Earth and Space Science

RESEARCH ARTICLE
10.1029/2022EA002679

**Key Points:**

- LMR-LIM trained on paleoclimate-data exhibits greater low-frequency variability than a GCM-LIM trained on a long climate model simulation
- Both LMR-LIM and GCM-LIM exhibit forecast skill to 10-year lead
- LMR-LIM outperforms GCM-LIM for most out-of-sample forecasting experiments

**Supporting Information:**

Supporting Information may be found in the online version of this article.

**Correspondence to:**

L. M. Taylor,
ltaylo@uw.edu

**Citation:**

Taylor, L. M., & Hakim, G. J. (2023). Skillful coupled atmosphere-ocean forecasts on interannual to decadal timescales using a linear inverse model. *Earth and Space Science*, *10*, e2022EA002679. https://doi.org/10.1029/2022EA002679

Received 3 NOV 2022
Accepted 19 MAR 2023

**Author Contributions:**

**Conceptualization:** L. M. Taylor, G. J. Hakim
**Formal analysis:** L. M. Taylor, G. J. Hakim
**Funding acquisition:** G. J. Hakim
**Investigation:** L. M. Taylor, G. J. Hakim
**Methodology:** L. M. Taylor, G. J. Hakim
**Supervision:** G. J. Hakim
**Validation:** L. M. Taylor
**Visualization:** L. M. Taylor
**Writing – original draft:** L. M. Taylor
**Writing – review & editing:** L. M. Taylor, G. J. Hakim



# Skillful Coupled Atmosphere-Ocean Forecasts on Interannual to Decadal Timescales Using a Linear Inverse Model

**L. M. Taylor[1] and G. J. Hakim[1]**

[1]University of Washington, Seattle, WA, USA

**Abstract** There are two major challenges to improving interannual to decadal forecasts: (a) consistently initializing the coupled system so that variability is not dominated by initial imbalances, and (b) having a large sample of different initial conditions on which to test forecast skill. The second challenge requires consideration of time periods not only outside the recent period of intensive ocean observation, but also before the instrumental era, which increases the importance of the first challenge. Forecasts prior to the 1850s isolate internally generated sources of variability by removing the majority of anthropogenic forcing, and the sparse observational record during this time period motivates the use of paleoclimate proxy data. We address these issues by using a linear inverse model (LIM) approach and a recent proxy-based reconstruction over the last millennium at annual resolution. The reconstruction is used to train, initialize, and validate LIM forecasts. The LIM trained on paleo-data assimilated using a LIM trained on global climate model (GCM) simulation data outperforms a LIM trained on raw GCM data at forecast leads longer than 2 years for in-sample experiments, and beyond 4-year leads in most out-of-sample experiments validated on instrumental data. The most skillful normal mode of the paleo-data LIM for the instrumental experiment represents a persistent pattern with a longer decay time than for the GCM-LIM's modes, which accounts for the outperformance at longer leads. The paleo-data LIM is consequently more sensitive to ocean initialization, which is reflected in forecasts during the instrumental era where ocean reanalyses exhibit large uncertainty.

**Plain Language Summary** Forecasts on decadal timescales are important to understanding the predictability of natural climate variability prior to significant human influences. Improving forecasts on decadal timescales require long samples of global atmospheric and oceanic conditions, which have been limited and derive mainly from climate proxy data. Here, we use a forecast-model emulator that is trained, initialized, and validated on proxy-based climate reconstructions. We find this emulator to be skillful for 1- through 10-year forecasts, and also to outperform another emulator that is trained on a global climate model (GCM). The reconstruction-based emulator captures more longer-term climate behavior than the GCM-emulator and thus results in more skillful forecasts at longer leads. The reconstruction-based emulator is consequently more sensitive to how the ocean is initialized than the GCM-emulator.

## 1. Introduction

Decadal prediction lies between the two extremes of short and long-term forecasting and involves uncertainty tied to initial conditions (ICs), internal variability, and external forcings. Climate mitigation and adaptation strategies require accurate short-to-long-term climate change predictions, as changing temperatures, air quality, and precipitation directly impact agriculture, water security, and human health. However, uncertain, dynamically evolving climate states on these timescales hinder policy making (e.g., Füssel, 2007). While climate prediction on centennial timescales is mostly a boundary condition problem that depends on external forcing (e.g., climate-change projections; see Branstator & Teng, 2010), interannual to decadal prediction lies between short-term weather predictions and long-term climate change projections, and thus involves uncertainties tied to both initialization and boundary conditions (e.g., Meehl et al., 2009; Meehl et al., 2014; Meehl et al., 2021).

Distinguishing forced climate variability from internal variability is complicated by the short observational record, and different methods of separating the two tend to introduce biases (e.g., Frankcombe et al., 2015; Schurer et al., 2013). The sparse observational record, especially for the ocean prior to the late 20th century, leaves a small sample of dynamically consistent atmospheric and oceanic conditions to initialize and verify decadal forecasts from climate models. Furthermore, the short record fails to sample low-frequency variability on multidecadal timescales and has contributed to a lack of general understanding of internal variability (e.g.,

Trenberth et al., 2007). The ocean serves as a large source of decadal variability due to its high thermal inertia, so skillful interannual to decadal prediction (2–20 years) relies on proper initialization of the ocean state (Branstator & Teng, 2010, 2012). The inhomogeneous and sparse record of upper-ocean observational data prior to the Argo float implementation, as well as differences in model resolution and physics, introduce large uncertainties into historical estimates of ocean heat content (OHC) with significant differences between ocean reanalysis products (e.g., Palmer et al., 2017). Ocean analysis estimates of OHC through the later 20$^{th}$ century have large variations associated with time-dependent biases in observations (Carton & Santorelli, 2008). Additionally, model-based ocean reanalyses have large spread in their mean states that decreases over time until reaching a minimum value in the early 2000s when Argo floats were implemented (Xue et al., 2012). During the Argo era, ocean reanalyses were found to capture similar large-scale trends as seen in observations (Liao & Hoteit, 2022). However, coupled model simulations lack the ability to sufficiently represent deep ocean change which contributes to uncertainty in ocean reanalyses even during the Argo era (Storto et al., 2017, 2022). The uncertainty in the historical ocean record greatly impacts our ability to produce accurate coupled reanalysis data (Penny et al., 2019). Most reanalyses are uncoupled, and those that are coupled tend to use data assimilation (DA) methods that reconstruct the atmospheric and oceanic components independently ("weakly coupled" DA; see Penny et al., 2017), which does not impose coupled dynamical consistency between the analyses. For example, the Climate Forecast System currently operational at the National Centers for Environmental Prediction (NCEP) uses weakly coupled DA to create coupled reanalysis data (Saha et al., 2006, 2010, 2014).

Model initialization plays an important role in decadal forecasting, and there are various methods that have been investigated in the literature and applied operationally. Decadal forecasts are typically initialized by integrating the coupled model over surface fluxes defined by reanalysis fields (e.g., Yeager et al., 2018), where the fluxes in the model are set to equal those from the reanalysis. Additionally, one strategy to mitigate model bias is to constrain the model using observed anomalies (i.e., anomaly initialization). For example, the Met Office Hadley Centre Decadal Prediction System (DePreSys; Smith et al., 2007; DePreSys 2; Knight et al., 2014) uses anomaly initialization for decadal forecasts as opposed to full-field initialization, where model bias is removed during assimilation by applying an observational constraint. Anomaly initialization allows for a starting state that is more likely to be within the climate model's variability, but may deviate significantly from observations (e.g., Hazeleger et al., 2013). Observationally-based initialization for decadal forecasting is utilized within the Coupled Model Intercomparison Project Phase 5 and Phase 6 (CMIP5 and CMIP6; Taylor et al., 2012 and Eyring et al., 2016, respectively). These projects provide frameworks for comparing initialized forecasts to uninitialized, free-running integrations to assess the impact initialization has on decadal hindcasts. For example, Yeager et al. (2018) compared forecasts using the Community Earth System Model decadal prediction large ensemble (CESM-DPLE) to equivalent uninitialized historical simulations from the CESM Large Ensemble (CESM-LE) to quantify impacts of external forcing and initialization. The CESM-DPLE is composed of 40 ensemble members that are initialized from 1954 to 2015 using full-field initialization. They found that the CESM-DPLE had improved forecasts over the CESM-LE when external forcing was not a dominant driver.

An efficient alternative form of initialization that has been found to remove initialization shock uses a model-analog approach. For example, Ding et al. (2018) used model-analogs by initializing the forecast with a model state taken from a control simulation that was relatively close to the observed state at the time of interest. This model state serves as an "analog" to the actual initial state, and the forecast derives directly from the control simulation to generate ensembles in a computationally-efficient manner. Model-analogs created from a combined library of CMIP5 and CMIP6 simulations have been found to produce comparable skill to initialized and uninitialized forecasts for decadal forecasts of the North Atlantic subpolar gyre (Menary et al., 2021).

The short instrumental record provides an incomplete sample of the slow-varying components of the climate system over a limited period of time when the climate system was strongly forced. Longer samples, outside of the period of strong anthropogenic forcing from 1850 to 2000, necessitate climate proxies from ice cores, tree rings, and other geochemical and biological recorders. Dynamically consistent space-time gridded climate fields have recently been derived using paleoclimate DA. For example, the Last Millennium Reanalysis (LMR) project applied DA to paleoclimate reconstructions using an ensemble Kalman filter (Hakim et al., 2016; Tardif et al., 2019). Perkins and Hakim (2021a; PH21) used online DA with a linear inverse model (LIM) trained on CCSM4 to reconstruct coupled atmosphere-ocean fields over the last millennium (1000–2000 C.E.) at annual resolution. Compared to previous reconstructions using offline DA, these reconstructions exhibit better lead-lag

relationships (e.g., atmosphere-ocean coupling), enhanced decadal to centennial variability, and increased persistence of OHC.

Even with an extended observational record of coupled reanalysis data, running ensembles of fully coupled global climate models (GCMs) across multiple decades is extremely computationally expensive. Furthermore, models have inherent biases that corrupt the forecast of internal modes. For example, Farneti (2017) found that GCMs have little agreement on the spatial variance of internal variability such as the Pacific Decadal Oscillation (PDO), as well as varying power spectra and persistence, which yield different predictive ability across models. In addition, model prediction is sensitive to small perturbations in the ICs, which necessitates the use of large ensembles and compounds the computational expense of these forecasts (Meehl et al., 2014; Yeager et al., 2018). Skillful decadal prediction thus requires large samples of consistently coupled atmosphere-ocean ICs to properly estimate the climate system and to sample a wide range of different initial states of internal variability.

One way to circumvent high computational cost is to run decadal predictions at coarse resolution, a technique that has been applied using GCMs in CMIP5 and CMIP6. A more computationally efficient modeling approach, and the one adopted here, is to emulate GCMs through statistical methods. A widely-used empirically based emulator is a linear inverse model (LIM; e.g., Penland & Sardeshmukh, 1995), which is attractive due to its computational efficiency, distinct timescale separation, and flexibility of calibration. Previous studies have found skillful decadal forecasts of regional sea surface temperatures using LIMs (e.g., Foster et al., 2020; Hawkins & Sutton, 2009). The LIM has also proven to be a suitable benchmark on decadal timescales that exceeds persistence and has comparable skill to Phase 5 of the Coupled Model Intercomparison Project model hindcasts for annual global surface temperature forecasts (Newman, 2013). A LIM can thus be used in place of more comprehensive models for a low-cost alternative.

Here we address the limitations of forecasting on interannual to decadal timescales with a short observational record by using a multivariate LIM and the PH21 climate reconstruction (henceforth referred to as LMR data). The LMR data provides 1,000 years of coupled atmosphere-ocean global grids at annual resolution for initializing and verifying decadal forecasts. We investigate whether forecast skill can be improved by training a second LIM on the LMR data (hereafter, LMR-LIM). Thus, a comparative study between the LMR-LIM and the GCM-LIM used in the PH21 DA procedure provides insight into the forecast skill from dynamics learned empirically from proxy data, as compared to a GCM simulation. We anticipate that LMR-LIM skill will differ from the GCM-LIM due to climate variability learned from the proxy records.

The remainder of the paper is organized as follows. Section 2 details the methods and data used in this study. Section 3 presents the results of several forecasting experiments involving single-domain and coupled-domain forecasts. Section 4 explores potential sources of skill for a single forecast experiment in terms of the modes of the LIM. Finally, our conclusions are presented in Section 5.

## 2. Methods

### 2.1. Linear Inverse Modeling

A LIM approximates a nonlinear dynamical system as linear processes plus stochastic white-noise forcing:

$$\frac{d\boldsymbol{x}}{dt} = \mathbf{L}\mathbf{x} + \xi. \tag{1}$$

Here, the time tendency of the anomaly state vector, $\mathbf{x}$, taken about a reference mean climate, is represented by slow-varying climate dynamics, $\mathbf{L}\mathbf{x}$, plus fast timescale processes that are defined as stochastic noise forcing, $\xi$. The noise forcing is drawn from a normal distribution that is white in time yet may have structure in the spatial dimension. This distinct timescale separation allows for application of the central limit theorem and, thus, statistical closure (Hasselman, 1976).

Assuming constant $\mathbf{L}$, Equation 1 may be integrated to forecast the state at time $t + \tau$ in the future,

$$\mathbf{x}(t+\tau) = \mathbf{G}_\tau \mathbf{x}_t + \boldsymbol{\epsilon}, \tag{2}$$

where the propagation matrix, $\mathbf{G}_\tau$, is related to $\mathbf{L}$ by $\mathbf{G}_\tau = \exp(\mathbf{L}\tau)$. The LIM is empirically determined by solving for $\mathbf{G}_\tau$ based on the lag-covariance statistics of a sample of training data to minimize the error variance, $\boldsymbol{\epsilon}$, in Equation 2 (Penland, 1989):

$$\mathbf{G}_\tau = \mathbf{C}_\tau \mathbf{C}_0^{-1} = \left\langle \mathbf{x}(t+\tau)\mathbf{x}^T(t) \right\rangle \left( \left\langle \mathbf{x}(t)\mathbf{x}^T(t) \right\rangle \right)^{-1}. \tag{3}$$

Here, the angle brackets denote an expectation, taken as a sample average in time, and $\mathbf{C}$ is an autocovariance matrix.

The LIM assumes that the system has stationary statistics, meaning they are independent of time. This requires energy conservation via a balance equation described by the Fluctuation-Dissipation Relationship (FDR; Penland & Matrosova, 1994). The FDR maintains a stable system by establishing a statistical balance between the energy lost through the LIM's decaying modes and gained by the stochastic forcing:

$$\frac{d\mathbf{C}_o}{dt} = \mathbf{L}\mathbf{C}_o + \mathbf{C}_o\mathbf{L}^T + \mathbf{Q} = 0. \tag{4}$$

Here, $\mathbf{Q} = \left\langle \xi\xi^T \right\rangle$ is the stochastic noise covariance matrix.

We may assess the sources of skill in LIM forecasts by analyzing the LIM's empirical normal modes (ENMs). Empirical normal modes are solutions to the LIM's deterministic dynamics (Equation 1 without the forcing term) and take the general form $\mathbf{e}_j \exp(\lambda_j t)$, where $\mathbf{e}_j$ represents the $j$th eigenvector of $\mathbf{L}$ and $\lambda_j$ is the $j$th eigenvalue. The eigenvectors and eigenvalues are recovered via an eigendecomposition of the linear operator matrix, $\mathbf{L}$:

$$\mathbf{L} = \mathbf{E}\mathbf{\Lambda}_L\mathbf{E}^{-1}. \tag{5}$$

Here, $\mathbf{E}$ is a matrix with the eigenvectors of $\mathbf{L}$ as columns, and $\mathbf{\Lambda}_L$ is a diagonal matrix containing the eigenvalues of $\mathbf{L}$. The propagation matrix, $\mathbf{G}_\tau$, from Equation 2 shares the same eigenvectors as $\mathbf{L}$, and the eigenvalues are related by $G_{\lambda_j} = e^{L_{\lambda_j}\tau}$. Both the eigenvectors, $\mathbf{e_j}$, and eigenvalues, $\lambda_j$, of $\mathbf{L}$ are complex, and the eigenvectors may come in complex conjugate pairs.

Each ENM has two properties that are directly retrieved from the eigenvalues of $\mathbf{L}$: the decay time and the period. The e-folding decay time is calculated as $-\frac{1}{\sigma}$ and the period as $\frac{2\pi}{\omega}$, where the real and imaginary parts of the eigenvalue are $\sigma$ and $\omega$, respectively. Stationary statistics require $\sigma < 0$, so all modes decay with time. However, ENMs are non-orthogonal and may interfere in combination with one another, resulting in transient anomaly growth (decay) via constructive (destructive) interference between modes (e.g., Alexander et al., 2008; Farrell & Ioannou, 1995).

### 2.2. Data

We compare LIMs trained on data from two sources: a global climate model (GCM-LIM) and paleo-informed data (LMR-LIM). While both LIMs contain information about coupled dynamics, the latter is informed by indirect estimates of actual climate variability, as resolved by paleoclimate proxies. We provide information on the data used to train each LIM in Section 2.2.1, and Section 2.2.2 presents the data used for testing LIM performance.

#### 2.2.1. Training Data

The data used to train the GCM-LIM comes from the Community Climate System Model version 4 (CCSM4; see Gent et al., 2011) last millennium simulation (850–1850 C.E.), which includes forcing from land-use change, volcanic eruptions, and greenhouse gasses. The GCM simulation provides full-field coupled atmosphere-ocean variables at monthly resolution spanning the last millennium, which are averaged to annual resolution to match the timescale resolved by the LMR (and proxy data). Variables considered in the forecast experiments include surface air temperature, sea level pressure, 500 hPa geopotential height, precipitation, ocean surface height, ocean surface temperature, and upper 700 m OHC.

The LMR-LIM is trained on data provided by the LMR reconstruction, which uses the GCM-LIM to assimilate paleoclimate proxy data. These reconstruction data provide a similarly large sample of globally gridded variables to train and initialize the LMR-LIM. In addition, although the reconstruction has dynamical aspects of the reference model used in assimilation, such as lead-lag relationships between the atmosphere and ocean (PH21), coupled dynamics of the LMR-LIM are distinctly different from those of the GCM-LIM. The differences derive from the influence of proxies in the reconstruction, which record the actual climate variability that the GCM simulation is not constrained to capture.

#### 2.2.2. Verification Data

This study presents several out-of-sample forecasts on reanalysis datasets, that is, any data that is not included within the LIM training. We first verify on reanalysis data that independently represent two components of the

system: one for the atmosphere and for the ocean. The atmosphere-only data are taken from the 20th Century Reanalysis version 3c (20CR; Slivinski et al., 2019). The 20CR data set contains four annually averaged atmospheric variables: surface air temperature, sea level pressure, precipitation, and 500 hPa heights. These data are available from 1836 to 2015 C.E. We also verify forecasts on the GISS Surface Temperature Analysis (GISTEMP; Lenssen et al., 2019), available from 1880 to 2019 C.E., ocean data using the Simple Ocean Data Assimilation (SODA) data set (Carton et al., 2018) from 1871 to 2008 C.E., and coupled reanalyses. One independent out-of-sample data set used here is the LMR data not included within the LMR-LIM training, that is, from 1851 to 2000 C.E. Another data set used for verification experiments is a coupled historical simulation from GISS ModelE2 with specified forcing from greenhouse gases and aerosols from 1850 to 2005 C.E (GISS-E2-R; Schmidt et al., 2014). This specific simulation is chosen for verification because all variables within the historical data set are the same as those within the training datasets and, thus, provides an opportunity to evaluate fully coupled out-of-sample forecasts, even if they do not exactly match observations over this time period.

#### 2.2.3. Data Processing

Both the training and verification datasets follow the same data processing steps. We first use bilinear interpolation to convert data to a common 2° by 2° latitude-longitude grid. The data set is then truncated to the time interval of interest and annually averaged. The annual data is then converted to gridded anomalies by removing the time mean at each grid point. We remove the linear trend from the data across the entire time interval by performing linear regression and subtracting the fit from the data. Trends are not captured by LIM dynamics, which assume stationary statistics. As a consequence, explicit modeling of anthropogenic trends is outside the scope of the LIMs used in this study.

### 2.3. LIM Calibration

Both LIMs are trained on a time series of coupled atmosphere-ocean data. LIM training employs a two-step EOF truncation that greatly reduces the data's dimensionality (Perkins & Hakim, 2020; PH20). The first step truncates each field to a smaller set of EOFs while the second step further compresses the data by truncating to a set of EOFs for the coupled covariance matrix, which removes collinearity among the predictor fields. Complete details of the two-step calibration method can be found in PH20 (Their Appendix A).

Training data are converted to a compressed state prior to calibration by applying the two-step EOF reduction. The number of EOFs included for each LIM depends upon the degrees of freedom in the calibration data set. For the GCM data, the first truncation retains above 90% of each variable's variance with 400 EOFs. The second truncation is set at 25 multivariate EOFs and retains around 75% of the variance (i.e., the shared variance between all variables); adding additional EOFs adds only a small percentage of the coupled variance (PH20). The LMR-LIM has fewer degrees of freedom than the GCM-LIM, most likely due to the limited dimensionality of the proxies and the 100-member ensemble used in the assimilation algorithm. For the LMR-LIM, the first step of the EOF reduction retains above 90% of the variance in each variable field with 15 variable EOFs, and the second reduction retains just above 90% of the combined field variance with 10 multivariate EOFs. When OHC is included in the LIMs, we retain 20 OHC EOFs as a separated field not subject to the second EOF truncation (PH20 find this is superior to including OHC with other variables) to capture more than 70% of the CCSM4 OHC variance and 10 OHC EOFs to capture about 94% of the LMR OHC variance. PH20 show 20 EOFs to be the optimal truncation for capturing OHC variability; retaining additional OHC EOFs does not capture significantly more variance. Sensitivity tests to the number of EOFs retained are discussed in Section 3.1.

### 2.4. Forecast Experiments

We perform several experiments to compare the performance of the LMR-LIM (i.e., trained on LMR data from 1000 to 1850) to a GCM-LIM (i.e., trained on CCSM4 data from 850 to 1850). We note that the variables used to train the LMR-LIM and GCM-LIM are constrained by those available withing the verification data set. The training variables thus vary by experiment, but the training datasets and times are consistent across all experiments and are as described in Section 2.2.1.

Table 1 summarizes all 5 experiments presented in Section 3. Experiments 1 and 2 are out-of-sample uncoupled forecasts on reanalysis datasets, where experiment 1 is atmosphere-only and verified on 20CR and experiment 2 is ocean-only and verified on SODA. The next 3 experiments are all coupled. Experiment 3 is initialized and

**Table 1**
*Experiment Design for Comparative Linear Inverse Model Study Presented in Section 3*

| Experiment | Verification dataset | Verification time | Variables (training and verification) | Classification |
|---|---|---|---|---|
| 1 | 20th Century Reanalysis (20CR) | 1851–2015 | TAS, PSL, PR, & ZG500 | Out-of-sample |
| 2 | Simple Ocean Data Assimilation (SODA) | 1871–2008 | SST & OHC | Out-of-sample |
| 3 | a) Last Millennium Reanalysis (LMR) | a) 1000–1850 | a & b) TAS, PSL, PR, ZG500, SST, ZOS, & OHC | a & b) In-sample |
| | b) Community Climate System Model version 4 (CCSM4) | b) 850–1850 | | |
| 4 | GISS ModelE2 (GISS-E2-R) | 1851–2005 | TAS, PSL, PR, ZG500, SST, ZOS, & OHC | Out-of-sample |
| 5 | Last Millennium Reanalysis (LMR) | 1851–2000 | TAS, PSL, PR, ZG500, SST, ZOS, & OHC | Out-of-sample |

*Note*. Shaded rows represent single-component experiments (i.e., uncoupled). Variables used across experiments include surface air temperature (TAS), surface level pressure (PSL), precipitation (PR), ocean surface temperature (SST), 500 hPa heights (ZG500), dynamic ocean surface heights (ZOS), and ocean heat content.

verified on in-sample data, i.e., those data used to train each LIM. Thus, experiment 3 has two additional experiments: 3a) LMR-LIM verified on its training data and 3b) GCM-LIM verified on its training data (see Table 1). Experiment 4 is out-of-sample coupled forecasts on the LMR data withheld from training, and experiment 5 is out-of-sample coupled forecasts on GISS-E2-R.

Forecast skill is evaluated over time using the anomaly correlation coefficient (ACC) and the standardized error variance (SEV). Both metrics are calculated at each grid point and either presented as spatial fields or area-weighted global means. The SEV metric is calculated after converting the forecast from the reduced LIM space into full lat-lon space; thus, all forecasts have truncation error. The global-mean SEV is calculated by taking the squared error at each grid point, standardizing by the variance of the verification data at that grid point, and then globally averaging. SEV calculations have 95% confidence bounds determined via bootstrap resampling based on 1,000 iterations that randomly sample 75% of the available forecast data with replacement. See Appendix A for further details on the ACC and SEV.

## 3. Forecast Results

### 3.1. Single-Component LIMs

Gridded samples of coupled reanalysis data are not available due to the lack of observations prior to 1850. Even after 1850, most reanalysis systems analyze the atmosphere and ocean independently, limiting opportunities for coupled forecast-verification experiments. Therefore, we first consider single component experiments (i.e., either atmosphere-only or ocean-only systems; Table 1 shaded rows) that initialize and verify forecasts based on reanalysis data, which is available after about 1850.

#### 3.1.1. Atmosphere-Only LIMs

Experiment 1 applies LIMs that are only trained on atmospheric variables (2 m air temperature, precipitation, sea level pressure, and 500 hPa heights; see Experiment 1 in Table 1). Forecasts are initialized and verified using data from 20CR during 1851–2015. We note that both LIMs perform better on the latter half of 20CR (1934–2015) when there are more observations to inform the reanalyses, yet the sample size is too small for either model to produce forecasts statistically different from climatology (not shown). We also conduct additional experiments using LIMs trained on only 2 m air temperature that are initialized/verified on GISTEMP data (see Figure S1 in Supporting Information S1), and the results are qualitatively similar to those shown subsequently. We choose to present only the 2 m air temperature forecast skill in the atmospheric forecast results, as it is representative of all atmospheric variables, and the most widely observed variable across the instrumental period.

The forecast results in Figure 1a show that the GCM-LIM performs better at leads less than 4 years, whereas the LMR-LIM is marginally better for leads longer than 5 years. Spatial skill at the 4-year lead, where globally averaged skill is comparable between the LIMs, reveals similar spatial patterns of forecast skill. Both frameworks share poor regional skill over parts of the Southern Ocean, with the LMR-LIM forecasts having more negative correlations; however, we note that this region is also poorly constrained in 20CR due to limited pressure observations and uncertain SST boundary conditions. Both LIMs also have similar areas of higher skill over the tropics, especially near the Indian Ocean.

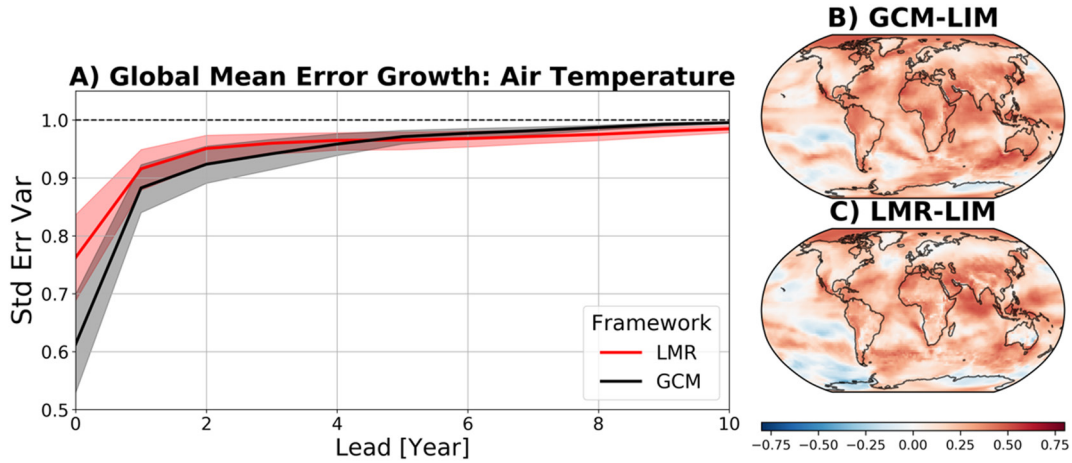


**Figure 1.** (a) Global mean of standardized error variance of surface air temperature for out-of-sample forecasts on 20CR data (1851–2015 C.E.). GCM-LIM forecasts are in black and LMR-LIM forecasts are in red. Solid lines represent the sample-mean linear inverse model forecasts with shading representing the 95% confidence range. Panels on the right show the 4-year-lead surface air temperature anomaly correlation coefficient for forecasts from the GCM-LIM (b) and the LMR-LIM (c).

### 3.1.2. Ocean-Only LIMs

We next consider companion experiments to those in the previous section by performing forecasting experiments for LIMs trained only on ocean variables (i.e., surface temperature and OHC). There exist a wide range of reanalysis products for forecast initialization and verification, including SODA (Carton et al., 2018), HadleyEN4 (Good et al., 2013), GECCO3 (Kohl, 2020), and ORAS4 (Balmaseda et al., 2013). Here, we conduct forecast experiments using SODA, as this data set provides the longest sample to initialize and verify forecasts of SST and OHC, spanning 1871–2008 C.E (see Experiment 2 in Table 1).

Skill for both LIMs is limited to mainly the first 2 years, with the GCM-LIM performing better than the LMR-LIM (Figure 2a). Spatial skill at 1-year lead shows that the GCM-LIM forecasts have higher correlation than LMR-LIM forecasts, most notably in the tropical Pacific and Atlantic, with a global-mean correlation coefficient of 0.40 compared to 0.30 for the LMR-LIM. We also test LIM performance on ocean data provided by several other data sources (see Text S1 in Supporting Information S1). All results are qualitatively similar for the

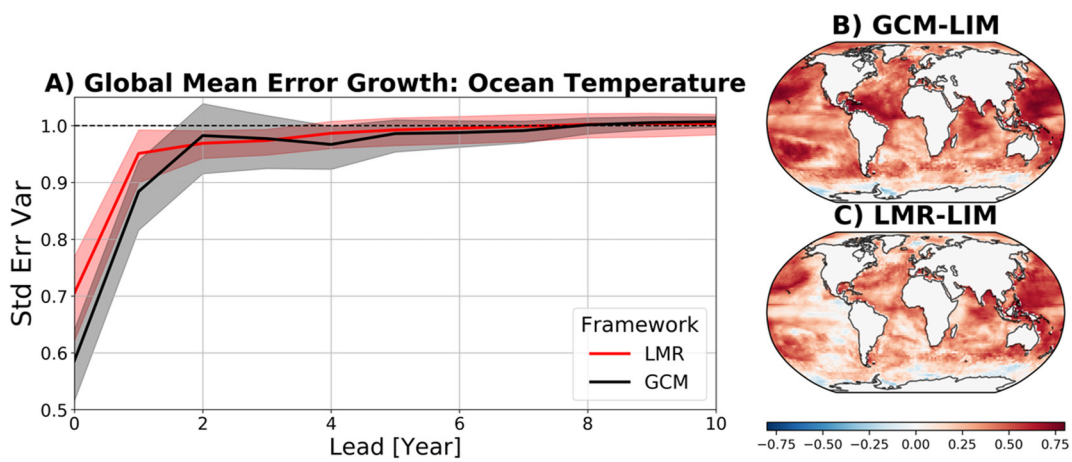


**Figure 2.** (a) Global mean of standardized error variance of sea surface temperatures for out-of-sample forecasts on ocean-only data provided by the Simple Ocean Data Assimilation data set (1871–2008 C.E.). GCM-LIM forecasts are in black and LMR-LIM forecasts are in red. Solid lines represent the sample-mean linear inverse model forecasts with shading representing the 95% confidence range. Panels on the right show the 1-year-lead sea surface temperature anomaly correlation coefficient for forecasts from the GCM-LIM (b) and the LMR-LIM (c).

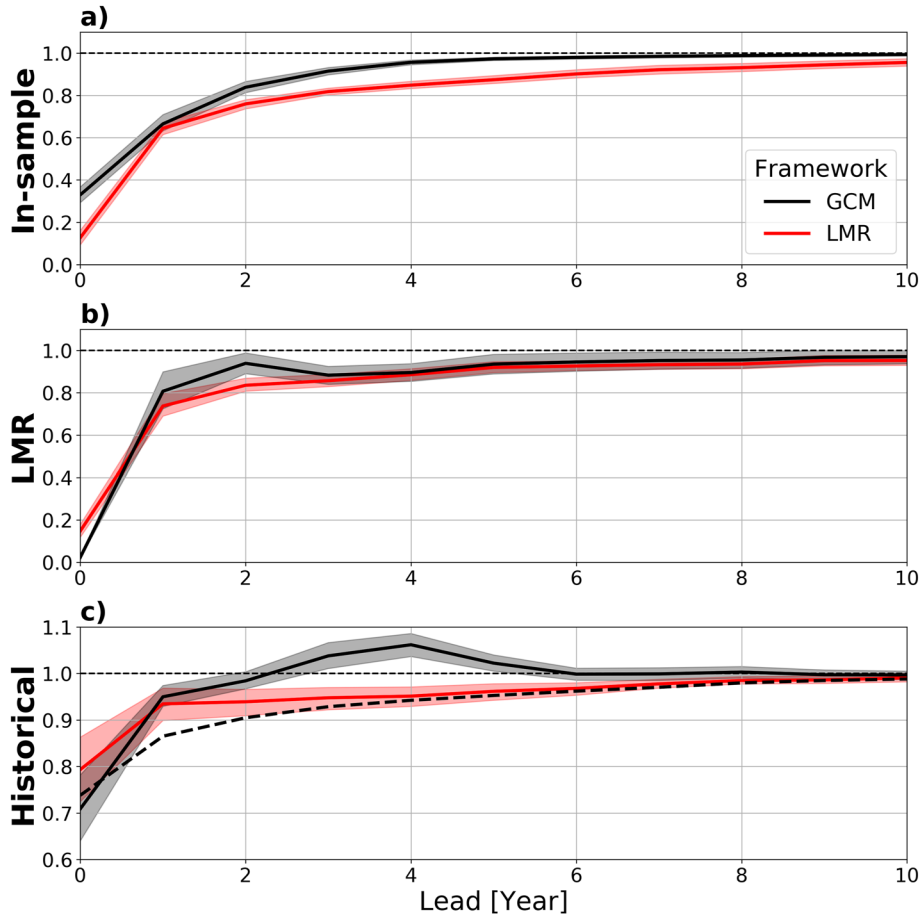


**Figure 3.** Global mean of the air temperature standardized error variance of the coupled LMR-LIM (red) and the GCM-LIM (black) forecasts on (a) in-sample data (b) the withheld LMR-data from 1851 to 2000 C.E., and (c) GISS-E2-R 1851–2005 C.E. Solid lines represent the ensemble-mean linear inverse model forecasts with shading representing the 95% confidence range. Also shown is GCM-LIM truncated at the same resolution as the LMR-LIM (dash-dot black line) for the historical simulation experiment.

GCM-LIM, but the LMR-LIM exhibits greater variability in skill, possibly due to sensitivity to OHC ICs, which is explored further below. We note that ocean-only reanalyses differ significantly in SST and OHC estimates prior to the late 20th century due to sparse and unreliable ocean measurements (e.g., Palmer et al., 2017). The correlation between global-mean SST (OHC) in SODA and HadleyEN4 is 0.61 (0.64) for the overlapping time period from 1900 to 2008, as compared to 0.84 (0.80) during 1950–2008. Thus, the uncertain analyses are reflected in less-skillful forecasts by both LIMs.

### 3.2. Coupled LIMs

We now move on to study the results of LIMs that are trained on both atmospheric and oceanic variables, that is, coupled LIMs (Table 1 Experiments 3–5). All LIMs for these experiments are trained and verified on 2 m air temperature, surface level pressure, 500 hPa heights, precipitation, 0–700 m OHC, dynamic ocean height, and sea surface temperature.

The first coupled experiment presents a comparison between the in-sample performance for each LIM. Thus, the LMR-LIM and GCM-LIM are verified on their respective training data (See Table 1; Experiment 3a and 3b). Results show that the LMR-LIM has more in-sample skill than the GCM-LIM across all leads (Figures 3a, 4a, and 4b). LMR-LIM also outperforms the GCM-LIM for the El Nino Southern Oscillation, the PDO, and the North Pacific indices (ENSO, PDO, and NPI; see Text S2 in Supporting Information S1). The 4-year-lead spatial correlation of air temperature forecasts shows that LMR-LIM forecasts have more skill than the GCM-LIM

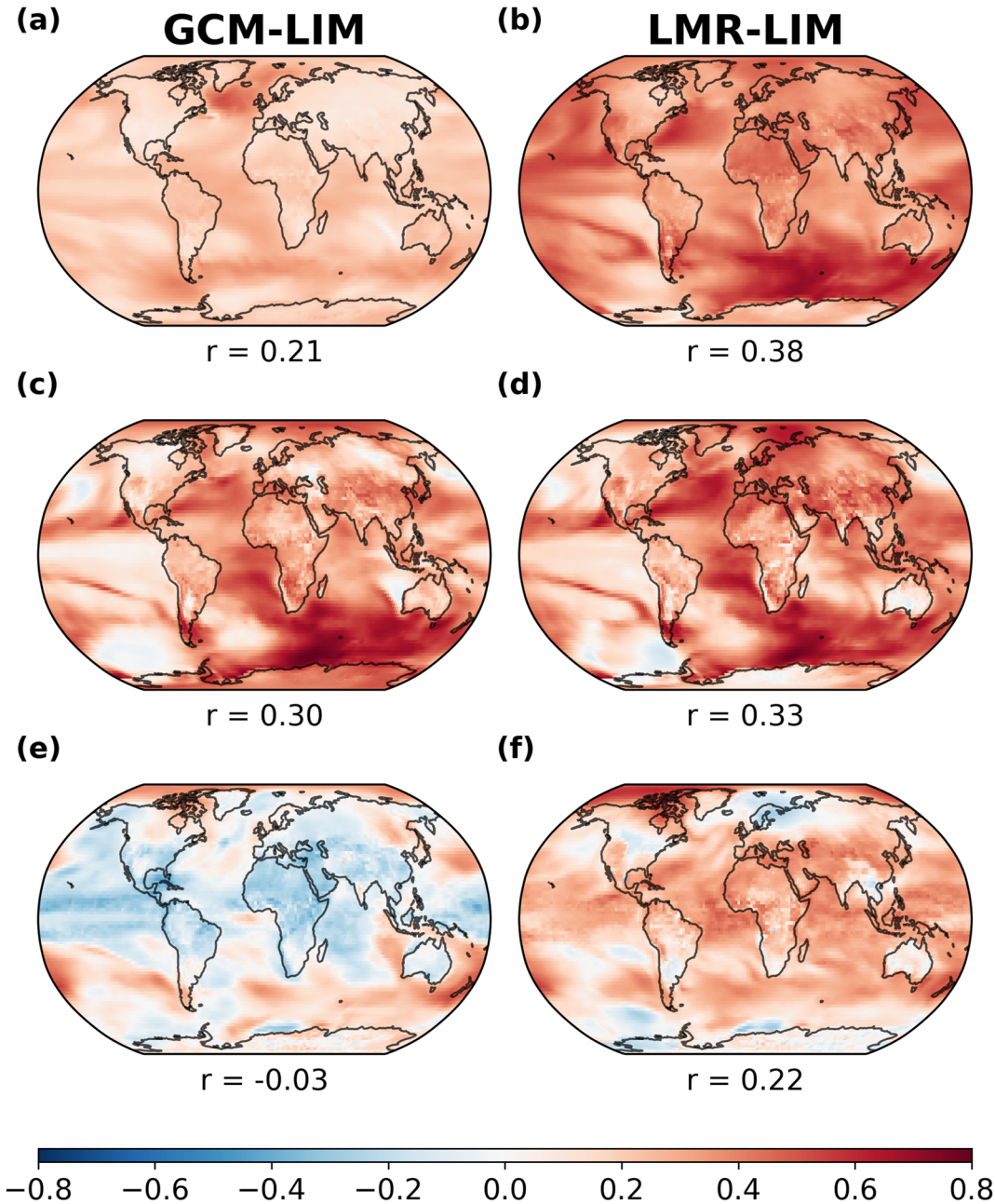


**Figure 4.** Average anomaly correlation coefficient values for 4-year lead air temperature forecasts provided by the GCM-LIM (left column) and the LMR-LIM (right column). Forecasts are verified on (a, b) in-sample data, (c, d) the withheld LMR-data from 1851 to 2000 C.E., and (e, f) GISS-E2-R. The global mean correlation coefficient is displayed beneath each panel.

forecasts nearly everywhere with a global-mean correlation coefficient of 0.38 as compared to 0.21 (Figures 4a and 4b).

We consider now two out-of-sample experiments to test and compare LIM performance (Table 1 Experiments 4 and 5). Experiment 4 uses the withheld LMR data for ICs and verification (Figures 3b, 4c and 4d). That is, the LMR-LIM is trained on the LMR-data from 1000 to 1850 C.E., and we use the remaining 150 years of LMR data from 1851 to 2000 C.E. for ICs and forecast verification for both the LMR-LIM and GCM-LIM. We remind the reader that the LMR data was determined based on DA using the GCM-LIM and differences in these LIMs are due to proxy DA.

The air temperature forecasts produced by the LMR-LIM are more skillful than the GCM-LIM for leads less than 3 years, where the GCM-LIM forecast skill is affected by the quasi-periodic ENSO behavior of the parent model (Figure 3b). The LIMs perform similarly beyond 3-year leads, with slightly lower, but not statistically significant, error for the LMR-LIM. Figures 4c and 4d show very similar spatial performance for both LIMs. Both models have high skill over the Southern Ocean and over the Indian Ocean, while the LMR-LIM outperforms the GCM-LIM over continental regions and most ocean basins. The LMR-LIM forecasts have slightly higher global-mean correlation of 0.33 compared to the GCM-LIM's 0.30 global mean. The skill difference in the global-mean correlation increases at longer lead years, for example, at 5-year lead, LMR-LIM at 0.28 as compared to the GCM-LIM at 0.22. This implies that the LMR-LIM forecasts have better timing than GCM-LIM forecasts, although skill in anomaly amplitude forecasts are little improved as measured by SEV (Figure 3b).

Experiment 5 is the final out-of-sample experiment and involves a forecast on a coupled historical simulation (Figures 3c, 4e, and 4f). Specifically, this experiment compares the LIM performance between frameworks for forecasts on a single historical simulation provided by the GISS-E2-R model. The results show that the LMR-LIM outperforms the GCM-LIM across all leads (Figure 3c), with the GCM-LIM losing all skill after 2 years; the LMR-LIM retains skill to 10-year leads. The GCM-LIM error exceeds the climatological variance in the historical simulation due to the regular ENSO behavior in the parent model, which is also noted by Perkins and Hakim (2020). Also shown is the result for a GCM-LIM that has the same truncation as the LMR-LIM (i.e., both LIMs have 20 total degrees of freedom) (Figure 3c, dash-dot line). Truncating the GCM-LIM to a smaller basis results in very different results from the original GCM-LIM, including modulating the regular ENSO signal of the parent model, which significantly improves the GCM-LIM's forecasts, such that they are better than the LMR-LIM. We note that this is the only experiment that the equally-truncated GCM-LIM showed significant improvements over the normal truncation (see Text S3 in Supporting Information S1).

The spatial performance at 4-year leads for the historical simulation experiment reflects the influence of different ENSO dynamics for the normal-truncation GCM-LIM (Figure 4e), with low correlation values across the tropics, which also extend into the extratropical regions, especially over land. The global-mean correlation captures the poor performance of the GCM-LIM with a value of −0.03, yet increases to 0.24 for the equal-truncation LIM. Figure 4f shows that the LMR-LIM not only has better skill in the tropics, but also improved forecasts over most of the continents, with a 0.22 global-mean correlation coefficient. We speculate that the continental air temperature skill may be due to the large number of proxies in these locations.

## 4. Sources of Forecast Skill

The LMR-LIM most notably outperforms the GCM-LIM for the in-sample coupled experiment. To begin to understand the reasons for this difference, we first decompose the LMR-LIM and GCM-LIM forecasts into contributions from individual variables. Specifically, we repeat the coupled in-sample experiment, but show the global-mean error growth for all variables as opposed to only air temperature as in Figure 3a. Additionally, we evaluate forecast sensitivity to variables included in the LIM by removing a single variable, retraining the model, and repeating the in-sample forecast experiments.

Comparing the global-mean error for all variables shows, as expected, that OHC is the most skillful variable for both frameworks (Figures 5a and 5b). SST and 2 m air temperature are most skillful for the LMR-LIM, and ocean-surface height and SST are most skillful for the GCM-LIM. Sea-level pressure and precipitation are the least skillful for both LIMs. Removing OHC from the state vector, the GCM-LIM loses skill by 6-year lead, similar to the control case (cf. Figures 5b and 5d). The LMR-LIM exhibits greater sensitivity to OHC, with larger error growth relative to the control than for the GCM-LIM, but also retains small skill beyond 6 years. These experiments reveal that the LMR-LIM has better in-sample forecasts than the GCM-LIM as well as increased skill of all other variables when OHC is part of the LIM state vector. The change in skill is smaller for all other variables when OHC is removed from the state vector, including sea surface temperature and ocean surface height.

OHC is a key contributor to differences in performance between the LMR-LIM and GCM-LIM. We now show why OHC results in these differences by using the ENMs of the LIMs as a basis for understanding sources of forecast skill. We focus on the out-of-sample experiment from Section 3.2 on LMR data during 1851–2000 (Figures 3b, 4c, and 4d) because these experiments pertain to fully coupled forecasts, and forecasts for both

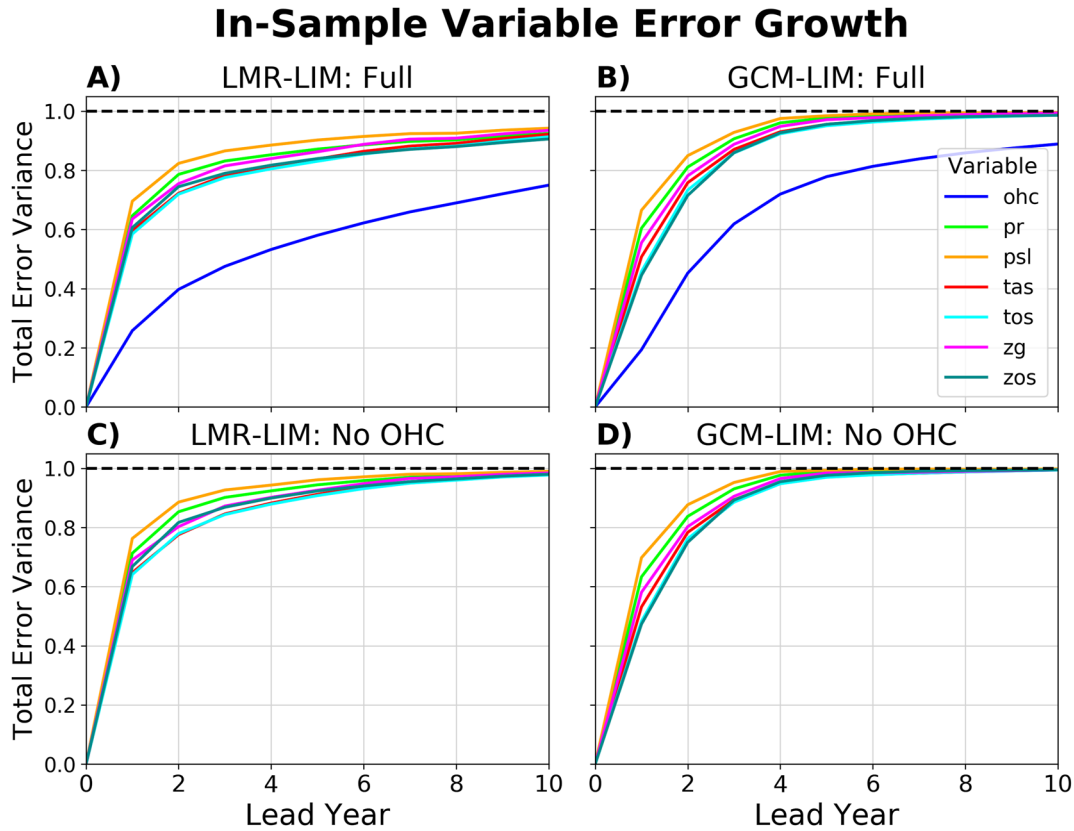


**Figure 5.** Global-mean standardized error variance (i.e., squared error at each grid point and globally averaged) for each variable in the LMR-LIM (left) and GCM-LIM (right) in-sample forecasts. Bottom row is the same as the top row but without ocean heat content (OHC) in the linear inverse model. Variables include OHC (ohc), precipitation (pr), surface level pressure (psl), surface air temperature (tas), ocean surface temperature (tos), 500 hPa heights (zg), and dynamic ocean surface heights (zos). All errors are normalized by the climatological variance for that variable.

LIMs are skillful to 10-year leads. We repeat this experiment but for single ENMs by projecting the ICs onto a single ENM and assessing forecast skill. We focus in particular on the ENMs that have the most forecast skill. For the LMR-LIM, the two most-skillful ENMs capture about half of the skill of the full model (Figure 6, left

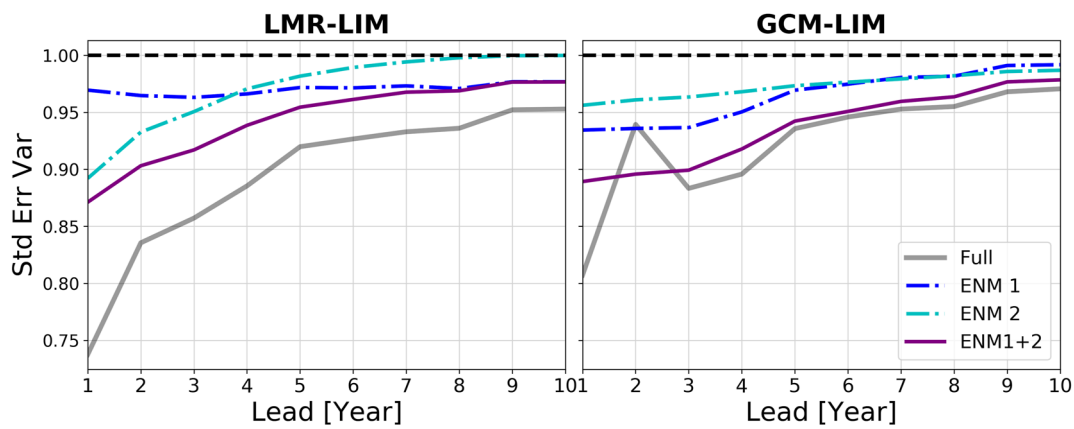


**Figure 6.** Global mean of air temperature standardized error variance for empirical normal mode (ENM) experiments using the LMR-LIM (left) and GCM-LIM (right) to forecast on Last Millennium Reanalysis data 1851–2000. Forecast pertain to the most skillful single eigenvector (ENM 1; dashed blue), the second most skillful single eigenvector (ENM 2; dashed cyan), and the sum of the two most-skillful ENMs (ENM1 + ENM2; solid purple); the full ENM forecast is shown in gray.

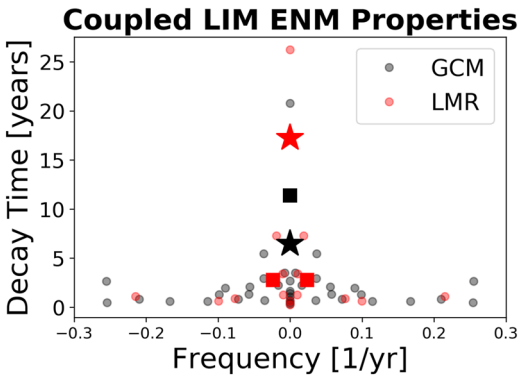


**Figure 7.** Empirical normal mode (ENM) properties of the LMR-LIM (red) and GCM-LIM (black). For both frameworks, the most skillful ENM found from single ENM experiments (ENM 1) is marked as a star, and the second most skillful ENM (ENM 2) is marked as a square.

panel); adding a third ENM adds only small additional skill (not shown). This implies that interactions between multiple ENMs in the LMR-LIM contributes to forecast skill of the full model, especially at longer leads. In contrast, for the GCM-LIM, the full forecast is nearly approximated by two ENMs, especially for forecasts past 4 years (Figure 6, right panel).

We next examine the decay time (measured by the e-folding time) and period of all ENMs to gain an understanding of how the ENMs differ between the coupled frameworks (Figure 7). In both frameworks, the least-damped modes are stationary, and the two least damped modes for LMR-LIM decay more slowly than those in GCM-LIM. For the LMR-LIM, the second least-damped mode is the most skillful ENM, with a decay time of 17.3 years, whereas for the GCM-LIM, the third least-damped mode is most skillful, with a decay time of 6.5 years (stars in Figure 7). Interestingly, the least damped modes were not the most skillful in this experiment, possibly because they are associated with longer timescales than are available for verification. The LMR-LIM's second most skillful mode has a period of about 42 years and a decay time of about 12 years, as compared to a stationary mode for the GCM-LIM with a decay time of about 3 years (Figure 7, squares). Overall, the LMR-LIM forecast skill appears to derive from a broader range of ENMs compared to the GCM-LIM, and since the LMR-LIM's least-damped modes decay slower than for the GCM-LIM, the LMR-LIM forecast skill at longer leads is more sensitive to the projection of ICs onto these ENMs.

## 5. Conclusions

We test the use of a LIM as a low-dimensional approximation to coupled atmosphere-ocean climate dynamics for use in forecasting experiments on interannual to decadal timescales. We use two LIMs, one derived from a coupled climate model (GCM), and one from a recent climate reconstruction of the last millennium based on assimilating paleoclimate proxy data (LMR). Both LIMs are skillful for a wide range of in- and out-of-sample forecasts on interannual to decadal timescales. We also show that the LIM calibrated on the proxy-informed reanalysis data (LMR) exhibits better forecast skill in some experiments, evidently because the proxies result in ENMs that are more persistent in the LIM dynamics. Conversely, the proxy-informed LIM performs worse in experiments where that increased persistence is inconsistent with the data used for initializing and verifying forecasts.

The LMR data set is one of the few fully coupled datasets available for model initialization and calibration on multi-centennial timescales. The single-component LIMs trained on just the atmosphere show the LMR-LIM performs similarly to the GCM-LIM up to 5-year leads, beyond which the LMR-LIM outperforms. The single-component LIMs on ocean-only data show both the GCM-LIM and LMR-LIM performing similarly across all leads with large uncertainty from ocean reanalysis products. Finally, the coupled LIMs trained on LMR data have better in-sample performance than the GCM-LIM, and out-of-sample experiments show similar performance between the two frameworks across all leads (when considering an equally-truncated GCM-LIM for the Historical experiment). The LMR-LIM skill for all experiments is more sensitive to ocean fields, especially OHC. Removing OHC from the LMR-LIM results in increased error growth for all variables relative to the GCM-LIM. Enhanced persistence in the leading ENMs of the LMR-LIM appears to derive from OHC, and results in forecasts at long leads being more sensitive to the projection of ICs and verification data onto these modes relative to the GCM-LIM. The enhanced persistence in the LMR-LIM ties back to the LMR-data itself, which has enhanced low-frequency variability and atmosphere-ocean lead-lag relationships as compared to offline reconstructions (Perkins & Hakim, 2021a).

All forecast experiments highlight the need for large samples of consistently coupled fields for initializing and evaluating forecasts on decadal timescales. Coupled climate field reconstructions are needed for decadal predictability studies, as they provide full-fields of atmosphere-ocean data that capture the uncertainty in ICs, as well as offer continuous samples of data to initialize models and verify forecasts. In particular, future work that adds seasonality to these reconstructions would allow forecast skill to be evaluated at finer time resolution.

## Appendix A: Skill Metrics

The correlation between the forecast anomalies, $x$, and observed anomalies, $y$, is calculated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{A1}$$

where summations are taken over select forecast leads. The correlation coefficient is important in diagnosing the phasing between the observations and forecast, but does not provide information on errors in the amplitude and bias.

The standardized error variance (SEV) takes into account amplitude information and biases between the forecast and verification fields, and is calculated at a single grid point by:

$$\mathrm{SEV} = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n-1} * \left(\frac{\sum_{i=1}^{n}(y_i)^2}{n-1}\right)^{-1}. \tag{A2}$$

Here, $n$ represents a summation over all forecasts for the selected lead. After calculating the SEV at each grid point, we take the area-weighted global mean for each forecast lead time. The final metric thus represents the global mean of the SEV field.

## Data Availability Statement

The code used to create a LIM can be found at https://doi.org/10.5281/zenodo.3243749 (Perkins, 2019). The LMR reconstruction data used to train the LMR-LIM can be found at https://doi.org/10.5281/zenodo.4626197 (Perkins & Hakim, 2021b).

**Acknowledgments**
This paper derives from the first author's Master's Thesis at the University of Washington. Funding was provided by NSF Grant 2002276 and NOAA Grant NA20NWS4680053 awarded to the University of Washington. We thank Andre Perkins for code development and discussion. Additionally, we thank Robert Tardif for assistance with data preparation.

## References

Alexander, M. A., Matrosova, L., Penland, C., Scott, J. D., & Chang, P. (2008). Forecasting Pacific SSTs: Linear inverse model predictions of the PDO. *Journal of Climate*, *21*(2), 385–402. https://doi.org/10.1175/2007JCLI1849.1

Balmaseda, M. A., Trenberth, K. E., & Kallen, E. (2013). Distinctive climate signals in reanalysis of global ocean heat content. *Geophysical Research Letters*, *40*(9), 1754–1759. https://doi.org/10.1002/grl.50382

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

Branstator, G., & Teng, H. (2012). Potential impact of initialization on decadal predictions as assessed for CMIP5 models. *Geophysical Research Letters*, *39*(12), L12703. https://doi.org/10.1029/2012GL051974

Carton, J. A., Chepurin, G. A., & Chen, L. (2018). SODA3: A new ocean climate reanalysis. *Journal of Climate*, *31*(17), 6967–6983. https://doi.org/10.1175/JCLI-D-18-0149.1

Carton, J. A., & Santorelli, A. (2008). Global decadal upper-ocean heat content as viewed in nine analyses. *Journal of Climate*, *21*(22), 6015–6035. https://doi.org/10.1175/2008JCLI2489.1

Ding, H., Newman, M., Alexander, M. A., & Wittenberg, A. T. (2018). Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs. *Journal of Climate*, *31*(14), 5437–5459. https://doi.org/10.1175/JCLI-D-17-0661.1

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Farneti, R. (2017). Modelling interdecadal climate variability and the role of the ocean. *WIREs Climate Change*, *8*(1). https://doi.org/10.1002/wcc.441

Farrell, B. F., & Ioannou, P. J. (1995). Stochastic dynamics of the midlatitude atmospheric jet. *Journal of the Atmospheric Sciences*, *52*(10), 1642–1656. https://doi.org/10.1175/1520-0469(1995)052%3C1642:SDOTMA%3E2.0.CO;2

Foster, D., Comeau, D., & Urban, N. M. (2020). A Bayesian approach to regional decadal predictability: Sparse parameter estimation in high-dimensional linear inverse models of high-latitude sea surface temperature variability. *Journal of Climate*, *33*(14), 6065–6081. https://doi.org/10.1175/JCLI-D-19-0769.1

Frankcombe, L. M., England, M. H., Mann, M. E., & Steinman, C. A. (2015). Separating internal variability from the externally forced climate response. *Journal of Climate*, *28*(20), 8184–8202. https://doi.org/10.1175/JCLI-D-15-0069.1

Füssel, H. M. (2007). Adaptation planning for climate change: Concepts, assessment approaches, and key lessons. *Sustainability Science*, *2*, 265–275. https://doi.org/10.1007/s11625-007-0032-y

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jyane, S. R., et al. (2011). The Community Climate System Model version 4. *Journal of Climate*, *24*(19), 4973–4991. https://doi.org/10.1175/2011JCLI4083.1

Good, S. A., Martin, M. J., & Rayner, N. A. (2013). EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, *118*(12), 6704–6716. https://doi.org/10.1002/2013JC009067
Hakim, G., Emile-Geay, J., Steig, E., Noone, D., Anderson, D., Tardif, R., et al. (2016). The last millennium climate reanalysis project: Framework and first results. *Journal of Geophysical Research*, *121*(12), 6745–6764. https://doi.org/10.1002/2016JD024751
Hasselmann, K. (1976). Stochastic climate models Part 1. Theory. *Tellus*, *28*(6), 473–485. https://doi.org/10.1111/j.2153-3490.1976.tb00696.x
Hawkins, E., & Sutton, R. (2009). Decadal predictability of the Atlantic Ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling. *Journal of Climate*, *22*(14), 3960–3978. https://doi.org/10.1175/2009BAMS2607.1
Hazeleger, W., Guemas, V., Wouters, B., Corti, S., Andreu-Burillo, I., Doblas-Reyes, F. J., et al. (2013). Multi-year climate predictions using two initialization strategies. *Geophysical Research Letters*, *40*(9), 1794–1798. https://doi.org/10.1002/grl.50355
Knight, J. R., Andrews, M. B., Smith, D. M., Colman, A. W., Dunstone, N. J., Eade, R., et al. (2014). Predictions of climate several years ahead using an improved decadal prediction system. *Journal of Climate*, *27*(20), 7550–7567. https://doi.org/10.1175/JCLI-D-14-00069.1
Kohl, A. (2020). Evaluating the GECCO3 1948–2018 ocean synthesis—A configuration for initializing the MPI-ESM climate model. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 2250–2273. https://doi.org/10.1002/qj.3790
Lenssen, N., Schmidt, G., Hansen, J., Menne, M., Persin, A., Ruedy, R., & Zyss, D. (2019). Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6307–6326. https://doi.org/10.1029/2018JD029522
Liao, F., & Hoteit, I. (2022). A comparative study of the Argo-era Ocean heat content among four different types of data sets. *Earth's Future*, *10*(9), e2021EF002532. https://doi.org/10.1029/2021EF002532
Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., et al. (2014). Decadal climate prediction: An update from the trenches. *Bulletin America Meteorology Social*, *95*(2), 243–267. https://doi.org/10.1175/BAMS-D-12-00241.1
Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., et al. (2009). Decadal prediction: Can it be skillful? *Bulletin America Meteorology Social*, *90*(10), 1467–1485. https://doi.org/10.1175/2009BAMS2778.1
Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., et al. (2021). Initialized Earth System prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment*, *2*(5), 340–357. https://doi.org/10.1038/s43017-021-00155-x
Menary, M. B., Mignot, J., & Robson, J. (2021). Skillful decadal predictions of subpolar North Atlantic SSTs using CMIP model-analogues. *Environmental Research Letters*, *16*(6), 064090. https://doi.org/10.1088/1748-9326/ac06fb
Newman, M. (2013). An empirical benchmark for decadal forecasts of global surface temperature anomalies. *Journal of Climate*, *26*(14), 5260–5269. https://doi.org/10.1175/JCLI-D-12-00590.1
Palmer, M. D., Roberts, C. D., Balmaseda, M., Chang, Y. S., Chepurin, G., Ferry, N., et al. (2017). Ocean heat content variability and change in an ensemble of ocean reanalyses. *Climate Dynamics*, *49*(3), 909–930. https://doi.org/10.1007/s00382-015-2801-0
Penland, C. (1989). Random forcing and forecasting using principal oscillation pattern analysis. *Monthly Weather Review*, *117*(10), 2165–2185. https://doi.org/10.1175/1520-0493(1989)117%3C2165:RFAFUP%3E2.0.CO;2
Penland, C., & Matrosova, L. (1994). A balance condition for stochastic numerical models with application to the El Nino-Southern Oscillation. *Journal of Climate*, *7*(9), 1352–1372. https://doi.org/10.1175/1520-0442(1994)007%3C1352:ABCFSN%3E2.0.CO;2
Penland, C., & Sardeshmukh, P. D. (1995). The optimal growth of tropical sea surface temperature anomalies. *Journal of Climate*, *8*(8), 1999–2024. https://doi.org/10.1175/1520-0442(1995)008%3c1999:TOGOTS%3E;2.0.CO;2
Penny, S. G., Akella, S., Alves, O., Bishop, C., Buehner, M., Chevallier, M., et al. (2017). *Coupled data assimilation for integrated Earth System analysis and prediction: Goals, challenges and recommendations*. World Meteorological Organization. WWRP_2017_3. Retrieved from https://library.wmo.int/doc_num.php?explnum?id=10830
Penny, S. G., Akella, S., Balmaseda, M. A., Browne, P., Carton, J. A., Chevallier, M., et al. (2019). Observational needs for improving Ocean and Coupled Reanalysis, S2S prediction, and decadal prediction. *Frontiers in Marine Science*, *6*(391). https://doi.org/10.3389/fmars.2019.00391
Perkins, W. A. (2019). frodre/pyLIM: JAMES paper LIM code (v0.9.1). [Software]. Zenodo. https://doi.org/10.5281/ZENODO.3243749
Perkins, W. A., & Hakim, G. (2020). Linear inverse modeling for coupled atmosphere–ocean ensemble climate prediction. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001778. https://doi.org/10.1029/2019MS001778
Perkins, W. A., & Hakim, G. (2021a). Coupled atmosphere–ocean reconstruction of the last millennium using online data assimilation. *Paleoceanography and Paleoclimatology*, *36*(5). https://doi.org/10.1029/2020PA003959
Perkins, W. A., & Hakim, G. (2021b). Reconstructions, code, and analysis for "coupled atmosphere-ocean reconstruction of the last millennium using online data assimilation". [Dataset]. Zenodo. https://doi.org/10.5281/ZENODO.4626197
Saha, S., Moorthi, S., Pan, H., Wu, X., Wang, J., Nadiga, S., et al. (2010). The NCEP climate forecast system reanalysis. *Bulletin America Meteorology Social*, *91*(8), 1015–1058. https://doi.org/10.1175/2010BAMS3001.1
Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1
Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., et al. (2006). The NCEP climate forecast system. *Journal of Climate*, *19*(15), 3483–3517. https://doi.org/10.1175/JCLI3812.1
Schmidt, G. A., Kelley, M., Nazarenko, L., Reudy, R., Russell, G. L., Aleinov, I., et al. (2014). Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, *6*(1), 141–184. https://doi.org/10.1002/2013MS000265
Schurer, A. P., Hegerl, G. C., Mann, M. E., Tett, S. F. B., & Phipps, S. J. (2013). Separating forced from chaotic climate variability over the past millennium. *Journal of Climate*, *26*(18), 6954–6973. https://doi.org/10.1175/JCLI-D-12-00826.1
Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., et al. (2019). Towards a more reliable historical reanalysis: Improvements for version 3 of the twentieth century reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908. https://doi.org/10.1002/qj.3598
Smith, D. M., Cusack, S., Colman, A. W., Folland, C. K., Harris, G. R., & Murphy, J. M. (2007). Improved surface temperature prediction for the coming decade from a global climate model. *Science*, *317*(5839), 796–799. https://doi.org/10.1126/science.1139540
Storto, A., Cheng, L., & Yang, C. (2022). Revisiting the 2003–2018 deep ocean warming through multiplatform analysis of the global energy budget. *Journal of Climate*, *35*(14), 4701–4717. https://doi.org/10.1175/JCLI-D-21-0726.1
Storto, A., Masina, S., Balmaseda, M., Guinehut, S., Xue, Y., Szekely, T., et al. (2017). Steric sea level variability (1993–2010) in an ensemble of ocean reanalyses and objective analyses. *Climate Dynamics*, *49*(3), 709–729. https://doi.org/10.1007/s00382-015-2554-9
Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., et al. (2019). Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling. *Climate of the Past*, *15*(4), 1251–1273. https://doi.org/10.5194/cp-15-1251-2019
Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin America Meteorology Social*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1
Trenberth, K. E., Jones, P. D., Ambenje, P., Bojariu, R., Easterling, D., Klein Tank, A., et al. (2007). Observations: Surface and atmospheric climate change. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, et al. (Eds.), *Climate change 2007: The physical*

*science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. Chapter 3.

Xue, Y., Balmaseda, M. A., Boyer, T., Ferry, N., Good, S., Ishikawa, I., et al. (2012). A comparative analysis of upper-ocean heat content variability from an ensemble of operational ocean reanalyses. *Journal of Climate*, *25*(20), 6905–6929. https://doi.org/10.1175/JCLI-D-11-00542.1

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., et al. (2018). Predicting near-term changes in the Earth system: A large ensemble of initialized decadal prediction simulations using the community Earth System Model. *Bulletin America Meteorology Social*, *99*(9), 1867–1886. https://doi.org/10.1175/BAMS-D-17-0098.1