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

Lindsey Michelle Taylor

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2021

Committee:

Gregory Hakim

Cecilia Bitz

Dale Durran

Program Authorized to Offer Degree:

Atmospheric Sciences

University of Washington

Abstract

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

Lindsey Michelle Taylor

Chair of the Supervisory Committee:
Gregory Hakim
Department of Atmospheric Sciences

Improvements to forecasts on interannual to decadal timescales face two major challenges: (1) consistently initializing the coupled system so that variability is not dominated by initial imbalances, and (2) 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. Forecasting atmospheric and oceanic conditions prior to the 1850s isolates internally generated sources of variability by removing the majority of anthropogenic forcing, yet the sparse observational record cannot capture low-frequency variability, further emphasizing the importance of both challenges and paleoclimate proxy data.

This research addresses these two challenges by using a multivariate linear inverse model (LIM) and recent data assimilation (DA) results that extend the observational record with annually-resolved atmospheric and oceanic variables via a low-cost forecast that taps into ocean memory. The reconstructions provide data throughout the last millennium to initialize, validate,

and calibrate the LIM. This work tests the forecast skill of LIMs trained on GCM simulations and on paleo-data assimilated reconstructions. Forecasts are initialized and verified on the reconstructions over 1000-2000 C.E. Both the DA and GCM-analog LIMs are found to have skill on interannual to decadal timescales that surpasses damped persistence for global mean sea surface temperature, as well as widespread significant positive spatial skill for 1-year forecasts of all atmosphere and ocean variables. For cross validation on global mean instrumental data, the LIM trained on paleo-data outperforms a LIM trained on the CCSM4 last millennium simulation beyond 4-year lead forecasts, with the CCSM4-LIM reaching climatological variance before the paleo-informed LIM. The paleo-data LIM requires consistent OHC states that, when provided, increase forecast skill outperformance over the GCM-informed LIMs.

TABLE OF CONTENTS

List of Figures iii

List of Tables vi

Chapter 1. Introduction 1

1.1 Decadal Prediction 1

1.1.1 Limitations 3

1.2 Modeling Approaches 3

1.3 Proxies and Data Assimilation 4

1.3 A Proposed Solution 5

Chapter 2. Methods 6

2.1 Linear Inverse Modeling 6

2.1.1 Theory 6

2.1.2 Calibration 7

2.1.3 Climate Emulation and Model Design 8

2.2 Forecast Experiment Setup 11

2.2.1 Verification Data 12

2.2.2 Dynamical Features in a LIM 13

2.2.3 Skill Metrics 14

Chapter 3. Forecast Results 16

3.1 In-sample Coupled Experiments 16

3.1.1 Ocean Field Contribution to Forecast Skill 19

3.2 Out-of-sample Coupled Experiments 21

3.3 Out-of-sample Experiments on Atmosphere-only Data 23

3.3.1 Adding an Inconsistent Ocean 26

3.4 Out-of-sample Experiments on Ocean-only Data 28

3.5 Historical Simulation Experiments 30

Chapter 4. Sources of Skill 33

4.1 Empirical Normal Modes 33

4.2 Single ENM Experiments 34

4.3 Physical Mechanisms 36

Chapter 5. Conclusions 40

References 42

LIST OF FIGURES

Figure 1. Scalar index correlation coefficient for the LMR-LIM (red) and CCSM4-LIM (black). Shown are in-sample forecasts using both models when forecasting global mean sea surface temperature (upper left), Pacific Decadal Oscillation (upper right), Nino 3.4 index (lower left), and North Pacific Index (lower right). The dashed-dot lines are AR1 forecasts trained on the reference data for either model. Horizontal dashed lines represent the 95% confidence threshold 17

Figure 2. 5-year lead air temperature correlation coefficient for in-sample forecasts from the CCSM4-LIM (left; 855-1850 C.E.) and the LMR-LIM (right; 1005-1850 C.E.)..... 18

Figure 3. Total error variance growth by variable for the LMR-LIM (left) and CCSM4-LIM (right) in-sample forecasts. Bottom row is the same as the top row but without OHC in the LIM's calibration. Variables included are ocean heat content (ohc), precipitation (pr), surface level pressure (psl), surface air temperature (tas), ocean surface temperature (tos), 500hPa heights (zg), and dynamic ocean surface heights (zos). Climatological variance is represented by the horizontal dashed line at y=1..... 20

Figure 4. Standardized error variance growth of global mean air temperature for coupled forecasts on LMR-data from 1850-2000 C.E. CCSM4-LIM forecasts are in black and LMR-LIM forecasts are in red. Solid lines represent the mean LIM forecasts with clouds representing the 95% confidence threshold (with a lower bound of 2.5% and upper bound of 97.5%). The horizontal dashed black line is the climatological variance..... 22

Figure 5. Same as in Figure 2 but for out-of-sample forecasts on the withheld LMR data from 1850-2000 C.E. 23

Figure 6. Same as in Figure 3 but for forecasts on GISTEMP (1880-2019 C.E) and 20CR data (1836-2015 C.E.). The forecasts on GISTEMP use a LIM calibrated on air temperature only (left) and the forecasts on 20CR use a LIM calibrated on four atmospheric variables (right).. 24

Figure 7. 5-year lead surface air temperature correlation coefficient using an LMR-LIM (left) and a CCSM4-LIM (right). The top row are the temperature-only LIMs forecasting on GISTEMP and the bottom row are the atmosphere-only LIMs forecasting on 20CR...25

Figure 8. Correlation coefficient of global mean sea surface temperature forecasts on ocean-only instrumental data. Solid lines are mean LIM forecasts for the LMR-LIM (red) and CCSM4-LIM (black) trained on OHC and SST. Both LIMs forecast on the SODA (1871-2008 C.E.) and the HADLEY EN4 (1900-2010 C.E) data sets. Clouds represent the 95% confidence bounds and the dashed red line is the 95% significance threshold for both frameworks 28

Figure 9. 1-year spatial correlation coefficient of sea surface temperatures using a CCSM4-LIM (left) and the LMR-LIM (right). Results are shown for forecasts on the HADLEY EN4 (top row) and SODA (bottom row) data sets..... 29

Figure 10. Same as in Figure 3 but for forecasts on GISS-2E-R 1951-2005 C.E model simultation data. These forecasts use LIMs that are trained on fully coupled data as for the in-sample experiments 31

Figure 11. Gridded correlations for 1-year forecasts (top row) and 4-year forecasts (bottom row) provided by the CCSM4-LIM (left) and the LMR-LIM (right). Forecasts are verified against the GISS-2E-R 1851-2005 C.E. model simulation data 32

Figure 12. Global mean temperature error variance growth for ENM experiments using the LMR-LIM (left column) and CCSM4-LIM (right column) to forecast on GISTEMP (top row) and 20CR (bottom row). The LIM is run with single ENMs and the top two skillful ENM forecasts are shown in dot dashed lines. The top two skillful ENMs are run together (opaque blue line) and compared to the full ENM forecast (opaque black line). The dashed black line represents climatological variance (i.e., when the forecast is indistinguishable from climatology) 35

Figure 13. ENM properties of the LMR-LIM (red) and CCSM4-LIM (black) that are trained on a) surface air temperature and b) surface air temperature, sea level pressure, precipitation, and 500hPa heights. The most skillful ENMs found from single ENM experiments in section 4.2 are marked by stars 37

Figure 14. Most skillful ENM patterns represented in the air temperature field. Results are shown for the temperature-only calibration (top row) and the atmosphere-only calibration (bottom row) for the CCSM4-LIM (left column) and the LMR-LIM (right column). Each panel is a single phase of the stationary modes (4 modes in total). All data are standardized by the largest amplitude to have a maximum value of +/-1..... 38

LIST OF TABLES

Table 1. 10-year GMT forecast correlation coefficient for 10-year forecasts spanning 1880-2019 C.E on GISTEMP and 1846-2015 C.E. on 20CR. Skill is provided for three experiments: 1) a full LIM that includes all available variables in the verification data set (i.e., No OHC added), 2) a full LIM that is initialized with zero anomaly OHC conditions (i.e., + 0 OHC), and 3) a full LIM that is initialized with OHC ICs from random states in the LMR data set (i.e., + LMR OHC)..... 27

ACKNOWLEDGMENTS

I would first like to thank my advisor, Greg Hakim, for his patience, guidance, and constant support throughout this process. I would also like to thank my committee members, Dale Durran and Cecilia Bitz, for their time and feedback on this work.

This research uses code designed by Andre Perkins, who has provided me with much needed assistance when I was starting out. I would also like to thank my officemate, Katie Brennan, for her extremely thoughtful discussions and for being a constant source of positivity. Furthermore, I would like to thank all members of the Hakim research group, Jessica Badgeley, Anna Black, Vince Cooper, Chang Liu, Gemma O'Connor, Luke Parsons, Sara Sanchez, Robert Tardif, Kinya Toride, and Molly Wieringa, for carefully listening to my research and offering their advice. In addition, I thank my entire cohort for making the first year of graduate school memorable and for their unending support.

I would not be here today if it were not for my undergraduate mentors, Yutian Wu and Ernest Agee, who introduced me into the amazing world of research and continue to reach out to me with interest in my work. Furthermore, I would like to express my deepest gratitude to my mom and dad, who have provided me with so much love throughout this entire process.

Chapter 1. INTRODUCTION

1.1 DECADAL PREDICTION

Climate mitigation and adaptation strategies require accurate short to long-term climate change predictions, as changing temperatures, air quality, and regional precipitation directly impact agriculture, water security, and human health. However, unprecedented climate conditions and uncertain, dynamically evolving climate change hinder policy making (e.g., Fussel, 2007). Furthermore, earth-system predictions have time-dependent uncertainties. For example, weather prediction is skillful up to about two weeks due to imperfect initial conditions that corrupt the forecast, otherwise known as predictability of the first kind (Lorenz, 1963). On the other hand, predictability of the second kind involves model sensitivities to perturbations in the boundary values and can be represented as a forcing term (e.g., Chu, 1999). Long-term climate prediction on centennial timescales is mostly a boundary condition problem and tends to only consider external forcing (e.g., climate projections). Decadal prediction lies in-between short-term weather predictions and long-term climate change projections, and thus considers a blending of uncertainties tied to both initialization and boundary conditions (e.g., Meehl et al., 2009).

The ocean is an important source of predictability on seasonal to decadal timescales due to the high thermal inertia of the upper-ocean layer that holds memory of the system (e.g., Smith et al., 2019). The high heat capacity of the upper ocean serves to integrate over the noisier atmosphere and regulate climate variability (e.g., Hasselman, 1976; Deser et al., 2010). Previous studies have found the North Atlantic and North Pacific to exhibit strong decadal variability via internally generated modes such as the Pacific decadal oscillation (PDO; e.g., Mantua et al., 1997) and Atlantic meridional overturning circulation (AMOC; e.g., Delworth et al., 1993). Accurately capturing internal variability would improve decadal forecasts, yet representations of these internal modes in the models are impacted by initial value and boundary condition uncertainty. The pioneering work of Branstator and Teng (2010, 2012) was the first to quantify predictability limits in the Pacific and Atlantic Ocean basins, finding that initialization is an important source of skill up to 6- and 8-year leads, respectively. Information from the initial values then becomes secondary to information provided by the forced response. Thus, skillful interannual to decadal

prediction (2 to 20 years) relies on the proper initialization of ocean states and capturing the correct phase of internal variability.

Extracting the forced signal from background internal variability is complicated by the short observational record, and different methods of separating the two tend to introduce biases (e.g., Schurer et al., 2013; Frankcombe et al., 2015). The sparse observational record ties to an inability to provide large samples of consistent atmospheric and oceanic conditions to initialize and verify climate models. Furthermore, the short record fails to sample low-frequency variability on multidecadal timescales which has overall contributed to the lack of general knowledge on internal variability (e.g., Trenberth et al., 2007). The ocean serves as a large source of decadal variability and having large observational samples, particularly of ocean heat content (OHC), would serve to benefit long-term forecasting. However, the ocean has been relatively poorly observed prior to the Argo float implementation in the early 2000s. The inhomogeneous record of upper-ocean observational data as well as differences in model configurations introduce large uncertainties into historical estimates of OHC with significant differences between ocean reanalysis products (e.g., Palmer et al., 2017). In addition, many reanalyses are created using data assimilation methods that reconstruct the atmospheric and oceanic components of the system independently. However, predictability on decadal timescales rely on properly initializing coupled modes of the climate system such that the ocean may inform the atmosphere and vice versa (e.g., Penny et al., 2019).

Even with an extended observational record of properly coupled data, running fully coupled global climate models (GCMs) across multiple decades is computationally expensive. Furthermore, models have initialization issues and inherent biases that corrupt the forecast of internal modes. For example, GCMs have little agreement on the spatial variance of internal variability such as PDO, as well as varying power spectrum and persistence that yield different predictive ability across models (e.g., Farnetti, 2017). These models tend to drift towards their imperfect climatology and require bias corrections to limit the drift through either full-field or anomaly initialization (Hazeleger et al., 2013). In addition, model prediction is sensitive to small perturbations in the initial conditions, which necessitates the application of large ensembles of data to sample initial state uncertainties (Meehl et al., 2014). Skillful decadal prediction thus requires large samples of coupled atmosphere-ocean initial conditions (ICs) to properly estimate

the climate system and sample different initial states of internal variability. Initial imbalances may then be lessened by consistently initializing the coupled system.

1.1.1 *Limitations*

The previous section provides several limitations of predictability on interannual to decadal timescales (2 to 20-year leads) that may be summarized by four points:

1. Computational cost of running fully-coupled GCMs
2. Limited data prior to the instrumental era
3. Inaccurate model representation of internal variability
4. Inability to initialize models with large samples of data

The following section presents solutions to each of the above limitations. However, we note there are additional limitations on decadal prediction not being addressed here, such as teasing apart external forcing from internal variability. The focus of this work lies more on capturing internal variability by training on data prior to the largely anthropogenically-influenced time beginning in the 1850s. The instrumental data sets used in verification will, undoubtedly, contain some influence from external forcing not considered within the proposed model, yet we limit the influence of anthropogenic external forcing in these data prior to running forecast experiments.

1.2 MODELING APPROACHES

Decadal predictions are limited by the computational cost of running a model representing the coupled interactions within the system over multiple years. For this reason, many GCMs, including those within phase 5 of the Coupled Model Intercomparison Project (CMIP5), run decadal predictions at coarse resolutions. However, a high-resolution is required to resolve regional-scale processes such as precipitation events (e.g., Salvi et al., 2017). As opposed to running the flag-ship GCMs, a more computationally efficient modeling approach is to emulate the climate system through statistical methods. A widely-used empirically-based climate emulator is a linear inverse model (LIM; Penland and Sardeshmukh, 1995). The LIM is an attractive model due to its computational efficiency, distinct timescale separation, and flexibility of calibration. Multiple studies have found skillful decadal forecasts of regional sea surface temperatures using LIMs (e.g., Hawkins and Sutton, 2009; Foster et al., 2020). The LIM has also proven to be a suitable benchmark that exceeds persistence and has comparable skill to CMIP5

models for global surface temperature forecasts (Newman, 2013). These studies establish the LIM as a reliable forecasting model and illustrate the model's ability to diagnose forecast skill. While computationally efficient, the LIM is able to capture the power spectra of internally-generated modes in the reference data (Perkins and Hakim, 2020; PH20). Properly representing internal variability is a necessity for skillful predictions beyond seasonal timescales.

1.3 Proxies and Data Assimilation

The observational record provides an incomplete sample of the slow-varying components of the climate system and leads to large uncertainties in model simulations of internal variability. However, the Earth contains a record of past climate captured by ice cores, tree rings, and more, otherwise referred to as climate proxies. Proxies are snapshots of past climate in time and space that allow us to see how atmospheric and oceanic variables changed through time. While proxies represent past observations, they do have several limitations. For one, proxies are time-integrated states with varying temporal resolution. Secondly, proxies have greatly varying spatial resolution, and proxy-availability is highly time-dependent (see the Pages2k Consortium). For example, a majority of proxies are located in the Northern Hemisphere, specifically over North America, yet other regions, such as the Southern Ocean, have limited spatial coverage by proxies. Thus, in order to retrieve a dynamically consistent view of past climate through time and space, proxies are combined with climate models through data assimilation (DA).

The Last Millennium Reanalysis (LMR) project applied data assimilation to paleoclimate reconstructions using an ensemble Kalman filter (Hakim et al., 2016; Tardif et al., 2019). These reconstructions weight the proxy data against the prior model to yield a posterior analysis probability density function with less variance than either the prior or proxies. However, data assimilation methods are restricted by the computational cost of running coupled GCM forecasts between reconstruction times and overall low forecast skill at proxy-time resolution. This is circumvented by using offline assimilation, where all temporal information comes from the proxies and reconstructed states are independent of one another. Offline reconstructions save the expense or running an ensemble forecast at the proxy resolution to get the next prior state, yet do not provide temporal constraints that including a forecast step otherwise would. These no-forecast reconstructions are unable to tap into ocean memory and miss out on this key source of decadal predictability.

Perkins and Hakim 2017 first looked at performing online DA using a LIM to emulate the dynamics of a GCM. This allowed for a computationally efficient forecasting step to connect the posterior analysis state to the next prior state estimate, and showed overall more agreement with observations than the offline reconstructions. Perkins and Hakim 2021 (PH21) performed online reconstructions using the LIM to reconstruct coupled atmosphere-ocean fields over the last millennium (1000-2000 C.E.). These reconstructions have better dynamics than the offline reconstructions (e.g., atmosphere-ocean coupling) as well as enhanced decadal to centennial variability and memory of ocean heat content. This work has opened the way for studying decadal predictability by providing large samples of coupled initial conditions to verify and initialize models.

1.4 A PROPOSED SOLUTION

We address the limitations of forecasting on interannual to decadal timescales with a short observational record by using a multivariate linear inverse model and online reconstructions from PH21. These methods tap into ocean memory via coupled atmosphere-ocean field reconstructions and allow one to understand the sources of forecast skill based on the linear modes of the system. The PH21 data (henceforth referred to as LMR data) provides 1,000 years of coupled atmosphere-ocean data to consistently initialize and verify decadal forecasts. We investigate whether a LIM's skill can be improved by calibrating on paleo-data as compared to GCM-analog LIMs. Both LIMs are expected to capture large-scale coupled dynamics of the climate system, yet the paleo-LIMs are trained on proxy-weighted reconstructions that are anticipated to partially correct for model misrepresentation of the spatial variance in internally generated modes. We anticipate to see a large portion of the paleo-LIM skill attributed to the ocean fields provided by the LMR data due to the large decadal variability and high persistence of the reconstructed ocean heat content.

Chapter 2. METHODS

2.1 LINEAR INVERSE MODELING

2.1.1 *Theory*

A linear inverse model (LIM) approximates a nonlinear dynamical system as linear processes plus stochastic white noise forcing.

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

Here, the time tendency of the anomaly state vector, **x**, is represented by slow-varying climate, **Lx**, plus fast timescale weather processes that are defined as stochastic forcing, ξ. This distinct timescale separation allows for application of the central limit theorem and, thus, statistical closure (Hasselman, 1976).

Equation 1 is integrated to get a forecast equation at a time t+τ in the future.

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

The propagation matrix, $\mathbf{G}_\tau$, is a Green's function related to the linear operator by $\mathbf{G}_\tau = \exp(\mathbf{L}\tau)$ and is calculated based on the lag-covariance statistics of the system that minimizes the error variance, **ϵ**, in equation 2 (Penland, 1989):

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

The angle brackets denote an expectation, taken as a sample average in time and **C** is an autocovariance matrix. $\mathbf{G}_\tau$ is independent of the current time t, so we may calculate a $\mathbf{G}_\tau$ for every lag, τ. This allows for faster processing when running forecasts and a determination of the dynamical behavior at all times given a single time, τ.

The LIM assumes that the system has stable dynamics and stationary statistics, meaning that the statistics must not change in time. This requires the system conserves energy via an energy

balance equation described by the Fluctuation-Dissipation Relationship (FDR; Penland and 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 a stochastic forcing:

$$\frac{\mathrm{d}\mathbf{C}_{\mathrm{o}}}{\mathrm{dt}} = \mathbf{L}\mathbf{C}_{\mathrm{o}} + \mathbf{C}_{\mathrm{o}}\mathbf{L}^{\mathrm{T}} + \mathbf{Q} = 0 \tag{4}$$

In equation 4, $\mathbf{Q} = \langle \xi\xi^{\mathrm{T}} \rangle$ is the stochastic noise covariance matrix.

2.1.2 *Calibration*

The computational cost of calibrating a model and forecasting with large samples of full-field data requires working in a reduced space. The LIM calibration employs a two-step EOF reduction that greatly reduces the data's dimensionality (developed by Perkins and Hakim, 2020). The first step is intuitive in that all variable fields are converted to a reduced space for faster processing. The second step of the reduction not only further compresses the data for additional computational ease, but also prevents redundant information within the LIM calibration.

The first step of the EOF reduction compresses individual variables by performing singular value decomposition (SVD) of the latitude-weighted fields, $\mathbf{X}^{\mathrm{W}}$.

$$\mathbf{X}^{\mathrm{W}}_{\mathrm{V},1} = \mathbf{U}_{\mathrm{V}}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$$

This first reduction truncates to the leading K modes, retaining above 90% variance of the original fields. The variables are then projected onto their respective EOFs, $\mathbf{U}_{\mathrm{v}}$, and standardized by the total component standard deviation across K modes.

$$\widehat{\mathbf{X}}_{\mathrm{V1}} = \mathbf{U}^{\mathrm{T}}\mathbf{X}_{\mathrm{V1}}$$

$$\widehat{\mathbf{X}}_{(\mathrm{V1})\sigma} = \frac{\widehat{\mathbf{X}}_{\mathrm{V1}}}{\sqrt{\sum_{j=1}^{K} \sigma_j^2}}$$

The standardization allows for equal consideration of all variables within the concatenated state prior to the second reduction.

Each of the n standardized variables are concatenated into a single state vector, $\mathbf{Y}$.

$$\mathbf{Y} = \begin{bmatrix} \widehat{\mathbf{X}}_{(\mathrm{V1})\sigma} \\ \vdots \\ \widehat{\mathbf{X}}_{(\mathrm{Vn})\sigma} \end{bmatrix}$$

We perform SVD again on the combined state vector to further reduce the fields. The multivariate EOFs (MVAR EOFs; $\mathbf{U}_\mathrm{M}$) eliminates the collinearity of variable fields and captures the coupled variability. Finally, the concatenated data, $\mathbf{Y}$, is projected onto the multivariate EOFs to retrieve the fully compressed state, $\widehat{\mathbf{Y}}$, also called the LIM space.

$$\widehat{\mathbf{Y}} = \mathbf{U}_\mathrm{M}^\mathrm{T}\mathbf{Y}$$

The data is mapped into LIM space via this two-step EOF reduction and, vice versa, returned to a reduced gridded space using the variable EOFs and multivariate EOFs obtained during calibration. While all variables follow this two-step reduction, the standardized OHC data are concatenated onto the end of the compressed multivariate state vector.

$$\widehat{\mathbf{Z}} = \begin{bmatrix} \widehat{\mathbf{Y}} \\ \widehat{\mathbf{X}}_{\mathrm{OHC}\sigma} \end{bmatrix}$$

PH20 found that separating the OHC EOFs as opposed to combining within the state vector prior to the multivariate reduction allows for larger e-folding times (EFT) and more system memory within the LIM calibration.

2.1.3 *Climate Emulation and Model Design*

The LIM is calibrated on time series of coupled atmosphere-ocean data. These calibration data sets provide full-field dynamics for the LIM to emulate. There are two main branches of LIMs

that are compared in this study: LIMs trained on global climate models (GCM-LIMs) and LIMs trained on paleo-informed data (LMR-LIMs). While both LIMs contain information about coupled dynamics, the latter is informed by actual climate variability before the instrumental era.

GCM-LIMs

A LIM that is calibrated on data from a GCM simulation acts as an analog to that GCM. This study uses GCM-LIMs that are trained on data from last millennium simulations (850-1850 C.E.). The first GCM-LIM is calibrated on Community Climate System Model version 4 (CCSM4) last millennium simulation, and the second GCM-LIM is calibrated on the MPI last millennium simulation. Both these data sources include forcing from land use change, volcanic eruptions, and greenhouse gasses. Comparing the performance between two GCM-LIMs tests the LIM's sensitivity to different model representations of variability within the calibration data sets.

The GCM simulations provide full-field coupled atmosphere-ocean variables at monthly resolution spanning the last millennium. The variables considered in forecast experiments are the surface air temperature, sea level pressure, 500hPa heights, precipitation, ocean surface height, ocean surface temperature, and upper 700m ocean heat content (OHC).

GCM-LIM calibration parameters

Prior to calibration, the GCM data is converted to a compressed state by applying the two-step EOF reduction from section 2.1.2. The LIM's calibration parameters vary based upon the degrees of freedom of 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, which retains around 75% of the coupled variance (i.e., the shared variance between all variables). Beyond 25 MVAR EOFs, the rate of variance increase is very slow such that any additional modes retained would only add a small percentage of the coupled variance (PH20). When OHC is included in the calibration, we retain 20 OHC EOFs as a separated field. The calibration parameters are consistent at 400 variable EOFs, 25 multivariate EOFs, and 20 OHC EOFs for the CCSM4-LIM and MPI-LIM. However, while the variables included within the calibration are from the same data source, the calibration variables must also be available within the verification data sets. Thus, the LIM calibration varies between forecast experiments, resulting in a different LIM in every experiment (see section 2.2.1 for further detail).

LMR-LIMs

The data from PH21 (the LMR data) provides a large sample of consistent, full-field variables to initialize and calibrate a LIM. While most DA uses an offline approach that reconstructs different fields separately, the LMR data are online coupled atmosphere-ocean reconstructions. PH21 retains memory between reconstruction times by applying a computationally efficient forecast-step using a CCSM4-LIM. The online reconstructions capture more ocean variability on decadal to centennial timescales than the offline reconstructions. In addition, the online reconstructions reproduce dynamical features that closely resemble the reference model, such as lead-lag relationships between the atmosphere and ocean (PH20).

The reconstructions use a 100-member ensemble with 30 Monte Carlo (MC) iterations, where each MC iteration samples 75% of available proxies to test proxy-sensitivity. Data has been reconstructed from 1000-2000 C.E., but the LMR-LIM is calibrated on the truncated time from 1000-1850 C.E. to exclude large trends tied to anthropogenic forcing. Note that the LMR-LIM is distinct from the LIM used in the construction of the PH21 data due to the weight that proxies hold within the LMR-LIM's calibration. The LIM used to make the PH21 data had only information from an imperfect model's dynamics, while the LMR-LIM has a better grasp on reality attributed to proxies that explain the system's actual variability.

LMR-LIM Calibration Parameters

The LMR-LIM calibration follows the same two-step EOF reduction as the GCM-analog LIMs. However, the LMR-LIM has fewer degrees of freedom than the GCM-LIMs, and, therefore, captures more variance in fewer modes. The availability of proxies most likely sets the degrees of freedom at a lower value where only a small number of modes matter. 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. The number of separated OHC EOFs is set at 20 as for the GCM-LIMs. However, retaining 20 OHC EOFs in the LMR framework results in a ratio of multivariate EOFs to OHC EOFs of 10/20 while the GCM-LIM ratio is 25/20. Thus, OHC is given more weight than the other variables within the LMR-LIM calibration. For this reason, we test the LIM

sensitivity to fewer OHC EOFs retained for a more equal consideration across all variables within the calibration.

Additional LMR-LIMs

The LMR-LIM trained on 1000-1850 C.E. reconstructions is the main focus of this study, but we consider the LIM's sensitivity to different variants of the LMR-data. One concern for the LMR-LIM was the reconstructions on proxy-sparse data within the earlier times of the calibration data set (~1000-1200 C.E.), which may corrupt variable EOFs of a LIM trained on a highly proxy-variable time period. A solution is to substitute the variable EOFs of the LMR-LIM with those provided from a GCM-LIM and project the LMR data onto these EOFs during the first step of the EOF reduction. This allows the LMR-LIM to be trained on patterns of variability from the GCM yet retain information on the coupled dynamics from the LMR data. The novelty of this methodology lies in the projection of the LMR data onto the GCM variable EOFs, so the LMR-LIM calibration following these methods is referred to as the separate projection LMR-LIM, or SPLMR-LIM. We also test the LIM's sensitivity to time truncation of the LMR data by considering LMR-LIMs trained on the later time periods of 1450-1850 C.E and 1650-1850 C.E.

We consider different variants of the LMR-LIM by calibrating on subsets of variables from the LMR-data. For example, the atmospheric component of the system may be emulated by only calibrating the LIM on the atmospheric variables of the LMR-data and, vice versa, for the ocean. The LIM will thus be emulating the uncoupled system, though the LMR-data does have an imprint of the coupled system through its construction via coupled data assimilation (e.g., the air temperature will have embedded persistence from the ocean). Separating the atmospheric and oceanic components of the system allow for the evaluation of the relative importance of each component to the overall skill of the LMR-LIM as well as a comparison between the component contribution between the LMR-LIM and GCM-LIM.

2.2 Forecast Experiment Setup

Section 2.2.1 provides a discussion on the data products used as verification in forecasting experiments, section 2.2.2 explains dynamical features we analyze in the LIM, and section 2.2.3 provides an explanation of skill metrics.

2.2.1 *Verification Data*

The LIM calibration is limited by the availability of the verification data. In other words, while the data sources that the LIM is calibrated on are the same between experiments, the variables included in the LIM calibration vary between forecasting experiments. For example, many instrumental data sets only have either an atmosphere or an ocean. If there are only atmospheric variables available to be verified on, the LIM can only be calibrated on atmospheric variables, and vice versa. Therefore, it is important to consider data availability in the verification data sets as the LIM's calibration (and, thus, the dynamics) will change.

There are a total of five forecast experiments within this study: 1) in-sample coupled, 2) out-of-sample coupled, 3) out-of-sample on atmosphere-only instrumental data, 4) out-of-sample on ocean-only instrumental data, and 5) historical experiments. In-sample forecasts refer to forecasting on the same data and time that the LIM is calibrated on while out-of-sample refers to forecasting on data not included in the LIM calibration. Out-of-sample experiments may refer to data from a completely different data source and/or data from a separate time that the LIM is calibrated on.

Experiment 1 is verified on in-sample data that are the same as described in the LIM calibration sections. Experiment 2 is on the LMR data not included within the LMR-LIM calibration, that is, 1850-2000 C.E. It may be argued that this data is not truly out-of-sample, since the CCSM4-LIM was used in the data's construction and the LMR-data shares large-scale dynamics regardless of time.

Experiment 3 is a truly out-of-sample forecast on instrumental data that contains atmospheric variables but no ocean data. The atmosphere-only forecasts are validated on data provided from the GISS Surface Temperature Analysis (GISTEMP) and the 20th Century Reanalysis (20CR). The GISTEMP data is referred to as TEMP-only because only surface air temperature is available. On the other hand, the 20CR data is referred to as ATMOS-only because there are four atmospheric variables available: surface air temperature, sea level pressure, precipitation, and 500hPa heights. GISTEMP data is available from 1880-2019 C.E. and 20CR data is available from 1836-2015 C.E. Missing data are replaced by nans and are thus ignored during LIM calibration.

Experiment 4 is an out-of-sample forecast on instrumental data that have only ocean variables. The ocean-only experiments use data provided from the Simple Ocean Data Analysis (SODA)

and HadleyEN4 data sets. The only variables available in these products that overlap with the calibration variables are sea surface temperature (SST) and ocean heat content (OHC). The SODA and HadleyEN4 data have monthly resolution and span from 1871-2008 C.E. and 1900-2010 C.E., respectively.

Finally, experiment 5 is verified on coupled historical simulation data sets. These historical data are provided by the GISS-2E-R from 1851-2005 C.E. All variables within the historical data set are the same as those from the in-sample experiment, and the historical data set thus contains all the available variables that both LIMs may be calibrated on. Experiment 5 is a truly out-of-sample experiment on fully coupled data that allows for proper comparison between the skill of different LIM frameworks.

Data Processing

All calibration and verification data sets go through the same processing steps prior to calibration and forecasting. First, we use bilinear interpolation to convert data to a common 2º by 2º grid. The data set is truncated if needed, annually averaged, and converted to gridded anomalies. Finally, data are detrended to remove any long-term trends and model drift. The detrending better illustrates the LIM's ability to capture internal variability, with a large portion of the trend attributed to anthropogenic external forcing for data post-1850s. Therefore, anthropogenic trends are outside the scope of the LIMs used in this study, though there is potential to add a forcing term to the LIM in future work.

2.2.2 *Dynamical Features in a LIM*

Previous studies have shown that LIM's are able to capture the overall power spectra of important dynamical features (e.g., PH20). Therefore, we later verify the LMR-LIM's ability to reproduce dynamical features in the fully coupled experiments. Three indices are calculated from full-field data, including the Nino 3.4 index, Pacific Decadal Oscillation (PDO), and the North Pacific Index (NPI). There is a lack of reliable full-field ocean observations, so these indices are limited to in-sample forecasting experiments. A common metric assessed in this study is the global mean air temperature (GMT), calculated from full-field forecasts. When there is no air temperature available, we consider the global mean sea surface temperature (SST).

2.2.3 *Skill Metrics*

Forecast quality is evaluated based on the anomaly correlation coefficient and the standardized error variance. These metrics are used to assess the LIM's ability to reproduce the verification data through time (i.e., across 1–20-year leads). Framework comparison also considers the statistical significance of skill metric calculations as well as model persistence.

Anomaly Correlation Coefficient

The anomaly correlation coefficient (ACC) is a comparison between the forecast and observational anomalies relative to climatology and measures the signal timing between two time series (or grid points). The correlation between the forecast anomalies, x, and observed anomalies, y, are calculated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

Where summations occur over select lead forecasts (e.g., for 1000 years of data to initialize with, n=999 for 1-year leads forecasts, n=998 for 2-year leads, and so on). The correlation coefficient is important in diagnosing the phasing between the observations and forecast, yet fails to provide information about the amplitude and biases. This metric may be compared to a 95% significance threshold that is calculated from the observations. When the forecast skill drops below this threshold, it cannot be distinguished from red noise.

Standardized Error Variance

The standardized error variance (SEV) takes into account amplitude information and biases between the forecast and observations. This metric is calculated from the forecast error and standardized by the observational variance following the equation below.

$$\text{std err var} = \frac{\sum_{i=1}^{n}(err_i - \overline{err})^2}{n-1} * \left(\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}\right)^{-1} \tag{6}$$

Where, once again, n represents a summation over all forecasts for the selected lead and err is the difference between the forecast and reference data. The bar over the err term represents the average forecast error taken through time (i.e., varying ICs) for the selected lead. The standardization sets the climatological error variance equal to 1, and the standardized error variance is bounded between 0 and 2, where a value of 0 indicates no error in the forecast. A value greater than 1 implies that the variance of the forecast error exceeds the observational variance, which occurs for noisier modes of the system. Furthermore, values that exceed 1 are unique to ensemble forecasting, whereas a single deterministic model has an upper bound of 1. However, the forecast error variance always converges to climatology as the forecast time goes to infinity. The forecast no longer has skill once the climatological variance has been reached.

Bootstrap Resampling

Comparing the forecast skill between multiple calibration frameworks requires confidence intervals to diagnose the statistical significance. In the following discussion, framework comparison using the ACC or SEV have 95% confidence bounds added via bootstrap resampling. This method uses 10,000 iterations that randomly samples 75% of the available data to estimate population statistics.

Autoregressive Forecasts

Forecast skill is compared to an autoregressive model that has dependence on the previous time step (AR1) and thus acts as damped persistence. Therefore, a forecast with more skill than the AR1 forecast may be capturing additional dynamical information above persistence. The deterministic AR1 forecast equation is:

$$\boldsymbol{X}_t = \varphi \boldsymbol{X}_{t-1} \tag{7}$$

The coefficients for the AR1 process, φ, are calculated from the lagged autocorrelation of the calibration data to provide a direct comparison to the respective LIM framework.

Chapter 3. FORECAST RESULTS

We hypothesize that the LMR-LIM will produce more skillful forecasts than the CCSM4-LIM due to information derived from actual climate variability contained in the paleoclimate proxy records. Furthermore, we expect that most of the skill from LMR-LIM forecasts will come from ocean fields. This chapter is focused on addressing these expectations by providing an in-depth framework comparison between the two types of LIMs and discussion on the ocean's contribution to overall forecast skill.

This chapter is split into 5 forecast experiments: in-sample coupled experiments in section 3.1, out-of-sample coupled experiments in section 3.2, instrumental forecast experiments in sections 3.3 and 3.4, and historical forecast experiments in section 3.5.

3.1 IN-SAMPLE COUPLED EXPERIMENTS

While GCM-LIMs have been thoroughly studied in past research (e.g., see PH20 on CCSM4- and MPI-LIMs), the LMR-LIM is a completely new model that has yet to be applied in any prediction studies. A way to test the null hypothesis that this model has skill comes from in-sample forecasting experiments and provides a base comparison between in-sample forecasting skill between the LMR- and GCM-LIMs. In other words, if the LMR-LIM is unable to predict data that the model has already seen, then the model will be most likely result in insignificant forecasts on out-of-sample data.

We first run the LMR-LIM to forecast the same data it is calibrated on, i.e., 1000-1850 C.E. of the LMR data, and compare the forecast performance to the CCSM4-LIM in-sample forecast on 850-1850 C.E. of the CCSM4 last millennium data. These in-sample forecasts are not limited in terms of calibration variables and thus include all the available variables described in section 2.1.3, making them fully coupled forecasts. For this experiment, the LIMs are initialized with all available times and run up to 20-year leads.

Forecast Skill

Figure 1 shows the average scalar correlation coefficient performance for four separate metrics summarized across all leads. An AR1 forecast is plotted in a dash-dot line as a measure of

damped persistence, and the 95% significance threshold is represented by the horizontal dashed red and black lines.

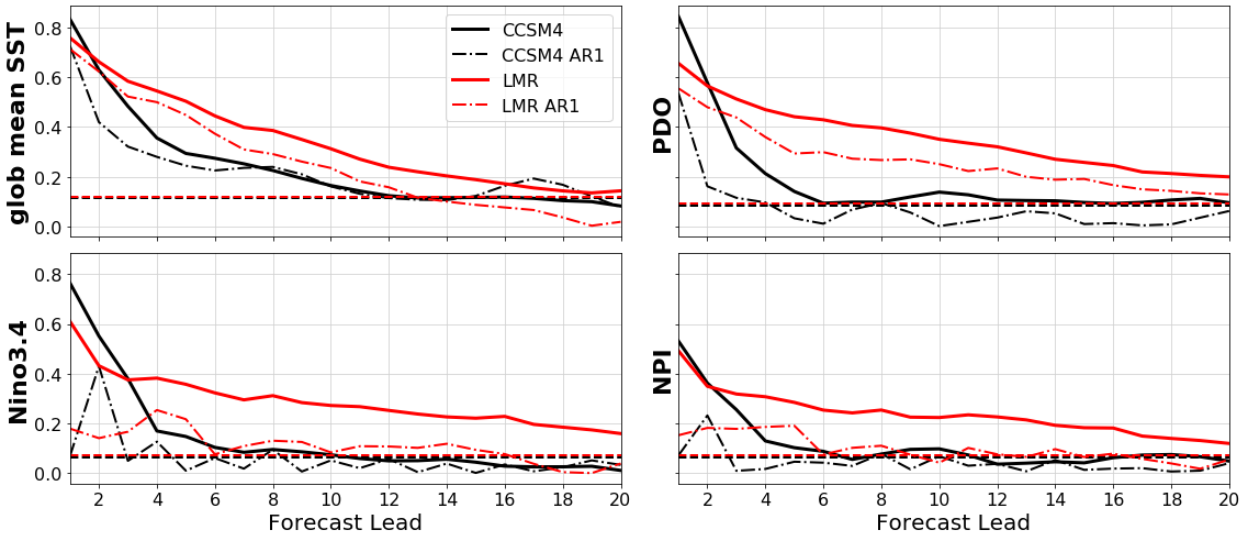

Figure 1: Scalar index correlation coefficient for the LMR-LIM (red) and CCSM4-LIM (black). Shown are in-sample forecasts using both models when forecasting global mean sea surface temperature (upper left), Pacific Decadal Oscillation (upper right), Nino 3.4 index (lower left), and North Pacific Index (lower right). The dashed-dot lines are AR1 forecasts trained on the reference data for either model. Horizontal dashed lines represent the 95% confidence threshold.

For one-year forecasts, the CCSM4-LIM has better forecast skill than the LMR-LIM for all four indices. However, the LMR-LIM begins to outperform the CCSM4-LIM at 2-year leads when forecasting global mean SST. The LMR-LIM outperforms the CCSM4-LIM beginning at 3-year leads for the PDO and NPI, and beginning at 4-year leads for the Nino3.4. The LMR-LIM thus has more in-sample skill at longer leads than the CCSM4-LIM. The LMR-LIM also remains significantly skillful for longer than the GCM-LIM remaining above the significance threshold for the full 20-year forecasts shown here.

The LMR-LIM has more persistence than the CCSM4-LIM, as noted by the difference in the AR1 forecast skill. The LMR-LIM's outperformance for these in-sample experiments may partially be explained by the additional persistence of the data it is calibrated on, yet all four indices in Figure 1 have higher skill than the AR1 model, indicating that the LIM is capturing additional dynamics on top of persistence. This is also true for the CCSM4-LIM at short leads, but the GCM-LIM forecast skill either becomes insignificant or matches persistence around 6- to 8-year leads for all four indices. This further suggests that the LMR-LIM better captures in-

sample internal variability than the CCSM4-LIM and this will be investigated for out-of-sample coupled data as well.

The LMR-LIM has been shown to overall reproduce important dynamical features for the in-sample forecasts, but it is important to consider the regional differences of this skill. Figure 2 shows the full-field air temperature correlation coefficient at 5-year lead.

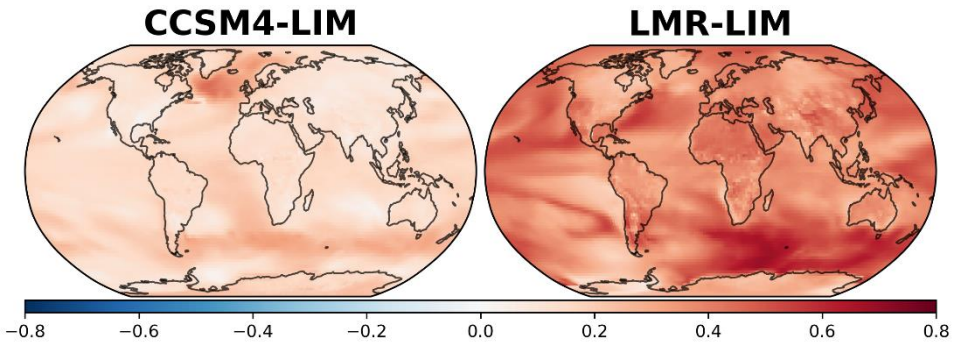

Figure 2: 5-year lead air temperature correlation coefficient for in-sample forecasts from the CCSM4-LIM (left; 855-1850 C.E.) and the LMR-LIM (right; 1005-1850 C.E.).

The LMR-LIM outperformance over the CCSM4-LIM is clearly evident in the air temperature field, represented by higher global correlation coefficient values. The CCSM4-LIM has maximum correlation located over the North Atlantic as well as some regionally high skill within the Southern Ocean storm tracks. The LMR-LIM has the best performance in the Southern Ocean as well as regionally high skill over the extratropical oceans. The Southern Ocean is thus an area of high air temperature skill at 5-year leads, most likely due to high persistence in this area. In addition, the Southern Ocean is a region of notably few proxies, so any signal there is attributed to the reference model's dynamics that is captured within the LIM's calibration. The CSM4-LIM does not perform as well in this region, hinting at large discrepancies between the LMR-data and the CCSM4 simulation data, perhaps tied to how information is spread through covariance weighting within the data assimilation scheme. The proxy-heavy locations are anticipated to contribute high skill in the LMR-LIM, yet this is not clearly evident in Figure 2 other than a small regionally skillful patch over the western US where there are abundant tree ring proxies. This may be due to the lack of persistence over land-based locations.

Sensitivity to Time Period for LIM training

The LMR-LIM is calibrated on data from 1000-1850 C.E. However, we test the LIM's sensitivity to proxy availability by considering the performance of a LIM that is trained on the LMR data truncated to the 200- and 400-year periods 1650-1850 C.E. and 1450-1850 C.E., respectively. These truncations exclude data based on proxy-sparse times before 1450 C.E. All versions of the LMR-LIM were run on the excluded LMR data from 1850-2000 C.E. for an out-of-sample assessment of the global mean sea surface temperature. For these temporally out-of-sample experiments, the 200-year LMR-LIM performed the worst, only remaining statistically significant until 4-year leads. The full time LMR-LIM had the highest correlation coefficient scores, with the 400-year LMR-LIM performing second-best. Therefore, we choose to run the full time LMR-LIM as opposed to truncated forms for all the following forecast experiments.

3.1.1 *Ocean Field Contribution to Forecast Skill*

The ocean is a large source of skill when forecasting on interannual to decadal timescales. The contribution of ocean fields to the forecast skill, specifically OHC, can be clearly seen within the LIMs. The most evident change in the model dynamics is between the EFT of the LIM's dynamical modes with and without OHC. The EFT of these modes changes significantly when ocean heat content is removed from the calibration. This is simple dynamics, stating that the models gain memory from the upper layer of the ocean due to its thermal inertia. However, the larger EFT of the LMR-LIM compared to the CCSM4-LIM suggests the potential of forecast outperformance on interannual and longer timescales due to increased memory pulled from the ocean fields. The improved dynamics of ocean states (as compared to OHC from offline DA; see PH21) may yield improved internal variability of LIM forecasts. This emphasizes the need to analyze the ocean field contribution to the forecast skill.

Error Growth

Figure 3 shows the error growth of individual variable fields for annual forecasts up to ten-year leads. The top row shows in-sample error growth when OHC is included in the LIM calibration while the bottom row shows variable error growth when there is no OHC within the calibration.

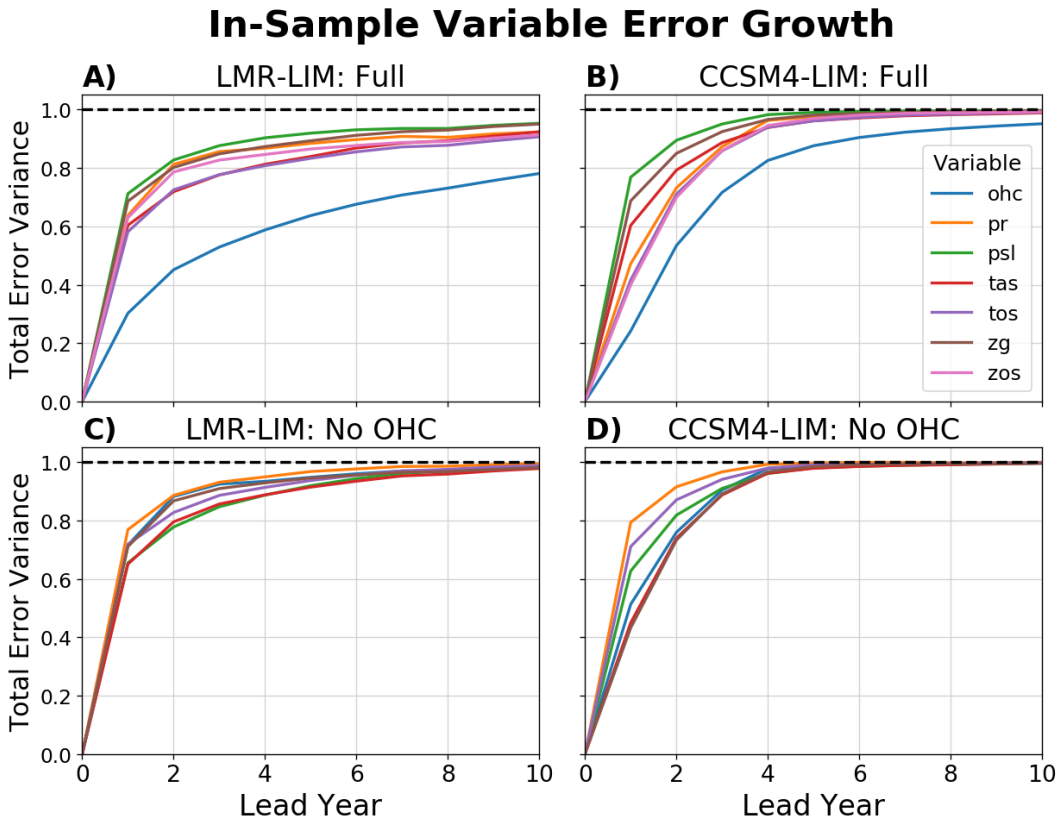

Figure 3: Total error variance growth by variable for the LMR-LIM (left) and CCSM4-LIM (right) in-sample forecasts. Bottom row is the same as the top row but without OHC in the LIM's calibration. Variables included are ocean heat content (ohc), precipitation (pr), surface level pressure (psl), surface air temperature (tas), ocean surface temperature (tos), 500hPa heights (zg), and dynamic ocean surface heights (zos). Climatological variance is represented by the horizontal dashed line at y=1.

The LMR-LIM with OHC has overall slower error growth than the CCSM4-LIM, allowing for skillful forecasts at longer leads. After removing OHC from the calibration as in Figure 3c, the LMR-LIM has increased error growth in all variables that reach climatological variance around 8-year leads. A comparison to Figure 3a shows that all variable fields within the full LMR-LIM do not reach climatology for the selected leads. Figure 3c thus displays a degradation in the forecast quality when there is no OHC within the LIM's calibration. Figure 3b shows that the CCSM4-LIM reaches climatological variance for all fields excluding OHC around 6-year leads. Figure 3d shows that the CCSM4-LIM only has a slight increase in variable field error growth when OHC is removed from the calibration, reaching climatological variance for all fields around 5-year leads. Overall, the LMR-LIM performance is dependent on whether or not OHC is

included in calibration while the CCSM4-LIM exhibits little sensitivity to variables within the calibration data set.

The error growth for both frameworks has little change when any other variable within the calibration is removed, including the sea surface temperature and ocean surface heights. These other ocean variables are overall noisier than the OHC and are more impacted by the atmosphere. This further highlights the need for consistent and reliable upper-ocean observational records to initialize, train, and verify models for decadal predictions.

3.2 Out-of-Sample Coupled Experiments

A proper comparison of framework performance requires out-of-sample data for verification. However, observational records are limited in terms of data extent and availability of coupled atmosphere-ocean fields. The only multi-centennial coupled data set is that provided by the LMR reconstructions. This second forecast experiment uses the withheld LMR data from 1850-2000 C.E. as verification, thus defining a temporally out-of-sample data set. Experiments that are out-of-sample in both time and data source are discussed in sections 3.3, 3.4, and 3.5. These experiments provide a comparison between the LMR-LIM and CCSM4-LIM for forecasts on out-of-sample LMR data. The variables included within each LIM's calibration are thus the same as in the in-sample coupled experiment.

Air Temperature Forecast Skill

Figure 4 shows the global mean air temperature skill calculated from full-field coupled forecasts. The clouds around the LIM forecasts represent the 95% significance threshold and thus provide a rough estimate of significant differences in the mean LIM performance between frameworks.

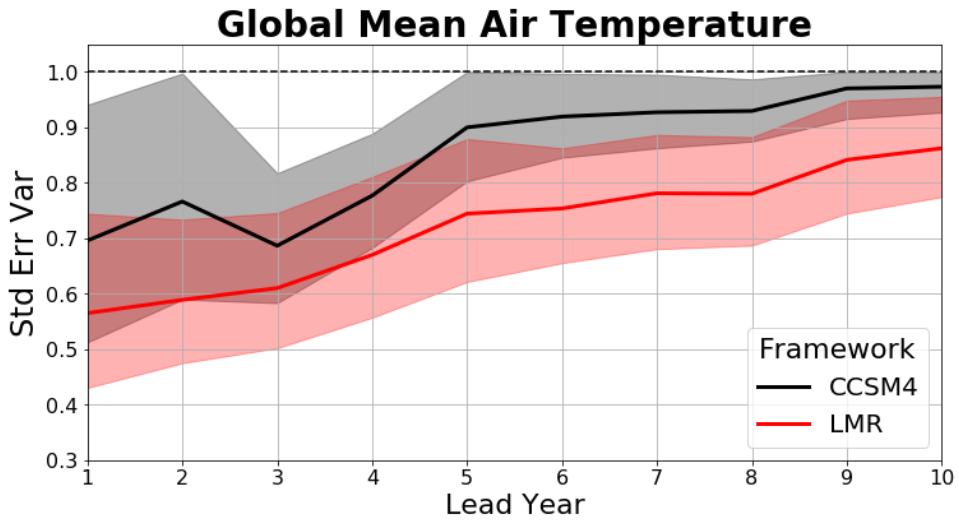

Figure 4: Standardized error variance growth of global mean air temperature for coupled forecasts on LMR-data from 1850-2000 C.E. CCSM4-LIM forecasts are in black and LMR-LIM forecasts are in red. Solid lines represent the mean LIM forecasts with clouds representing the 95% confidence threshold (with a lower bound of 2.5% and upper bound of 97.5%). The horizontal dashed black line is the climatological variance.

The LMR-LIM air temperature forecasts have overall lower error variance than the CCSM4-LIM's forecasts. The difference in the error variance is more significant at longer leads than at shorter leads, most notable beyond 5-year leads. The global mean SST forecasts also have the LMR-LIM outperforming the CCSM4-LIM at longer leads (not shown). However, the LMR-LIM air temperature forecasts have more statistically significant differences from the CCSM4-LIM forecasts than for the SST forecasts. The LMR-LIM outperforming the CCSM4-LIM implies that the LMR-LIM is better representing the climate system, yet it is important to note that this experiment is only temporally out-of-sample, and that commonalities exist between the calibration data set and the verification data set.

Figure 5 provides a look at the 5-year lead full-field air temperature correlation coefficient for this out-of-sample experiment.

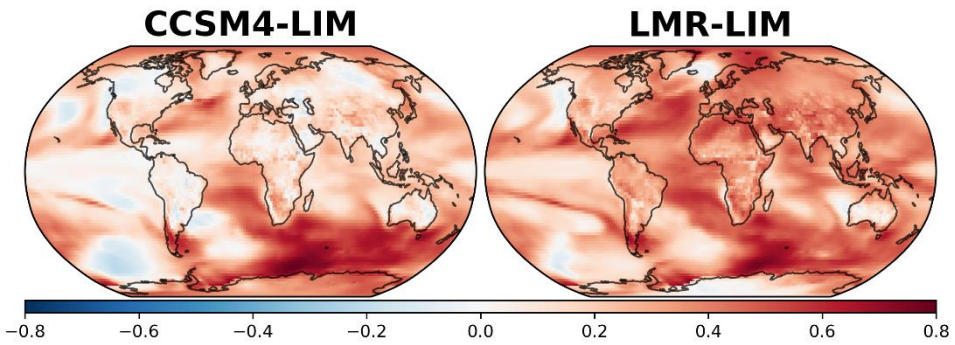

Figure 5: Same as in Figure 2 but for out-of-sample forecasts on the withheld LMR data from 1850-2000 C.E.

As in the in-sample experiments, the temperature over the Southern Ocean has high skill in limited regions. There are notable differences between the CCSM4-LIM and the LMR-LIM, mainly captured by regionally low skill over the extratropical Pacific Ocean and over North America in the CCSM4-LIM.

3.3 Out-of-Sample Experiments on Atmosphere-only Data

The LMR-LIM has skillful predictions on long timescales for in-sample assessment and for temporally out-of-sample LMR data. Here we assess performance on a separate data source that neither LIM has seen. Large samples of coupled data are not available due to the lack of observations prior to the instrumental data. There are, however, multiple instrumental products to perform experiments on, which is the focus of this section and section 3.4. While these data are uncoupled, the experiments allow for additional analysis of the LMR-LIM skill, its comparison to GCM-LIMs, and variable contribution to the overall skill.

This experiment considers verification data provided by GISTEMP and 20CR. The GISTEMP data is limited to surface air temperature and is thus referred to as the TEMP-only experiment. On the other hand, the 20CR has several available atmospheric variables on isobaric surfaces (see section 2.2.1) and is labelled the ATMOS-only experiment.

Air Temperature Forecast Skill

Figure 6 shows the global mean air temperature forecast for the TEMP-only calibration forecasting on GISTEMP (left) and for the ATMOS-only calibration forecasting on 20CR

(right). Once again, 95% confidence intervals are shown to highlight significant differences in model performance.

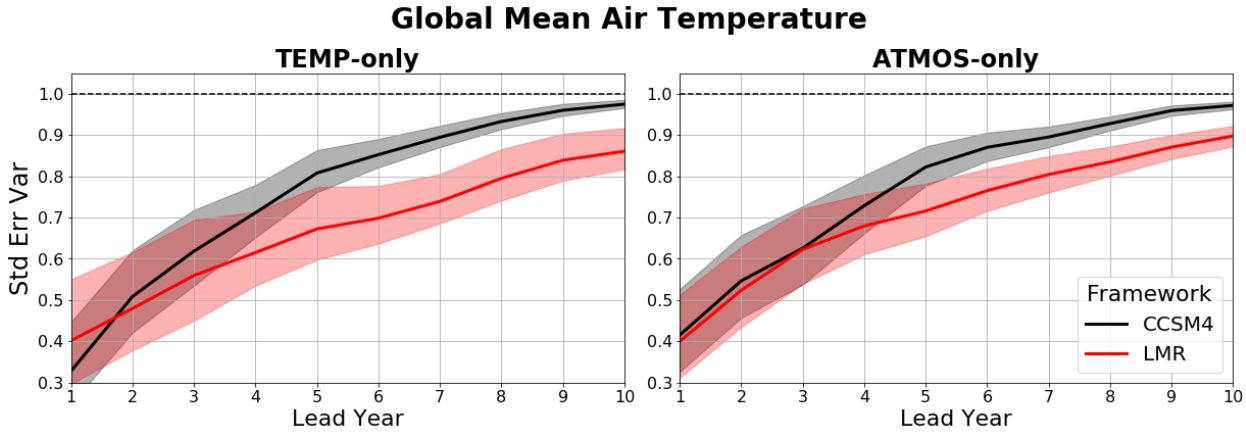

Figure 6: Same as in Figure 3 but for forecasts on GISTEMP (1880-2019 C.E) and 20CR data (1836-2015 C.E.). The forecasts on GISTEMP use a LIM calibrated on air temperature only (left) and the forecasts on 20CR use a LIM calibrated on four atmospheric variables (right).

The LIM's mean forecasts have similar error values and growth between the two forecasting experiments. The TEMP-only experiment shows that the CCSM4-LIM performs better at 1-year leads than the LMR-LIM. However, the LMR-LIM outperforms the CCSM4-LIM beginning at 2-year leads. The ATMOS-only experiment shows that both frameworks have similar skill up to about 4-year leads, where the LMR-LIM begins to outperform the CCSM4-LIM. The confidence bands of the LMR-LIM do not overlap with those of the CCSM4-LIM beginning around 5-year leads and beyond for both experiments, and the difference in the means are statistically significant. The LMR-LIM also outperforms the CCSM4-LIM in terms of a correlation coefficient (not shown), beginning at 4-year leads on 20CR and 3-year leads on GISTEMP, though these differences in the mean are not statistically significant at the 95% confidence level.

The LMR-LIM has skillful forecasts for longer than the CCSM4-LIM with the most statistically significant differences in error variance at the longest leads. Here, we only provide a brief outlook on the skill, yet we will further investigate the reason for LIM skill differences in Chapter 4.

As for the in-sample experiments, we test LIM sensitivity to the GCM calibration data, a 10 MVAR CCSM4-LIM, and the SPLMR-LIM for the same metric in Figure 6. The MPI-LIM performs similarly to the CCSM4-LIM when forecasting on GISTEMP, with smallest

improvement in forecast skill beginning at 4-year leads. The SPLMR-LIM performs slightly worse than the LMR-LIM at all leads beyond 1-year. However, the LMR-LIM consistently has the lowest error variance values at most leads. There is less sensitivity when forecasting on 20CR, with the SPLMR-LIM nearly identical to the LMR-LIM forecast skill metric and small differences between the GCM-analog LIMs. The reduced MVAR CCSM4-LIM performs slightly worse than the 25 MVAR CCSM4-LIM shown above, though this difference is insignificant. Therefore, LIMs exhibit little sensitivity to calibration methods and GCM dynamics for the atmosphere-only experiments described here. The LMR-LIM always has the best performance beyond 1-year leads regardless of these changes.

Figure 7 shows the air temperature spatial skill at 5-year leads for the TEMP-only experiment (top row) and ATMOS-only experiment (bottom row).

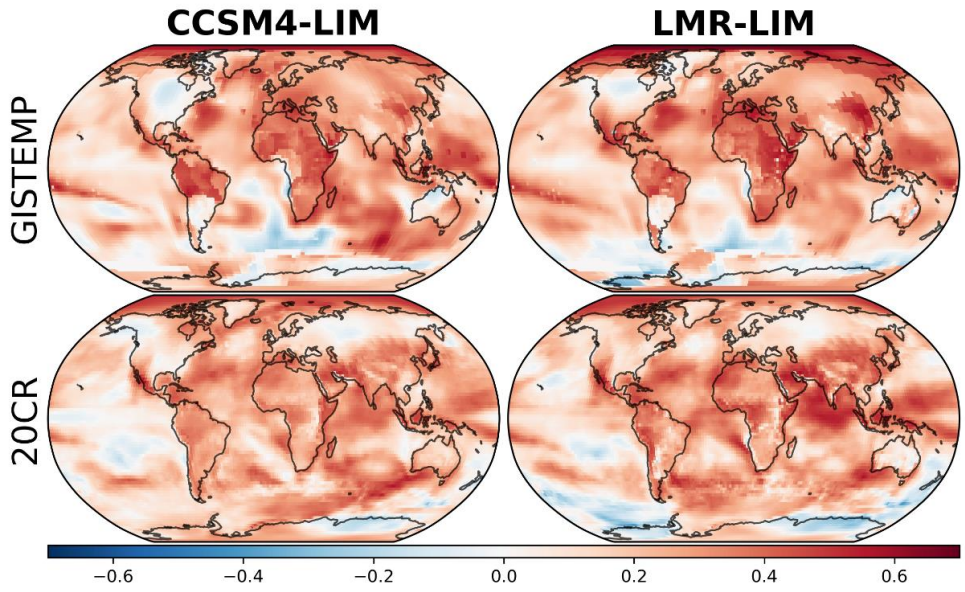

Figure 7: 5-year lead surface air temperature correlation coefficient using an LMR-LIM (left) and a CCSM4-LIM (right). The top row are the temperature-only LIMs forecasting on GISTEMP and the bottom row are the atmosphere-only LIMs forecasting on 20CR.

The TEMP-only experiments have similar regions of low and high skill between frameworks. The air temperature above the Southern Ocean tends to be unskillful in both frameworks, and North America also lacks skill. However, other continental regions such as Africa and the northern part of South America have regionally high skill, more notably in the GISTEMP results. This may suggest that the two data sets have a systematic difference in their verification statistics. For both the GISTEMP and 20CR, the air temperatures are more skillful over

continental regions than over most of the ocean basins. This contradicts the thought that the ocean is lending most of the air temperature field's skill through its high thermal inertia. Recall, though, that ocean variables are not included within the calibration and any influence of the ocean on air temperature shown here is due to the imprint during the construction of the data. We anticipate that including ocean variables in the calibration for out-of-sample experiments would result in the most skillful air temperatures over the ocean rather than over the continents.

The ATMOS-only experiments have more regional differences in skill between frameworks. Most of the LMR-LIM's poor skill is located in the Southern Ocean, which tends to be more skillful in the CCSM4-LIM. Both frameworks have regionally poor skill in the extratropical Pacific Ocean. The LMR-LIM slightly outperforms the CCSM4-LIM over the tropical Pacific and over North America. In addition, the LMR-LIM has better performance than in the CCSM4-LIM over some continental regions (e.g., over the southern portion of Eurasia).

3.3.1 *Adding an Inconsistent Ocean*

While these atmosphere-only experiments are skillful up to 10-year leads, they lack a coupled ocean which limits the LIM's calibration to include additional memory from ocean fields. The air temperature fields are clearly capturing persistence from the ocean, yet we anticipate that both models will see improved forecasts if ocean fields, specifically OHC, are included within the LIM's calibration

One solution is to provide ocean heat content initial conditions to the atmosphere-only instrumental experiments by adding inconsistent OHC to the GISTEMP and 20CR data sets. Here, inconsistency refers to an atmospheric component that is uncoupled from the ocean below in the initial state. Thus, the initial state exhibits large discrepancies between the atmosphere and ocean that a coupled system would otherwise not. With an inconsistent ocean, the LIMs are free to include OHC in their calibrations yet are verified on artificial or random OHC states. We consider two different types of inconsistent OHC fields. The first experiment adds zero OHC anomalies to the verification (i.e., artificial states). The second experiment adds OHC anomalies provided by LMR-data from the latest available time, which serve as random states. Table 1 summarizes the skill at 10-year leads for LIMs initialized with inconsistent OHC.

Table 1: 10-year GMT forecast correlation coefficient for 10-year forecasts spanning 1880-2019 C.E on GISTEMP and 1846-2015 C.E. on 20CR. Skill is provided for three experiments: 1) a full LIM that includes all available variables in the verification data set (i.e., No OHC added), 2) a full LIM that is initialized with zero anomaly OHC conditions (i.e., + 0 OHC), and 3) a full LIM that is initialized with OHC ICs from random states in the LMR data set (i.e., + LMR OHC).

Verified on:	**TEMP-only (GIS)**		**ATMOS-only (20CR)**	
Framework	CCSM4	LMR	CCSM4	LMR
No OHC added	0.47	0.58	0.50	0.60
+ 0 OHC	0.42	0.46	0.52	0.51
+ LMR OHC	0.46	0.18	0.54	0.28

The LMR-LIM correlation coefficient sees a large drop when the model is provided random OHC states. This decrease in skill is seen for both the TEMP-only calibrations and ATMOS-only calibration. There is a larger decrease in the LMR-LIM 10-year skill when providing inconsistent ICs from the LMR-data itself as compared to providing 0 OHC anomalies. When initialized with the LMR OHC fields, the LMR-LIM air temperature forecast correlation drops by over 50% for both calibrations. The LMR-LIM sensitivity to OHC initialization is larger for a LIM calibrated on air temperature only as compared to the multivariate atmosphere-only calibration. The LMR-LIM thus requires large samples of consistent OHC fields to initialize the model for optimized decadal forecasting.

The CCSM4-LIM sees little change in GMT performance regardless of whether artificial or random OHC states are provided or not. The TEMP-only calibration has slight drops in skill when provided inconsistent OHC, while the ATMOS-only calibration has slight increases in the skill.

The impacts of OHC were also studied when considering an LMR-LIM that retains 10 OHC EOFs as opposed to 20 OHC EOFs, resulting in a MVAR to OHC EOF ratio of 10/10 whereas the CCSM4-LIM ratio is constant at 25/20. Overall, reducing the retained OHC EOFs by half leads to a lower correlation coefficient than the values in Table 1 (e.g., for 20CR, +0 OHC drops from 0.51 to 0.42 and +LMR OHC drops from 0.28 to -0.04). Therefore, making the ratio of multivariate EOFs retained to OHC EOFs retained closer to 1 increases the LMR-LIM sensitivity to OHC initialization. This further highlights the importance of OHC fields in the LMR-LIM relative to the CCSM4-LIM.

3.4 Out-of-Sample Experiments on Ocean-only Data

There are instrumental sources of ocean data that come from reanalysis products such as SODA, HadleyEN4, GECCO3, and ORAS4 (Carton et al., 2018, Good et al., 2013, Kohl, 2020, and Balmaseda et al., 2013, respectively). Here, we conduct forecast experiments on SODA and HadleyEN4 since these two records provide the longest samples of ocean fields to initialize and verify forecasts. The SODA and HadleyEN4 products provide several ocean variables, including OHC, SST, sea surface salinity, and sea water potential temperature, yet only OHC and SST are available in the LMR and CCSM4 data for calibration. Therefore, experiment 4 considers LIMs with only these two variables in the calibration.

Sea Surface Temperature Forecast Skill

Figure 8 shows the forecast ACC for ocean-only experiments. There is little significant skill when analyzing the SEV, so we present the correlation coefficient because it is a less strict metric and shows more statistical significance for longer.

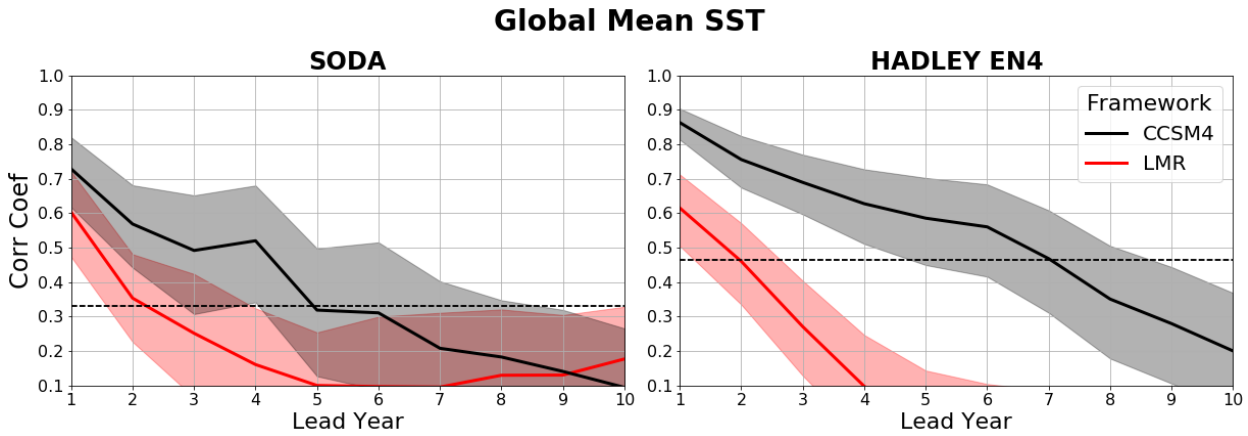

Figure 8: Correlation coefficient of global mean sea surface temperature forecasts on ocean-only instrumental data. Solid lines are mean LIM forecasts for the LMR-LIM (red) and CCSM4-LIM (black) trained on OHC and SST. Both LIMs forecast on the SODA (1871-2008 C.E.) and the HADLEY EN4 (1900-2010 C.E) data sets. Clouds represent the 95% confidence bounds and the dashed red line is the 95% significance threshold for both frameworks.

Both LIMs have skillful forecasts for a short time, with the CCSM4-LIM dropping below the significance threshold before 5-year leads on SODA and at 7-year leads on HadleyEN4. The LMR-LIM has less skill than the CCSM4-LIM as well as insignificant forecasts beginning as

early as 2-year leads on SODA and HadleyEN4. However, it is important to note that the significance threshold is much higher for these ocean-only experiments than for the in-sample forecasts because the LIM forecasts are being verified and initialized with a small sample of data (in this case, about 100 years). This highlights the need for large samples of data for proper analysis of the LIM's performance.

The LMR-LIM has less skill than the CCSM4-LIM for these ocean-only experiments, though previous experiments involving LMR-LIM's ocean fields show the opposite result. One possibility for this discrepancy is disagreement between ocean-only products due to sparse and unreliable ocean measurements prior to the 21st century (e.g., Palmer et al., 2017). Thus, the LMR-LIM may have unskillful forecasts either due to the model capturing the wrong dynamics or the verification data itself failing to represent the true, observable ocean fields. In addition, having an ocean coupled to the atmosphere better constrains ocean fields. We address these issues in section 3.5 by comparing the LIMs on truly out-of-sample coupled historical data.

Figure 9 shows the 1-year full-field sea surface temperature correlation for the ocean-only calibrations forecasting on the Hadley EN4 (top row) and on the SODA (bottom row).

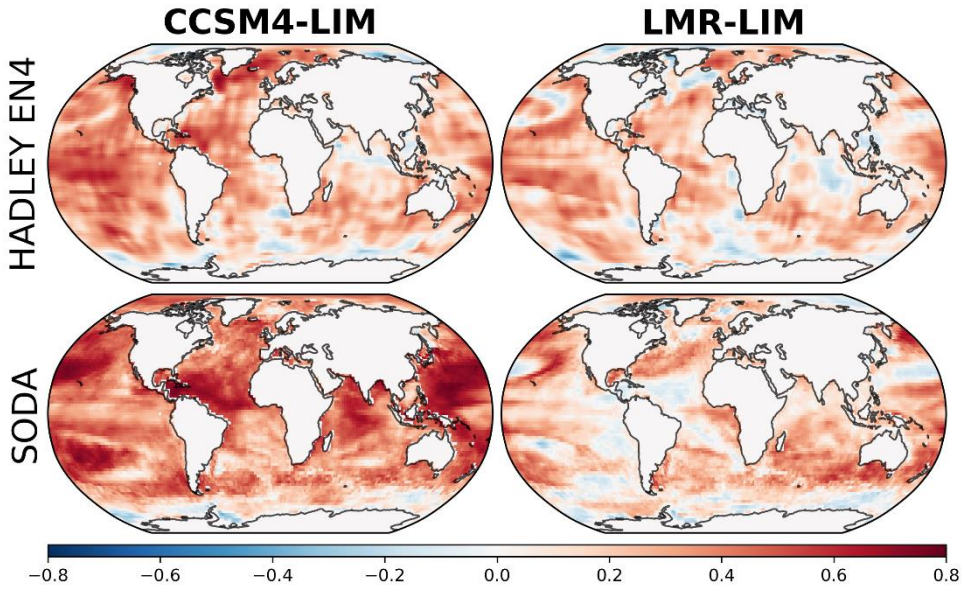

Figure 9: 1-year spatial correlation coefficient of sea surface temperatures using a CCSM4-LIM (left) and the LMR-LIM (right). Results are shown for forecasts on the HADLEY EN4 (top row) and SODA (bottom row) data sets.

The spatial results shown here are the least skillful and significant of any of the previous forecasting experiments. The CCSM4-LIM clearly outperforms the LMR-LIM as early as 1-year leads. Both frameworks have similar unskillful regions in the Southern Ocean, yet the CCSM4-LIM has higher skill in all other ocean basins as compared to the LMR-LIM. Past 1-year leads, the forecast skill is highly insignificant, more so than any of the past experiments. This lack of skill illustrates that the LIMs may have different dynamics than those implied by the ocean analyses. Furthermore, the lack of agreement between OHC within ocean analysis products (correlations between 0.6-0.7) point to a little understanding of upper-ocean dynamics within reanalysis products.

3.5 Historical Simulation Experiments

The most useful form of verification for comparing LIM performance is on out-of-sample data. Experiment 2 considered temporally out-of-sample data, though the source of this data was the same as what one of the LIM's was calibrated on. On the other hand, experiments 3 and 4 were both temporally out-of-sample and from a source that neither LIM has seen before. However, both experiments were uncoupled, which is unpreferred for evaluating the LMR-LIM performance as the highlights of this model are its ability to pick up low-frequency variability of the coupled system. These experiments therefore fail to fully illustrate the LMR-LIM's novelty.

Historical simulations provide completely out-of-sample as well as a coupled system for LIM verification. Experiment 5 compares the LIM performance between framework for forecasts on a single historical simulation provided by the GISS-2E-R model. The simulation has all available variables as within the LMR-LIM, so the models are trained on both atmospheric and oceanic data.

Air Temperature Forecast Skill

The first comparison of skill for global mean air temperature forecasts is presented in Figure 10. It is important to note that, while not shown, both LIM forecasts converge towards the climatological value of 1 as the forecast time approaches 20-year leads.

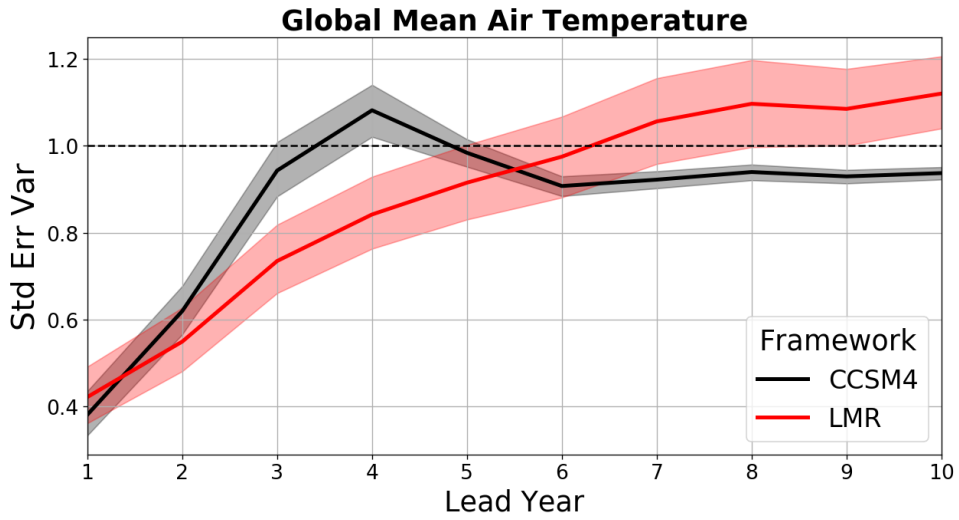

Figure 10: Same as in Figure 3 but for forecasts on GISS-2E-R 1951-2005 C.E model simulation data. These forecasts use LIMs that are trained on fully coupled data as for the in-sample experiments.

The LMR-LIM outperforms the CCSM4-LIM for 2 through 5-year lead forecasts yet the CCSM4-LIM experiences a recovery of forecast skill for longer leads. The CCSM4-LIM has a drop-off in skill at 4-year leads associated with the clockwork ENSO signal. However, this ENSO mode quickly decays past 4-year lead forecasts and persistence takes over, leading to a slight recovery of skill at 5-year leads. The LMR-LIM outperformance at 4-year leads points to proxies pulling away from the improper model dynamics of the prior and overall, more reliable forecasting.

Another interesting aspect of the LMR-LIM forecast skill is the overshoot of SEV beyond 6-year lead forecasts. The overshoot is most likely due to long-timescale oscillations that cause skill to vary about the climatological variance, a signal that does not appear within the CCSM4-LIM forecasts. In addition, SEV larger than the climatological variance is associated with the differences in ratio of variance between the training data and verification. These experiments therefore illustrate the difference of variance within the training data sets between CCSM4 and the LMR, where the LMR data has larger variance than the CCSM4.

We next examine the spatial properties of air temperature forecasts between frameworks by considering the gridded correlations presented in Figure 11.

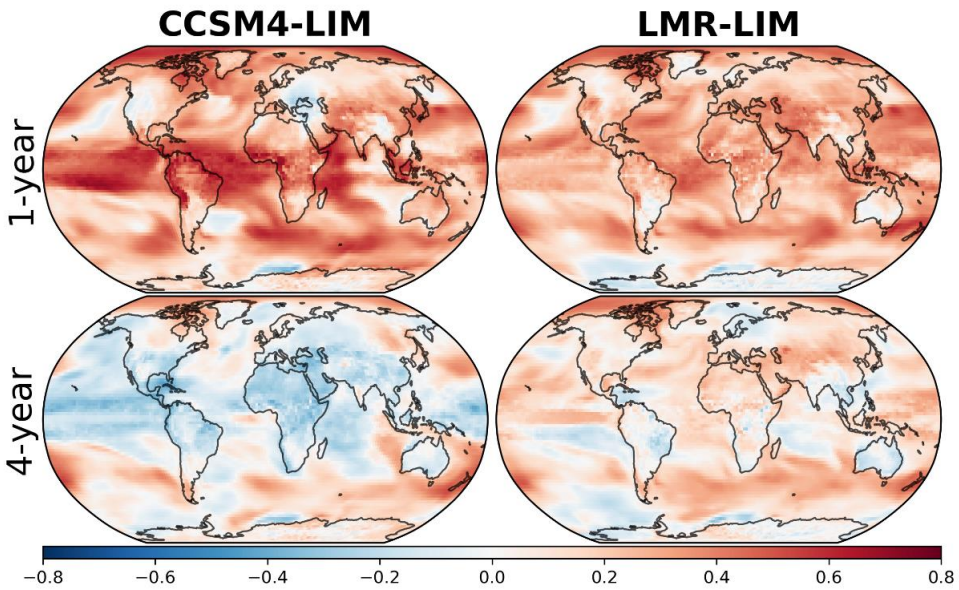

Figure 11: Gridded correlation coefficient values for 1-year air temperature forecasts (top row) and 4-year air temperature forecasts (bottom row) provided by the CCSM4-LIM (left) and the LMR-LIM (right). Forecasts are verified against the GISS-2E-R 1851-2005 C.E. model simulation data.

The CCSM4-LIM slightly outperforms the LMR-LIM at 1-year leads. This outperformance is clear when comparing correlations between frameworks in tropical regions, with the CCSM4-LIM having better performance. There are also regions of more skill in the Southern Ocean as well as parts of the Arctic. However, the LMR-LIM does have better performance over Northern continental regions such as over North America and Eurasia. Figure 11 further illustrates the CCSM4-LIM's dip in skill associated with ENSO at 4-year leads. As expected, there are low correlation values within the tropical Pacific that extends into the extratropical regions. The LMR-LIM has better skill here, but also improved forecasts over most of the continental regions. The continental air temperature skill could either be due to helpful information that the proxies contribute or tied to persistence through teleconnections.

Chapter 4. SOURCES OF SKILL

We next examine the linear modes of the LIM to understand the dynamical sources of forecast skill. In this chapter, we will focus on the atmosphere-only experiments and interpret the physical mechanisms that contribute to the forecast skill. The atmosphere-only experiment is selected because it is out-of-sample in regards to both the data source and time in addition to having the highest forecast performance amongst all other out-of-sample experiments. Section 4.1 provides an explanation on how we obtain dynamical modes and their properties. Section 4.2 investigates each dynamical mode's contribution to the overall forecast skill by considering single mode forecasts. Section 4.3 describes the properties of the most skillful mode and provides physical interpretation of the mode's air temperature pattern.

4.1 EMPIRICAL NORMAL MODES

Empirical normal modes (ENMs) are solutions to the LIM's deterministic update equation (Equation 1 without the forcing term). ENMs take the general form $\mathbf{e}_j \exp(\beta_j t) c_j$ where **e** represents the j^{th} eigenvector of **L**, β is the j^{th} eigenvalue, and c_j is an arbitrary complex constant. The eigenvectors and eigenvalues are recovered via an eigendecomposition of the linear operator matrix, **L**:

$$\mathbf{L} = \mathbf{E}\boldsymbol{\Lambda}_L\mathbf{E}^{-1} \tag{8}$$

Here, **E** is a matrix with the eigenvectors of **L** as columns, and $\boldsymbol{\Lambda}_L$ is a diagonal matrix containing the eigenvalues of **L**. The propagation matrix, $\mathbf{G}_\tau$, from equation 2 shares the same eigenvectors as **L**, but the eigenvalues are related by $\lambda_g = e^{\lambda_l \tau}$. Both the eigenvectors, **e**, and eigenvalues, λ, of **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 **L**: the decay time and the period. The decay time is calculated as $-\frac{1}{\sigma}$ and the period is $\frac{2\pi}{\omega}$, where the real and imaginary parts of the eigenvalue are σ and ω, respectively. The real part of the eigenvalue must be negative to maintain stationary statistics. Thus, all modes are damped and decay over time. However, ENMs are non-orthogonal and free to interfere in combination with one another. The

system may then experience transient anomaly growth (decay) via constructive (destructive) interference between modes (e.g., Farrell and Ioannou, 1995; Alexander et al., 2008).

Given the eigen-decomposition of **L**, any solution of the LIM can be written as a linear combination of the ENMs. Both members of complex conjugate pairs are required to retrieve a full solution. Here, we are interested in only the real solutions:

$$\mathbf{x}_j(t) = \{\mathbf{a}_j \cos(\omega_j t) + \mathbf{b}_j \sin(\omega_j t)\} \exp(\sigma_j t) \tag{9}$$

In the above equation, coefficients $\mathbf{a}_j$ and $\mathbf{b}_j$ contain information from the real and imaginary part of the eigenvector, respectively. Equation 8 considers both members of the complex conjugate pair and provides a measure of the oscillation frequency as well as the EFT for each mode. Complex-conjugate pairs share the same decay time as well as the same zero and quarter-period patterns due to their phase-quadrature relationship (i.e., one member is associated with a cosine function and the other member is associated with a sine function). There are special cases where the eigenvectors do not come as a complex conjugate pair. These eigenvectors are stationary and thus represented as pure exponential decay (i.e., $\mathbf{b}_j = 0$ and $\omega_j = 0$).

4.2 Single ENM Experiments

The LIM's dynamics are explained by non-orthogonal modes that oscillate and decay in time. We may investigate the contribution of an individual ENM to the overall forecast skill by isolating single eigenvectors prior to forecast verification. This dynamical mode breakdown allows the identification of ENMs that have large contributions to the forecast skill. It is important to note that single ENM experiments consider both members of complex-conjugate pairs, as both members are required to define the mode's oscillation.

An equation for forecasts represented in LIM space is attained by substituting the eigen-decomposition of $\mathbf{G}_\tau$ into equation 2.

$$\delta\hat{\boldsymbol{x}}_\tau = \boldsymbol{\Lambda}_G \delta\hat{\boldsymbol{x}}_o \tag{10}$$

In the above equation, $\delta\hat{\boldsymbol{x}}_n = \boldsymbol{E}^{-1}\delta\boldsymbol{x}_n$ is a projection of the state onto the eigenvector matrix of $\mathbf{G}_\tau$ where n is a time subscript. The forecast is thus a reweighting of the projection of initial conditions onto the eigenvectors. We may isolate an eigenvector in equation 9 to run single ENM experiments by simply using a column from the forecast $\delta\hat{\boldsymbol{x}}_\tau$ prior to verification. Skillful ENMs

are those with the lowest error variance values (largest correlation coefficients), defined as the smallest (largest) sum of the error variance (correlation coefficient values) across all leads. The ENMs are not orthogonal, so the skill from individual ENMs will not sum to the total correlation.

We conduct single ENM experiments on the atmosphere-only experiments for GMT forecasts in order to diagnose the sources of skill of the out-of-sample results with the best LIM performance. We use SEV as in Chapter 3.3 to assess individual ENM skill and compare to the full GMT SEV. In addition to individual ENM experiments, we run two ENM experiments with the top two ENMs, allowing for interactions between the most skillful modes. Figure 12 summarizes the single and two-ENM experiments. In this figure, ENM 1 is the most skillful ENM for the GMT metric and ENM 2 is the second most skillful, though the labels 1 and 2 are arbitrarily defined.

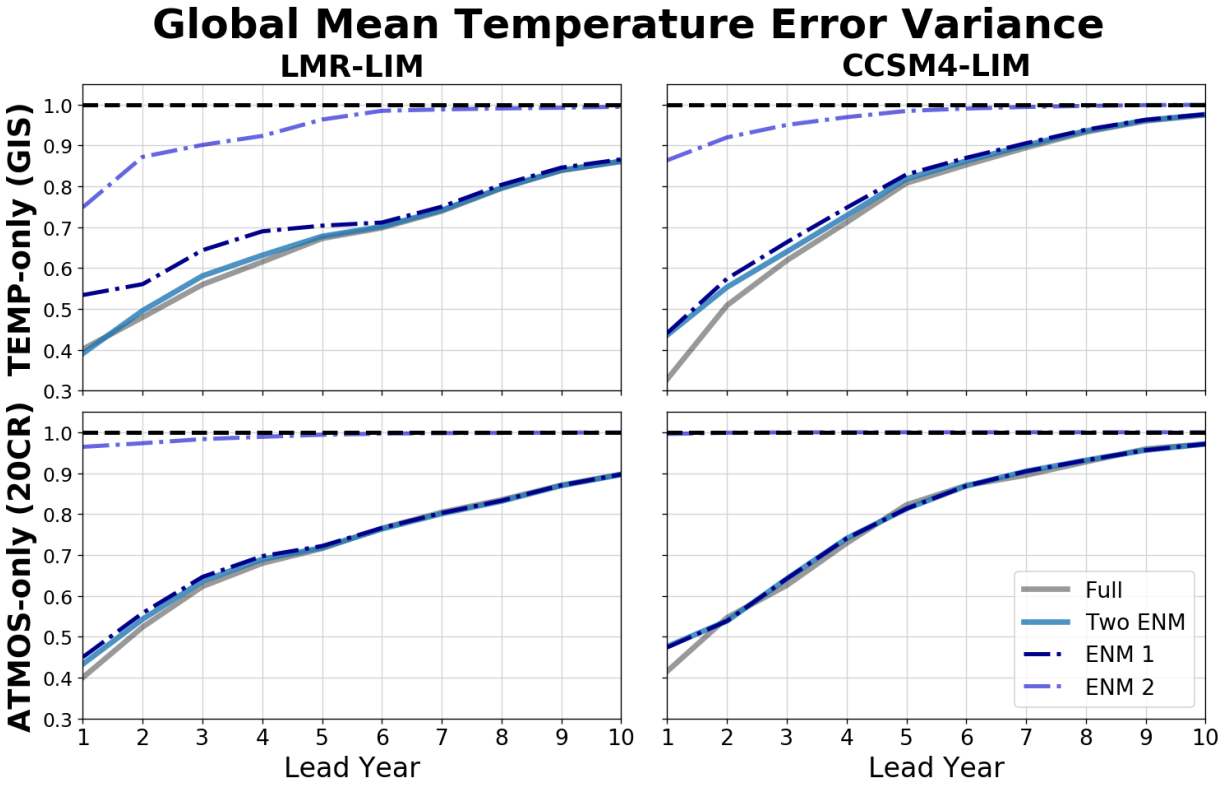

Figure 12: Global mean temperature error variance growth for ENM experiments using the LMR-LIM (left column) and CCSM4-LIM (right column) to forecast on GISTEMP (top row) and 20CR (bottom row). The LIM is run with single ENMs and the top two skillful ENM forecasts are shown in dot dashed lines. The top two skillful ENMs are run together (opaque blue line) and compared to the full ENM forecast (opaque black line). The dashed black line represents climatological variance (i.e., when the forecast is indistinguishable from climatology).

The most skillful ENM in the ATMOS-only experiments nearly captures the same SEV as in the full ENM forecast. The second most skillful ENM has very little skill and is at climatological variance for most leads. Combining the two most skillful ENMs only has slight improvement over the ENM 1 forecast in the LMR-LIM. ENM 1 in the TEMP-only experiments resembles the SEV of the full forecast, more so at longer leads, but there is more contribution from ENM 2 than in the ATMOS-only experiments. For both experiments, the largest discrepancies between the two-ENM and full forecasts tend to be during the early leads when the skill relies more on nonlinear interactions between noisier modes of variability.

Both LIMs are able to nearly approximate the full ENM forecast skill with just a single mode of variability. There is more improvement in the LMR-LIM skill when including a second ENM than in the CCSM4-LIM. However, the 1-2 ENM approximation of the full forecast skill may not hold when considering other indices outside of the global mean, where more ENMs may be required to explain the system's dynamics.

4.3 Physical Mechanisms

Properties of skillful ENMs

Figure 13 summarizes the ENM properties for the TEMP-only and ATMOS-only calibration, where stars represent the most skillful ENMs found in the previous section. Note that the ENM properties are purely dependent on the calibration and do not take the verification data sets into consideration.

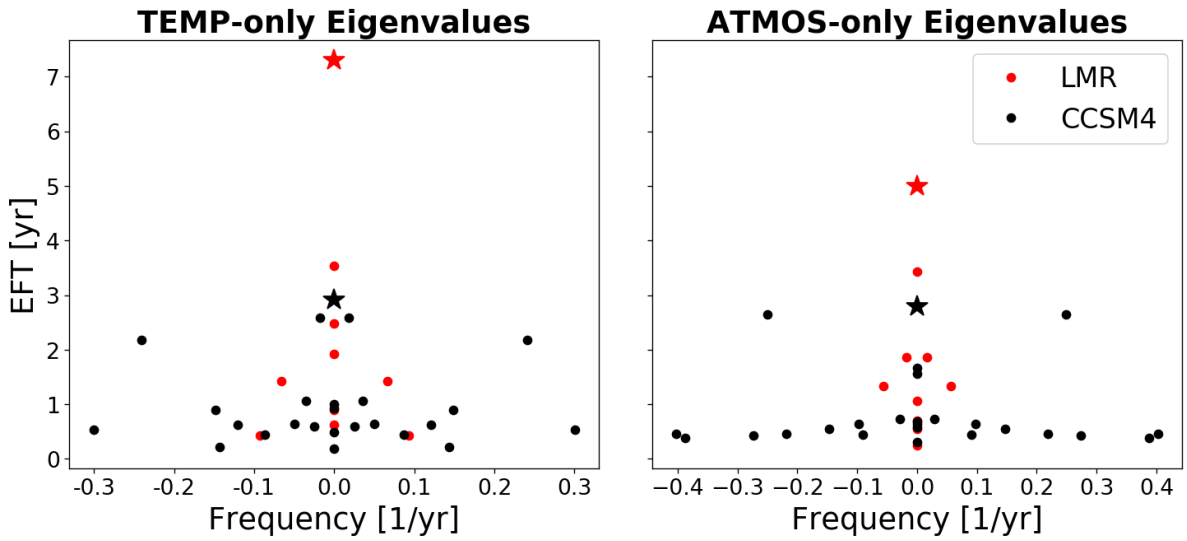

Figure 13: ENM properties of the LMR-LIM (red) and CCSM4-LIM (black) that are trained on a) surface air temperature and b) surface air temperature, sea level pressure, precipitation, and 500hPa heights. The most skillful ENMs found from single ENM experiments in section 4.2 are marked by stars.

The most skillful modes exhibit two common features. First, they are stationary modes that do not propagate over time. Second, they are the least-damped modes within the system, defined as having the largest EFT. A comparison to the AR1 forecast supports the fact that the global forecast skill is mostly persistence. For the ATMOS-only forecasts, the total forecast skill is almost identical to damped persistence, regardless of framework. This is also true for the CCSM4-LIM TEMP-only forecasts. The LMR-LIM does outperform the AR1 forecast, but the LMR-LIM also shows the most disagreement between the one ENM and full ENM forecasts.

The difference in the most skillful mode's EFT between the LMR-LIM CCSM4-LIM is represented by the vertical difference between the stars on each plot. This explains the LMR-LIM outperformance for the atmosphere-only forecast experiments, because the LMR-LIM has higher persistence in its least damped mode than the CCSM4-LIM, even when ocean fields are excluded from the calibration. The persistence within the atmosphere-only calibrations suggests that the ocean is lending skill to the atmosphere not only through the LIM calibration but in the construction of the data itself. Therefore, the atmosphere is influenced more by the ocean fields in the LMR data than in the CCSM4 last millennium simulation. This conclusion is supported by the inconsistent OHC results from section 3.3.1., where the LMR-LIM was more sensitive to inconsistent OHC initialization than the CCSM4-LIM.

Skillful ENM Air Temperature Patterns

Figure 14 shows the most skillful ENM as an air temperature pattern for the TEMP-only calibration (top row) and the ATMOS-only calibration (bottom row). All values are standardized such that the maximum absolute value is 1 and amplitudes are thus arbitrary.

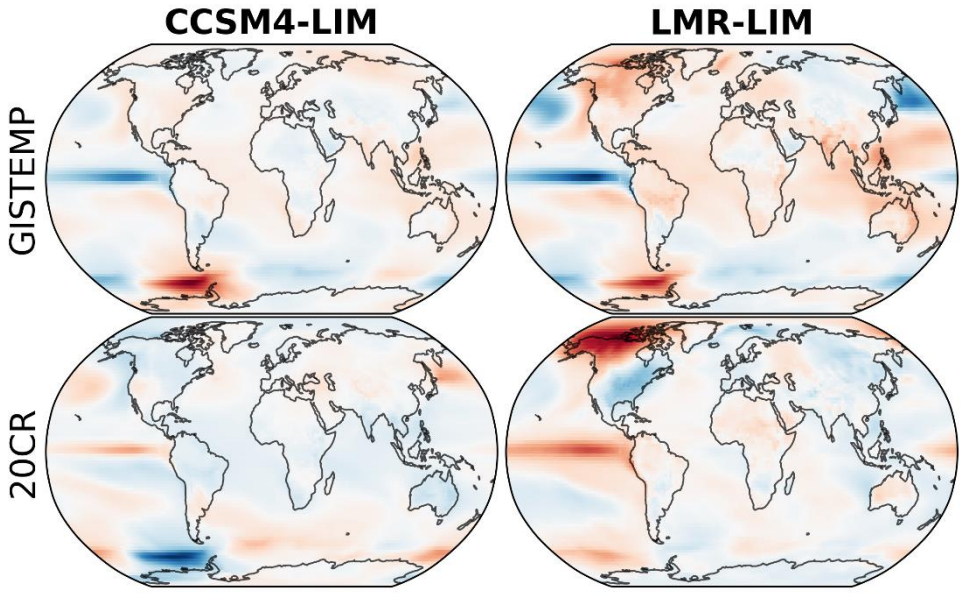

Figure 14: Most skillful ENM patterns represented in the air temperature field. Results are shown for the temperature-only calibration (top row) and the atmosphere-only calibration (bottom row) for the CCSM4-LIM (left column) and the LMR-LIM (right column). Each panel is a single phase of the stationary modes (4 modes in total). All data are standardized by the largest amplitude to have a maximum value of +/-1.

The LMR-LIM for the TEMP-only calibration has an ENSO-like signal with an overlapping PDO-like signal of opposite sign. In addition, the signal in the Southern Ocean represents a teleconnection of the Southern Annular Mode (SAM) with ENSO. Positive phases of ENSO drive warming trends over the Southern Ocean (Ferster et al., 2018), which tend to be strongly centered over the Amundsen Sea low as part of the influence of SAM (Turner et al., 2012). The CCSM4-LIM for the TEMP-only calibration has a similar pattern to the LMR-LIM, yet with reduced magnitude in the tropical and extratropical latitudes.

The CCSM4-LIM ENM for the ATMOS-only calibration has nearly the same pattern as the TEMP-only ENM yet with the opposite sign. On the other hand, the LMR-LIM has a different pattern to the new calibration. There is still an ENSO-like pattern and a weak PDO-like pattern, yet the strongest signal resembles the Pacific North Atlantic (PNA) pattern. The PNA oscillates

on multidecadal timescales between positive and negative patterns as defined by anomalies in the geopotential heights over the US (Trenberth and Hureel, 1994). Including geopotential heights may have helped define PNA as an important mode of variability within the 20CR experiments. Furthermore, the PNA exhibits strong teleconnections to ENSO and PDO, with the strongest PNA occuring when ENSO and PDO are in-phase (Yu and Zwiers, 2007; Trouet and Taylor, 2009). While the PDO pattern represented by the LMR-LIM skillful ENM is weak, the PDO and ENSO are in-phase, which may explain why the PNA has the largest variability within this mode. All the other skillful ENMs in figure 12 fail to capture an ENSO and PDO that are in-phase.

The Interdecadal Pacific Oscillation (IPO) and Atlantic Multidecadal Oscillation (AMO) are the leading modes that drive GMT on decadal timescales (Power et al., 1999; Kushnir, 1994). However, there appear to be little to no signals of AMO in the skillful ENMs, and the only pattern consistent with an IPO is the LMR-LIM skillful ENM on the 20CR data. Both data sets cover the global surface warming hiatus from 1998-2012. The hiatus is attributed to a transition of IPO from a positive to negative phase as well as a shift in the AMO from a negative to a positive phase that is associated with a La-Nina pattern in the tropical Pacific (Liu and Xie, 2018). However, the skillful ENMs picked up on the persistence rather than the oscillatory nature of internal variability.

The LMR-LIM skillful ENMs thus resemble some combination of ENSO, PDO, and the PNA. It is important to note that while these patterns may resemble internal variability, they are stationary and do not propagate in time. One potential is that volcanic eruptions have imprinted persistence into the system by providing long-term globally cooler temperatures. The patterns of air temperature following volcanic eruptions can be quite complex and exhibit seasonal dependencies as well as rely on the response of internal variability to the eruptions (Fujiwara et al., 2020).

Chapter 5. CONCLUSIONS

A linear inverse model is a way of approximating the climate system as linear processes plus stochastic white noise forcing. This model is able to capture large-scale dynamics of the underlying model that it is trained on and has been previously shown to have comparable skill to GCMs (e.g., Newman 2013). Calibrating a LIM on proxy-informed data results in more "realistic" dynamics due to weighting towards actual climate variability within the data assimilation scheme. Thus, a comparison of a LIM trained on the paleo-data from PH21 to GCM-LIMs highlights the usefulness of the proxies within the LIM calibration.

A proxy-informed LIM outperforms GCM-LIMs for coupled in-sample (experiment 1), temporally out-of-sample and coupled (experiment 2), and out-of-sample forecasts on atmosphere-only instrumental data (experiment 3). In addition, experiments on historical simulations (experiment 4) illustrated the LMR-data's ability to correct for the improper dynamics of the reference GCM. In this case, the LMR-LIM had better forecast skill on interannual timescales due to the proxy weighting away from the ENSO clockwork signal of the CCSM4 simulation data.

The LMR-LIM skill for all experiments depends on reliable ocean variable fields and exhibits high sensitivity to having consistently initialized states, specifically for OHC. Removing OHC from the LMR-LIM results in increased error growth for all other variables as well as a decrease in the EFT (i.e., the system memory) of the LIM's dynamical modes. The LMR-LIM has greater low-frequency variability from ocean states that allow for high performance at longer leads and slower error growth overall, though results from Chapter 4 also illustrate that the LMR-LIM outperformance over the GCM-LIMs is partially due to the LMR-LIM being calibrated on highly persistent fields.

The atmosphere-only experiments' global mean forecasts can be approximated by a single ENM, though fully coupled systems may require more ENMs as these systems are more complex. Regardless of the calibration, the most skillful ENM was the least damped mode and stationary. Further analysis of the most skillful ENM of the LMR-LIM shows an air temperature pattern that resembled some combination of ENSO, PDO, and PNA. However, the stationary mode could clearly not contain internal variability and only captures persistence. Thus, the most skillful LMR-patterns seem to be influenced by some forcing that added persistence to the

system, such as volcanic eruptions. The next immediate step for this project will be to quantify the projection of initial conditions onto the skillful ENM spatial patters. This will allow for a better comparison between dynamical modes and support further interpretation of sources of skill.

All experiments highlight the need for large samples of consistent coupled ICs for skillful forecasts on decadal timescales. Climate field reconstructions should play a key role in further decadal predictability studies, as they provide full-fields of coupled atmosphere-ocean data that capture large spreads of uncertainties in initial conditions, as well as provide continuous samples to initialize models and verify forecasts. A potential future endeavor is to consider a LIM trained on paleo-informed data retrieved using a superprior (i.e., a multimodel ensemble) as opposed to a single GCM prior, which would allow for consideration in the spread of model internal variability. Furthermore, including radiation variables may be useful for forecasting Earth's energy budget through time and provide an outlook on ocean heat uptake.

REFERENCES

Alexander, M. A., L. Matrosova, C. Penland, J. D. Scott, and P. Chang, 2008: Forecasting Pacific SSTs: Linear Inverse Model Predictions of the PDO. *J. Climate*, **21**, 385-402, https://doi.org/10.1175/2007JCLI1849.1.

Balmaseda, M. A., K. E. Trenberth, and E. Kallen, 2013: Distinctive climate signals in reanalysis of global ocean heat content. *Geophys. Res. Lett.*, **40**, 1754-1759, https://doi.org/10.1002/grl.50382.

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.

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

Carton, J. A., G. A. Chepurin, and L. Chen, 2018: SODA3: A New Ocean Climate Reanalysis. *J. Climate*, **31**, 6967-6983, https://doi.org/10.1175/JCLI-D-18-0149.1.

Chu, P. C., 1999: Two Kinds of Predictability in the Lorenz System. *J. Atmos. Sci*., **56**, 1427-1432, https://doi.org/10.1175/1520-0469(1999)0562.0.

Delworth, T., S. Manabe, and R. J. Stouffer, 1993: Interdecadal variations of the thermohaline circulation in a coupled ocean-atmosphere model. *J. Climate*, **6**, 1993-2011, https://doi.org/10.1175/1520-0442(1993)006<1993:IVOTTC>2.0.CO;2.

Deser, C., M. A. Alexander, S. Xie, and A. S. Phillips, 2010: Sea Surface Temperature Variability: Patterns and Mechanisms. *Annu. Rev. Mar. Sci*., **2**, 115-143, https://doi.org/ 10.1146/annurev-marine-120408-151453.

Farrell, B. F., and P. J. Ioannou, 1995: Stochastic Dynamics of the Midlatitude Atmospheric Jet. *Journal of the Atmospheric Sciences*, **52**, 1642-1656, https://doi.org/10.1175/1520-0469(1995)052%3C1642:SDOTMA%3E2.0.CO;2.

Farnetti, R., 2017: Modelling interdecadal climate variability and the role of the ocean. *WIREs Clim. Change*, **8**, https://doi.org/10.1002/wcc.441.

Ferster, B. S., B. Subrahmanyam, and A. M. Macdonald, 2018: Confirmation of ENSO-Southern Ocean Teleconnections Using Satellite-Derived SST. *Remote Sens.*, **10**, 331, https://doi.org/10.3390/rs10020331.

Foster, D., D. Comeau, and N. M. Urban, 2020: A Bayesian Approach to Regional Decadal Predictability: Sparse Parameter Estimation in High-Dimensional Linear Inverse Models of High-Latitude Sea Surface Temperature Variability. *J. Climate*, **33**, 6065-6081, https://doi.org/10.1175/JCLI-D-19-0769.1.

Frankcombe, L. M., M. H. England, M. E. Mann, and C. A. Steinman, 2015: Separating Internal Variability from the Externally Forced Climate Response. *J. Climate*, **28**, 8184-8202, https://doi.org/10.1175/JCLI-D-15-0069.1.

Fujiwara, M., P. Martineau, and J. S. Wright, 2020: Surface temperature response to the major volcanic eruptions in multiple reanalysis data sets. *Atmos. Chem. Phys.*, **20**, 345–374, https://doi.org/10.5194/acp-20-345-2020.

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

Good, S. A., M. J. Martin, and N. A. Rayner, 2013: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res.: Oceans*, **118**, 6704-6716, https://doi.org/10.1002/2013JC009067.

Hakim, G., J. Emile-Geay, E. Steig, D. Noone, D. Anderson, R. Tardif, N. Steiger, and W. Perkins, 2016: The Last Millennium Climate Reanalysis Project: Framework and First Results. *J. Geophys. Res.*, **121**, https://doi.org/10.1002/2016JD024751.

Hasselmann, K., 1976: Stochastic climate models Part 1. Theory. *Tellus*. **28**, 473-485, https://doi.org/10.1111/j.2153-3490.1976.tb00696.x.

Hawkins, E., and R. Sutton, 2009: Decadal predictability of the Atlantic Ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling. *J. Climate*, **22**, 3960–3978, https://doi.org/10.1175/2009BAMS2607.1.

Hazeleger, W., and Coauthors, 2013: Multi-year climate predictions using two initialization strategies. *Geophys. Res. Lett.*, **40**, 1794–1798. https://doi.org/10.1002/grl.50355.

Kohl, A., 2020: Evaluating the GECCO3 1948-2018 ocean synthesis – a configuration for initializing the MPI-ESM climate model. *Q. J. R. Meteorol. Soc.*, **146**, 2250-2273, https://doi.org/10.1002/qj.3790.

Kushnir, Y., 1994: Interdecadal variations in the North Atlantic sea surface temperature and associated atmospheric conditions. *J. Climate*, **7**, 141–157, https://doi.org/10.1175/1520-0442(1994)007<0141:IVINAS>2.0.CO;2.

Liu, W. and S. Xie, 2018: An Ocean View of the Global Surface Warming Hiatus. *Oceanography*, **31**, 72-79, https://doi.org/10.5670/oceanog.2018.217.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 131–140, https://doi.org/10.1175/1520-0469(1963)020%3C0130:DNF%3E2.0.CO;2.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. *Bull. Amer. Meteor. Soc.*, **78**, 1069-1080, https://doi.org/10.1175/1520-0477(1997)078%3C1069:APICOW%3E2.0.CO;2.

Meehl, G. A., and Coauthors, 2009: Decadal prediction: Can it be skillful? *Bull. Amer. Meteor. Soc.*, **90**, 1467–1485, https://doi.org/10.1175/2009BAMS2778.1.

Meehl, G. A, L. Goddard, G. Boer, and R. Burgman, 2014: Decadal Climate Prediction: An Update from the Trenches. *Bull. Amer. Meteor. Soc.*, **95**, 243-267, https://doi.org/10.1175/BAMS-D-12-00241.1.

Newman, M., 2013: An Empirical Benchmark for Decadal Forecasts of Global Surface Temperature Anomalies. *J. Climate*, **26**(14), 5260–5269, https://doi.org/10.1175/JCLI-D-12-00590.1.

Pages2k Consortium, 2017: Data descriptor: A global multiproxy database for temperature reconstructions of the Common Era. *Nature: Scientific Data*.

Palmer, M. D., C. D. Roberts, M. Balmaseda, et al., 2017: Ocean heat content variability and change in an ensemble of ocean reanalyses. *Clim Dyn*, **49,** 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**, 2165-2185, https://doi.org/10.1175/1520-0493(1989)117%3C2165:RFAFUP%3E2.0.CO;2.

Penland, C. and L. Matrosova, 1994: A Balance Condition for Stochastic Numerical Models with Application to the El Nino-Southern Oscillation. *J. Climate*, **7**, 1352-1372, https://doi.org/10.1175/1520-0442(1994)007%3C1352:ABCFSN%3E2.0.CO;2.

Penland, C., and P. D. Sardeshmukh, 1995: The Optimal Growth of Tropical Sea Surface Temperature Anomalies. *J. Climate*, **8**(8), 1999–2024, https://doi.org/10.1175/1520-0442(1995)008¡1999:TOGOTS¿2.0.CO;2.

Penny, S. G., S. Akella, M. A. Balmaseda, P. Browne, J. A. Carton, M. Chevallier, F. Counillon, C. Domingues, S. Frolov, P. Heimbach, P. Hogan, I. Hoteit, D. Iovino, P. Laloyaux, M. J. Martin, S. Masina, A. M. Moore, P. de Rosnay, D. Schepers, B. M. Sloyan, A. Storto, A. Subramanian, S. Nam, F. Vitart, C. Yang, Y. Fujii, H. Zuo, T. O'Kane, P. Sandery, T. Moore, and C. C. Chapman, 2019: Observational Needs for Improving Ocean and

Coupled Reanalysis, S2S Prediction, and Decadal Prediction. *Front. Mar. Sci.*, **6**(391), https://doi.org/10.3389/fmars.2019.00391.

Perkins, W. A., and G. Hakim, 2017: Reconstructing paleoclimate fields using online data assimilation with a linear inverse model. *Clim. Past*, **13**, 421-436, https://doi.org/10.5194/cp-13-421-2017.

Perkins, W. A., and G. Hakim, 2020: Linear Inverse Modeling for Coupled Atmosphere–Ocean Ensemble Climate Prediction. *Journal of Advances in Model Earth Systems*, **12**(1), https://doi.org/10.1029/2019MS001778.

Perkins, W. A., and G. Hakim, 2021: Coupled Atmosphere–Ocean Reconstruction of the Last Millennium Using Online Data Assimilation. *Paleoceanography and Paleoclimatology*, in review.

Power, S., T. Casey, C. Folland, A. Colman, and V. Mehta, 1999: Inter-decadal modulation of the impact of ENSO on Australia. *Clim. Dynamics*, **15**, 319–324, https://doi.org/10.1007/s003820050284.

Salvi, K., G. Villarini, G. A. Vecchi, 2017: High resolution decadal precipitation predictions over the continental United States for impacts assessment. *Journal of Hydrology*, **553**, 559-573, https://doi.org/10.1016/j.jhydrol.2017.07.043.

Schurer, A. P., G. C. Hegerl, M. E. Mann, S. F. B. Tett, and S. J. Phipps, 2013: Separating Forced from Chaotic Climate Variability over the Past Millennium. *J. Climate*, **26**, 6954-6973, https://doi.org/10.1175/JCLI-D-12-00826.1.

Smith, D. M., R. Eade, R., A. A. Scaife, et al., 2019: Robust skill of decadal climate predictions. *npj Clim Atmos Sci*, **2**, https://doi.org/10.1038/s41612-019-0071-y.

Tardif, R., G. J. Hakim, W. A. Perkins, K. A. Horlick, M. P. Erb, J. Emile-Geay, D. M. Anderson, E. J. Steig, and D. Noone, 2019: Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling. *Clim. Past*, **15**, 1251-1273, https://doi.org/10.5194/cp-15-1251-2019.

Trenberth, K. E. and J. W. Hurrell, 1994: Decadal atmosphere-ocean variations in the Pacific. *Clim. Dynamics*, **9**, 303–319, https://doi.org/10.1007/BF00204745.

Trenberth, K. E., and Coauthors, 2007: Observations: Surface and atmospheric climate change. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 235–336.

Trouet, V. and A. H. Taylor, 2009: Multi-century variability in the Pacific North American circulation pattern reconstructed from tree rings. *Clim. Dynamics*, **35**, 953-963, https://doi.org/ 10.1007/s00382-009-0605-9.

Turner, J., T. Phillips, J. S. Hosking, G. J. Marshall, and A. Orr, 2012: The Amundsen Sea Low. *International Journal of Climatology*, **33**(7), 1818-1829, https://doi.org/10.1002/joc.3558.

Yu, B. and F. W. Zwiers, 2007: The impact of combined ENSO and PDO on the PNA climate: a 1,000-year climate modeling study. *Clim. Dynamics*, **29**, 837-851, https://doi.org/ 10.1007/s00382-007-0267-4.