doi.org/10.1007/s00382-013-1708-x.

Boer, G. J., and Coauthors, 2016: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geosci. Model Dev.*, **9**, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016.

——, V. V. Kharin, and W. J. Merryfield, 2019: Differences in potential and actual skill in a decadal prediction experiment. *Climate Dyn.*, **52**, 6619–6631, https://doi.org/10.1007/s00382-018-4533-4.

Branstator, G., and H. Teng, 2010: Two limits of initial-value decadal predictability in a CGCM. *J. Climate*, **23**, 6292–6311, https://doi.org/10.1175/2010JCLI3678.1.

——, and ——, 2012: Potential impact of initialization on decadal predictions as assessed for CMIP5 models. *Geophys. Res. Lett.*, **39**, L12703, https://doi.org/10.1029/2012GL051974.

——, ——, G. A. Meehl, M. Kimoto, J. R. Knight, M. Latif, and A. Rosati, 2012: Systematic estimates of initial-value decadal predictability for six AOGCMs. *J. Climate*, **25**, 1827–1846, https://doi.org/10.1175/JCLI-D-11-00227.1.

Christensen, H. M., I. M. Moroz, and T. N. Palmer, 2015: Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quart. J. Roy. Meteor. Soc.*, **141**, 538–549, https://doi.org/10.1002/qj.2375.

Collins, M., and Coauthors, 2006: Interannual to decadal climate predictability in the North Atlantic: A multimodel-ensemble study. *J. Climate*, **19**, 1195–1203, https://doi.org/10.1175/JCLI3654.1.

Corti, S., and Coauthors, 2015: Impact of initial conditions versus external forcing in decadal climate predictions: A sensitivity experiment. *J. Climate*, **28**, 4454–4470, https://doi.org/10.1175/JCLI-D-14-00671.1.

Danabasoglu, G., and Coauthors, 2016: North Atlantic simulations in Coordinated Ocean-ice Reference Experiments phase II (CORE-II). Part II: Inter-annual to decadal variability. *Ocean Modell.*, **97**, 65–90, https://doi.org/10.1016/j.ocemod.2015.11.007.

Delworth, T. L., F. Zeng, L. Zhang, R. Zhang, G. A. Vecchi, and X. Yang, 2017: The central role of ocean dynamics in connecting the North Atlantic Oscillation to the extratropical component of the Atlantic multidecadal oscillation. *J. Climate*, **30**, 3789–3805, https://doi.org/10.1175/JCLI-D-16-0358.1.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, https://doi.org/10.1002/2014GL061146.

Enfield, D., A. M. Mestas-Nuñez, and P. J. Trimble, 2001: The Atlantic Multidecadal Oscillation and its relationship to rainfall and river flows in the continental U.S. *Geophys. Res. Lett.*, **28**, 2077–2080, https://doi.org/10.1029/2000GL012745.

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866.

Frankignoul, C., G. Gastineau, and Y. O. Kwon, 2017: Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic multidecadal oscillation and the Pacific decadal oscillation. *J. Climate*, **30**, 9871–9895, https://doi.org/10.1175/JCLI-D-17-0009.1.

García-Serrano, J., F. J. Doblas-Reyes, and C. A. Coelho, 2012: Understanding Atlantic multi-decadal variability prediction skill. *Geophys. Res. Lett.*, **39**, L18708, https://doi.org/10.1029/2012GL053283.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Griffies, S. M., and K. Bryan, 1997: Predictability of the North Atlantic multidecadal climate variability. *Science*, **275**, 181–184, https://doi.org/10.1126/science.275.5297.181.

IPCC, 2014: *Climate Change 2014: Synthesis Report*. R. K. Pachauri and L. A. Meyer, Eds., IPCC, 151 pp., https://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_FINAL_full.pdf.

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. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1.

Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, **59**, 2057–2072, https://doi.org/10.1175/1520-0469(2002)059<2057:MDPUUR>2.0.CO;2.

Kumar, A., P. Peng, and M. Chen, 2014: Is there a relationship between potential and actual skill? *Mon. Wea. Rev.*, **142**, 2220–2227, https://doi.org/10.1175/MWR-D-13-00287.1.

Kushnir, Y., and Coauthors, 2019: Towards operational predictions of the near-term climate. *Nat. Climate Change*, **9**, 94–101, https://doi.org/10.1038/s41558-018-0359-7.

Leutbecher, M., 2010: Diagnosis of ensemble forecasting systems. *Seminar on Diagnosis of Forecasting and Data Assimilation Systems*, ECMWF, Shinfield Park, Reading, 235–266.

Lorenz, E. N., 1975: Climatic predictability. *The Physical Basis of Climate and Climate Modeling*, GARP Publication Series, Vol. 16, 132–136.

MacLeod, D., C. O. Reilly, T. Palmer, and A. Weisheimer, 2018: Flow dependent ensemble spread in seasonal forecasts of the boreal winter extratropics. *Atmos. Sci. Lett.*, **19**, 1–7, https://doi.org/10.1002/ASL.815.

Merryfield, W. J., and Coauthors, 2020: Current and emerging developments in subseasonal to decadal prediction. *Bull. Amer. Meteor. Soc.*, **101**, E869–E896, https://doi.org/10.1175/BAMS-D-19-0037.1.

Ossó, A., R. Sutton, L. Shaffrey, and B. Dong, 2018: Observational evidence of European summer weather patterns predictable from spring. *Proc. Natl. Acad. Sci. USA*, **115**, 59–63, https://doi.org/10.1073/pnas.1713146114.

Palmer, T. N., 2006: Predictability of weather and climate: From theory to practice. *Predictability of Weather and Climate*, T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 1–29.

——, and L. Zanna, 2013: Singular vectors, predictability and ensemble forecasting for weather and climate. *J. Phys.*, **46A**, 254018, https://doi.org/10.1088/1751-8113/46/25/254018.

Qasmi, S., C. Cassou, and J. Boé, 2020: Teleconnection processes linking the intensity of the Atlantic multidecadal variability

to the climate impacts over Europe in boreal winter. *J. Climate*, **33**, 2681–2700, https://doi.org/10.1175/JCLI-D-19-0428.1.

Robson, J., R. Sutton, and D. Smith, 2014: Decadal predictions of the cooling and freshening of the North Atlantic in the 1960s and the role of ocean circulation. *Climate Dyn.*, **42**, 2353–2365, https://doi.org/10.1007/s00382-014-2115-7.

Scaife, A. A., and D. Smith, 2018: A signal-to-noise paradox in climate science. *npj Climate Atmos. Sci.*, **28**, 1–28, https://doi.org/10.1038/s41612-018-0038-4.

——, and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, https://doi.org/10.1002/2014GL059637.

Sheen, K. L., D. M. Smith, N. J. Dunstone, R. Eade, D. P. Rowell, and M. Vellinga, 2017: Skilful prediction of Sahel summer rainfall on inter-annual and multi-year timescales. *Nat. Commun.*, **8**, 14966, https://doi.org/10.1038/ncomms14966.

Simpson, I. R., S. G. Yeager, K. A. McKinnon, and C. Deser, 2019: Decadal predictability of late winter precipitation in western Europe through an ocean–jet stream connection. *Nat. Geosci.*, **12**, 613–619, https://doi.org/10.1038/s41561-019-0391-x.

Slingo, J., and T. N. Palmer, 2011: Uncertainty in weather and climate prediction. *Philos. Trans. Roy. Soc.*, **A369**, 4751–4767, https://doi.org/10.1098/RSTA.2011.0161.

Smith, D. M., and Coauthors, 2019: Robust skill of decadal climate predictions. *npj Climate Atmos. Sci.*, **2**, 1–10, https://doi.org/10.1038/s41612-019-0071-y.

Sutton, R. T., and D. L. R. Hodson, 2005: Atlantic Ocean forcing of North American and European summer climate. *Science*, **309**, 115–119, https://doi.org/10.1126/science.1109496.

Teng, H., and G. Branstator, 2011: Initial-value predictability of prominent modes of North Pacific subsurface temperature in a CGCM. *Climate Dyn.*, **36**, 1813–1834, https://doi.org/10.1007/s00382-010-0749-7.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2009: Forced and internal twentieth-century SST trends in the North Atlantic. *J. Climate*, **22**, 1469–1481, https://doi.org/10.1175/2008JCLI2561.1.

Trenberth, K. E., and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, **33**, L12704, https://doi.org/10.1029/2006GL026894.

Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, https://doi.org/10.1098/RSIF.2013.1162.

Yeager, S. G., 2020: The abyssal origins of North Atlantic decadal predictability. *Climate Dyn.*, https://doi.org/10.1007/s00382-020-05382-4, in press.

——, and G. Danabasoglu, 2014: The origins of late-twentieth-century variations in the large-scale North Atlantic circulation. *J. Climate*, **27**, 3222–3247, https://doi.org/10.1175/JCLI-D-13-00125.1.

——, and J. I. Robson, 2017: Recent progress in understanding and predicting Atlantic decadal climate variability. *Curr. Climate Change Rep.*, **3**, 112–127, https://doi.org/10.1007/s40641-017-0064-z.

——, A. R. Karspeck, and G. Danabasoglu, 2015: Predicted slowdown in the rate of Atlantic sea ice loss. *Geophys. Res. Lett.*, **42**, 10 704–10 713, https://doi.org/10.1002/2015GL065364.

——, and Coauthors, 2018: Predicting near-term changes in the Earth system: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bull. Amer. Meteor. Soc.*, **99**, 1867–1886, https://doi.org/10.1175/BAMS-D-17-0098.1.