Impact of Systematic GCM Errors on Prediction Skill as Estimated by Linear Inverse Modeling

INGO RICHTER,^a PING CHANG,^b AND XUE LIU^b

^a *Application Laboratory, Research Institute for Value-Added-Information Generation, Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan;* ^b *Department of Oceanography and International Laboratory for High-Resolution Earth System Prediction (iHESP), Texas A&M University, College Station, Texas*

(Manuscript received 24 March 2020, in final form 2 August 2020)

Statistical prediction of tropical sea surface temperatures (SSTs) is performed using linear inverse models (LIMs) that are constructed from both observations and general circulation model (GCM) output of SST. The goals are to establish a baseline for tropical SST predictions, to examine the extent to which the skill of a GCM-derived LIM is indicative of that GCM's skill in forecast mode, and to examine the linkages between mean state bias and prediction skill. The observation-derived LIM is more skillful than a simple persistence forecasts in most regions. Its skill also compares well with some GCM forecasts except in the equatorial Pacific, where the GCMs are superior. The observation-derived LIM is matched or even outperformed by the GCM-derived LIMs, which may be related to the longer data record available for GCMs. The GCM-derived LIMs provide a fairly good measure for the skill achieved by their parent GCMs in forecast mode. In some cases, the skill of the LIM is actually superior to that of its parent GCM, indicating that the GCM predictions may suffer from initialization problems. A weak-to-moderate relation exists between model mean state error and prediction skill in some regions. An example is the eastern equatorial Atlantic, where an erroneously deep thermocline reduces SST variability, which in turn affects prediction skill. Another example is the equatorial Pacific, where skill appears to be linked to cold SST biases in the western tropical Pacific, which may reduce the strength of air–sea coupling.

KEYWORDS: Atlantic Ocean; Pacific Ocean; Statistics; Climate prediction; Model errors

1. Introduction

The tropical oceans are subject to intense year-to-year variability that exerts a strong influence on the surrounding continents, with El Niño–Southern Oscillation (ENSO) in the Pacific being the most prominent example. Predicting such oceanic variability patterns and their remote impacts has been a priority for many research centers. Such predictions usually rely on coupled general circulation models (GCMs), which are complex numerical models consisting, at minimum, of atmospheric and oceanic components. The skill of GCM-based seasonal predictions has increased considerably over the past few decades, particularly in the equatorial Pacific, where predictions at 6-month lead time tend to be highly successful (Barnston et al. 2019), with some studies suggesting that even 2-yr predictions with an anomaly correlation coefficient (ACC) above 0.5 are possible (Luo et al. 2008). There are cases, on the other hand, when GCM predictions do not live up to their high expectations, with the poor predictions of the 2014/15 El Niño event being one example (McPhaden 2015; Imada et al. 2016; Zhu et al. 2016; Chiodi and Harrison 2017). While failed predictions of the equatorial Pacific are the exception rather than

the rule, this cannot be said of other tropical regions. A region that is notoriously difficult to predict is the equatorial Atlantic (e.g., Penland and Matrosova 1998; Chang et al. 2006a; Richter et al. 2018). Variability there is marked by sea surface temperature (SST) changes in the eastern equatorial Atlantic that starts in boreal spring and terminates in early fall [e.g., Merle 1980; Zebiak 1993; Carton and Huang 1994; see reviews by Chang et al. (2006b), Lübbecke et al. (2018), and Richter and Tokinaga (2020)]. Due to its apparent similarity with El Niño, this phenomenon has been dubbed “Atlantic Niño” (Merle 1980), although in recognition of the fundamental differences that have come to light (e.g., Richter et al. 2013) we refer to this variability pattern as the Atlantic zonal mode (AZM).

Prediction of the AZM remains a challenge with little improvement over the years (Stockdale et al. 2006; Richter et al. 2018). At the same time, the equatorial Atlantic is subject to severe mean state biases in SST, surface winds, and precipitation (Richter et al. 2014a), which also have seen little improvement over the last few decades (Davey et al. 2002; Richter and Xie 2008; Richter et al. 2014a). Given these facts, it may be natural to link the current low prediction skill for the AZM to the tropical Atlantic GCM biases but the few studies that have attempted to address this issue point to a weak link between the two (Richter et al. 2018; S. Koseki et al. 2020, unpublished manuscript). Likewise, most studies of prediction skill in the tropical Pacific and other tropical regions have only found a weak-to-moderate relation with mean state biases, particularly for SST [Gualdi et al. 2005; Manganello and Huang

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JCLI-D-20-0209.s1>.

Corresponding author: Ingo Richter, richter@jamstec.go.jp

2009; Del Sole and Shukla 2010; Lee et al. 2010; Magnusson et al. 2013; Ding et al. 2015; see Richter et al. (2018) for a discussion of these studies]. Recently, Ding et al. (2020) have used analog forecasts based on GCM output to examine the relation between bias and prediction skill in the equatorial Pacific. They find a strong relation for precipitation but a relatively weak one for SST. This is consistent with Richter et al. (2018), who found a stronger impact on precipitation than on SST. Li et al. (2020) use a different statistical technique to make forecasts based on output from two versions of a GCM with differing resolutions. They find that the high-resolution version of their GCM has smaller biases and that the statistical model built from it has higher prediction skill in the tropical Atlantic.

Given the uncertain role of model biases, it remains an open question whether prediction skill for the AZM, and the tropical Atlantic in general, is held back by model deficiencies or by fundamental predictability limits. These limits are essentially set by the amount of internal variability, that is, variability unpredictable at the time scales of interest, such as atmospheric weather events. Past studies have taken a look at the role of internal variability in the tropical Atlantic. Chang et al. (2001) examine a variability pattern known as the Atlantic meridional mode (AMM; Servain et al. 1999), in which subtropical SST anomalies of opposite sign straddle the equator. Chang et al. (1997) suggest that the AMM relies on an air–sea coupling mechanism called the wind–evaporation–SST (WES) feedback (Xie and Philander 1994). The study by Chang et al. (2001) suggests that coupled air–sea feedbacks are not strong enough for a self-sustained oscillation and that stochastic forcing related to the North Atlantic Oscillation (NAO; e.g., Hurrell 2001) is crucial for the development of the AMM [see Amaya et al. (2017) for more discussion on this topic]. Likewise, Zebiak (1993) found that coupled air–sea feedbacks in the form of the Bjerknes feedback (Bjerknes 1969) are not sufficient to maintain the AZM, suggesting that stochastic forcing plays an important role. More recently, Richter et al. (2014b) and Richter and Doi (2019) have provided additional evidence that stochastic forcing plays an important role in the evolution of the AZM.

The present study aims to further investigate the link between mean state biases and prediction skill in GCMs. Since the mean state (including prediction drift) is routinely removed from seasonal predictions it does not figure directly into prediction errors. Rather, mean state biases must affect prediction skill indirectly through variability errors, which include errors in stochastic noise forcing and errors in coupled feedbacks and variability patterns (e.g., the strength of the Bjerknes feedback and the geographical location of ENSO variability). Our focus is on the latter; that is, we would like to examine the linkage from mean state biases to errors in coupled feedbacks and variability patterns, and further on to prediction skill.

One way of examining the link between mean state biases is empirically by comparing mean state biases and prediction skill across a large model ensemble. While a few multimodel archives of GCM reforecasts are available [e.g., the Climate-system Historical Forecast Project (CHFP; Kirtman and Pirani 2009) and the North American Multi-Model Ensemble (NMME; Kirtman et al. 2014)], there are typically no corresponding free-running

TABLE 1. Summary of all the types of all the linear inverse models (LIMs) used in this study.

LIM name	Description
O-LIM	LIM built from SST observations in the global tropics
O-LIM_TATL	Like O-LIM, but using tropical Atlantic SST only
M-LIM	LIM built from model SST output in the global tropics
M-LIM_TATL	Like M-LIM, but using tropical Atlantic SST only
M-LIM_50yr	Like M-LIM, but uses only 50 years of model output for training

control simulations to assess the underlying mean state biases. We therefore choose a different approach that relies on the phase 5 of the Coupled Model Intercomparison Project (CMIP5) archive of preindustrial control (piControl) simulations. Using SSTs from these GCMs, we train a relatively simple statistical model to predict observed SSTs. The statistical technique in question is called linear inverse modeling (LIM; e.g., Penland and Matrosova 1998, hereafter PM98) and has been applied by numerous authors to predict variability patterns such as the AMM (Vimont 2012), ENSO (Tang 1995), the tropical Atlantic (PM98; Li et al. 2020), the Pacific decadal oscillation (Alexander et al. 2008), or annular modes (Sheshadri and Plumb 2017). This technique will be described in section 2.

Typically, LIMs are trained on a subset of observations (also referred to as O-LIM hereafter) and verified using the remaining data. This is also our starting point but, in addition, we construct one LIM for each model in the CMIP5 database, based on SST from the free-running piControl simulations only (also referred to as M-LIM hereafter; see Table 1 for a summary of all LIM types). These M-LIMs must reflect, to some extent, the errors in coupled feedbacks and the variability patterns of the parent GCM, such as an erroneous westward shift of ENSO-induced SST variability. Thus, if there is an influence of coupled feedbacks and variability pattern errors on prediction skill, it should be reflected in the performance of the M-LIM. At the same time, the free-running piControl simulations the LIMs are based on allow a straightforward estimation of the mean state biases.

The above approach can elucidate the link between mean state biases in the parent GCM and the skill of the associated M-LIM. The remaining question is to what extent the skill of each M-LIM is representative of that of the parent GCM when run in forecast mode. Since a few GCMs in the CMIP5 archive have corresponding seasonal forecasts available in the CHFP archive, we examine this question in section 5.

The goals of the present study can thus be summarized as follows:

- 1) Evaluate the performance of LIMs constructed from observations and GCMs, and establish a baseline to compare dynamical prediction systems against.
- 2) Examine the extent to which the skill of M-LIMs is representative of that of the parent GCM when run in forecast mode.
- 3) Examine the link between M-LIM performance and the mean state biases of the parent GCM.

TABLE 2. List of all CMIP5 coupled GCMs used to construct M-LIMs. The record length (years) of each dataset is noted in the rightmost column.

Model	Modeling center, country	No. of years
ACCESS1.0	CSIRO and BoM, Australia	250
ACCESS1.3	CSIRO and BoM, Australia	500
BCC-CSM1.1	Beijing Climate Center, China	500
BCC-CSM1.1-m	Beijing Climate Center, China	400
BNU-ESM	Beijing Normal University, China	559
CanESM2	Canadian Centre for Climate Modeling and Analysis, Canada	996
CCSM4	National Center for Atmospheric Research, United States	501
CESM1-CAM5	National Center for Atmospheric Research, United States	319
CMCC-CM	Centro Euro-Mediterraneo Cambiamenti Climatici, Italy	330
CNRM-CM5	Centre National de Recherches Météorologiques, France	850
CSIRO-Mk3.6.0	Commonwealth Scientific and Industrial Research Organization, Australia	500
FGOALS-g2	Institute of Atmospheric Physics, China	700
FGOALS-s2	Institute of Atmospheric Physics, China	501
FIO-ESM	First Institute of Oceanography, China	200
GFDL CM3	Geophysical Fluid Dynamics Laboratory, United States	500
GFDL-ESM2G	Geophysical Fluid Dynamics Laboratory, United States	500
GFDL-ESM2M	Geophysical Fluid Dynamics Laboratory, United States	500
GISS-E2-H	NASA Goddard Institute for Space Studies, United States	240
GISS-E2-R	NASA Goddard Institute for Space Studies, United States	550
HadGEM2-CC	Met Office Hadley Centre for Climate Change, United Kingdom	240
HadGEM2-ES	Met Office Hadley Centre for Climate Change, United Kingdom	575
INM-CM4	Russian Institute for Numerical Mathematics, Russia	500
IPSL-CM5A-LR	Institut Pierre Simon Laplace, France	1000
IPSL-CM5A-MR	Institut Pierre Simon Laplace, France	300
IPSL-CM5B-LR	Institut Pierre Simon Laplace, France	300
MIROC4h	University of Tokyo and NIES and JAMSTEC, Japan	100
MIROC5	University of Tokyo and NIES and JAMSTEC, Japan	670
MIROC-ESM	University of Tokyo and NIES and JAMSTEC, Japan	531

TABLE 2. (Continued)

Model	Modeling center, country	No. of years
MIROC-ESM-CHEM	University of Tokyo and NIES and JAMSTEC, Japan	255
MPI-ESM-LR	Max Planck Institute for Meteorology, Germany	1000
MPI-ESM-MR	Max Planck Institute for Meteorology, Germany	1000
MPI-ESM-P	Max Planck Institute for Meteorology, Germany	1156
MRI-CGCM3	Meteorological Research Institute, Japan	500
NorESM1-M	Bjerknes Centre for Climate Research, Norway	501
NorESM1-ME	Bjerknes Centre for Climate Research, Norway	252

After introducing the data and methods in [section 2](#), we will address goals 1 and 2 in [section 3](#). Goal 3 will be addressed in [section 4](#). A summary and conclusions are given in [section 5](#).

2. Data description and methods

We use a set of 35 models from the piControl simulation of the CMIP5 archive for our analysis (see [Table 2](#) for a list of the models). These models span a wide range of mean state biases and variability patterns ([Richter et al. 2014a](#)), which is essential to our study. Monthly mean SST anomalies were used to construct the LIMs. Since the radiative forcing is kept steady in piControl, simulations typically do not display pronounced long-term trends in SST, but we nevertheless performed linear detrending prior to analysis.

Our reference data are SSTs from the NCEP–NCAR reanalysis ([Kalnay et al. 1996](#)) for the period 1948–2017, with the climatological annual cycle and linear trend removed. These data serve as reference for assessing the skill of LIMs constructed from both the reference data itself and from model output. Both the reanalysis and GCM output are interpolated to the same 2° grid before analysis.

To examine the extent to which the skill of M-LIMs is representative to that of the parent model in forecast mode, we examine a set of four models that have corresponding hindcasts available in the CHFP archive. These models are CanESM2, MIROC5, MPI-ESM-LR, and MRI-CGCM3. The CHFP archive comprises hindcasts for the period 1979–2012, although the exact hindcast period varies by model.

The LIM technique has been extensively described in the literature (e.g., [Penland 1989](#); [Alexander et al. 2008](#)). Here we will only give a brief description and list the settings specific to our analysis. The basis of the LIM approach is the assumption that the evolution of a dynamical system can be approximated as the sum of a linear term and a residual

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}(t) + \xi(t), \quad (1)$$

where $\mathbf{x}$ is the state vector of the system, $\mathbf{A}$ is a matrix representing the deterministic linear dynamics of the system, and ξ is the residual, which includes nonlinear terms and other processes. The approach of the LIM assumes that ξ can be parameterized as a white-noise process with spatial structure. The state vector $\mathbf{x}$ describes the state of the relevant system variables at all locations and is a function of time. In this study, we primarily examine SST-based LIMs, in which case $\mathbf{x}$ contains only SST. In the online supplemental material, we also examine LIMs that are derived from combined SST and sea surface height (SSH). In that case, $\mathbf{x}$ contains both SST and SSH.

Equation (1) can be integrated to yield

$$\mathbf{x}(t + \tau_0) = \exp(\mathbf{A}\tau_0)\mathbf{x}(t) + \varepsilon. \quad (2)$$

The basic idea behind the LIM approach is to construct the linear operator $\mathbf{B} = \exp(\mathbf{A}\tau_0)$ from data. This can be achieved through the following covariance equation:

$$\langle \mathbf{x}(t + \tau_0)\mathbf{x}^T(t) \rangle = \mathbf{B}\langle \mathbf{x}(t)\mathbf{x}^T(t) \rangle + \langle \varepsilon\mathbf{x}^T(t) \rangle,$$

where $\langle \mathbf{x}(t)\mathbf{x}^T(t) \rangle$ is the covariance matrix of $\mathbf{x}(t)$, $\mathbf{C}$, and $\langle \mathbf{x}(t + \tau_0)\mathbf{x}^T(t) \rangle$ is the lag-covariance matrix at τ_0 , $\mathbf{C}\tau_0$. Assuming that the white-noise process ε is uncorrelated to the state vector $\mathbf{x}(t)$, this simplifies to

$$\mathbf{C}\tau_0 = \mathbf{B}\mathbf{C}. \quad (3)$$

These covariance matrices are readily calculated from the SST data. If $\mathbf{C}$ is well defined then Eq. (3) can be inverted to yield the linear operator $\exp(\mathbf{A}\tau_0) = \mathbf{B} = \mathbf{C}\tau_0\mathbf{C}^{-1}$.

If only predictions at lead time τ_0 were desired one could simply use $\mathbf{x}(t + \tau_0) = \mathbf{B}\mathbf{x}(t)$. The prediction at an arbitrary lead time τ is given by

$$\mathbf{x}(t + \tau) = \mathbf{B}^{\tau/\tau_0}\mathbf{x}(t). \quad (4)$$

While the linear operator could be estimated directly from SST data (in this case the dimension of $\mathbf{C}$ will be determined by the product of longitudinal and latitudinal grid points), inverting such a high-dimensional matrix is a nontrivial task. Furthermore, the existence of $\mathbf{C}^{-1}$ is not guaranteed in this case. Thus, the estimation of $\mathbf{B}$ is typically performed in reduced state space using principal component (PC) analysis. In this approach, the state vector $\mathbf{x}(t)$ is first expressed by the sum of empirical orthogonal functions (EOFs) times the corresponding PCs. Then, the summation is truncated at a certain leading number N of EOFs or PCs that explain a high percentage of variance. The truncation number N , which is typically a small fraction of the dimensionality of the physical space, is one of the adjustable parameters of the LIM technique. Here we use the minimum number of PCs that explain 80% of the SST variance, so that N is model dependent. The reason for choosing this criterion is to facilitate comparison of the performance of LIMs calculated from different datasets. Specifically, we want to ensure that performance differences are not due to the underlying PCs explaining disparate amounts of variance. The orthogonality of the PCs diagonalizes $\mathbf{C}$, greatly simplifying the calculation of $\mathbf{C}^{-1}$.

When computing the covariance matrices, one has to set a value for τ_0 . If the dynamics are linear, $\mathbf{A}$ should be independent

of the choice of τ_0 . This is checked with the so-called “tau test” (e.g., Penland and Sardeshmukh 1995). We have performed this test on the LIMs used in this study (see the online supplemental material) and found that they generally perform reasonably well, with $\mathbf{A}$ roughly constant for lags ranging from 3 to 9 months.

In practice, one also has to consider aliasing and the so-called Nyquist problem that arises when the lag approaches half the period of a mode of variability contained in the data [see Penland (2019) for a detailed discussion]. In the present study we set τ_0 as 7 months, which is a value often chosen for ENSO prediction. The AZM, however, which is also of interest here, has a shorter period and thus one may wonder if a smaller value for τ_0 would yield better prediction skill by avoiding potential Nyquist problems. We tested this but could not find substantial performance differences.

A technique that is closely related to LIM is principal oscillation patterns (POPs; von Storch et al. 1995). The POPs are defined as the eigenvectors of the linear operator $\mathbf{B}$. A practical reason for calculating the eigenvector decomposition of $\mathbf{B}$ is that it facilitates the exponentiation $\mathbf{B}^{\tau/\tau_0}$. Beyond that, however, the POPs yield information about the (presumably oscillatory) variability patterns of the system, with the generally complex eigenvalues of $\mathbf{B}$ indicating the damping rate and frequency of the individual modes. If the eigenvalues are real, this indicates a purely decaying pattern without any oscillatory component.

Since POPs are eigenvectors they are only determined to a complex scalar; that is, if $\mathbf{p}_i$ is an eigenvector to eigenvalue λ_i then $c\mathbf{p}_i$ is also an eigenvector to the same eigenvalue, where c is an arbitrary complex number. Thus, when comparing POPs obtained from different datasets (i.e., models in our case), normalization becomes an issue. Here we follow the procedure described by Gallagher et al. (1991), which normalizes POPs using the following two requirements: 1) $\mathbf{p}_i\mathbf{p}_i = N$, where N is the number of PCs, or, more generally the dimensionality of the space; and 2) $\text{Re}(\mathbf{p}_i)\text{Im}(\mathbf{p}_i) = 0$; that is, the real and imaginary components of the eigenvector are orthogonal.

3. LIM performance

a. LIM constructed from NCEP reanalysis

To establish a baseline, we first evaluate the O-LIM. This is done using a jackknifing procedure in which the LIM is constructed from 56 years of data and validated on the remaining 14 years. The validation period is shifted in 14-yr increments, resulting in five sets of skill metrics, and the average over these five sets is analyzed here. For an assessment of the O-LIM performance we calculate the anomaly correlation coefficient (ACC) between the O-LIM prediction at lead time 6 months and the reference data, without stratifying by season (Fig. 1). The spatial pattern of the ACC shows the highest values in the central tropical Pacific (Fig. 1a), consistent with the generally high predictability of ENSO. The ACC of the O-LIM is higher than that of persistence in most regions, with the South Atlantic and southeastern Indian Ocean being notable exceptions (Fig. 1b). The unstratified root-mean-square error (RMSE) shows the highest values along the equatorial Pacific (Fig. 1c), indicating that the O-LIM predictions have large errors there. This is to be

FIG. 1. Skill assessment of the LIM constructed from observed SST from NCEP–NCAR reanalysis (O-LIM). All metrics are calculated for 6-month lead time and not stratified by initialization month. Cross-validation is used to examine the forecast skill for the period 1948–2017. The individual panels show the (a) anomaly correlation coefficient (ACC), (b) ACC difference of the O-LIM and persistence forecasts, (c) root-mean-square error (RMSE; K), and (d) RMSE difference of the O-LIM and persistence forecasts.

expected because variability is also highest in the equatorial Pacific. When compared to the RMSE of the persistence forecast (Fig. 1d), the O-LIM shows superior performance everywhere, including the equatorial Pacific. Overall, the performance of the O-LIM compares reasonably well with that of other studies (Tang 1995; PM98; Newman et al. 2011; Newman and Sardeshmukh 2017) but tends to perform a little worse, particularly compared to the last two studies in the list. In the equatorial Pacific and Atlantic

the performance of the O-LIM may be affected by the relatively low variability in these regions over the last two decades (Tokinaga and Xie 2011; McGregor et al. 2018; Richter and Tokinaga 2020), which was not part of the validation period in some previous studies.

Adding SSH and atmospheric winds to LIMs has been shown to boost their performance in terms of SST predictions (e.g., Newman et al. 2011; Capotondi and Sardeshmukh 2015).

We tested the impact of including SSH but found only minor performance improvements in the tropical Pacific for lead times up to 9 months, while performance in the Atlantic tended to deteriorate (see the supplemental material). Finally, the performance of LIMs can also be sensitive to the number of PCs retained. Since our main goal is not to construct an optimized forecast model, we do not further analyze the performance differences.

We examine in more detail the skill for individual regions of interest. The ATL3 region (20°W–0, 3°S–3°N) is a commonly used indicator for equatorial variability and the AZM. The ACC of SST anomalies area-averaged over the ATL3 (Fig. 2a) shows that skill is slightly below persistence at short lead times (1–2 months), and above at longer lead times. For the northern tropical Atlantic (NTA; ocean points in 40°–10°W, 10°–20°N), ACC (Fig. 2b) is slightly below (above) persistence at lead times 1–4 (5–12), whereas for the southern tropical Atlantic (STA; ocean points in 20°W–20°E, 25°S–5°S) the ACC is below persistence at most lead times (Fig. 2c). The NTA and STA indices are important indicators for the state of the AMM (Chang et al. 1997; Amaya et al. 2017), a variability pattern that is associated with changes in ITCZ latitude and tropical Atlantic surface winds, among other things. The ACC for the NTA and STA indicates that our O-LIM has no useful skill in their prediction. For RMSE, on the other hand, the O-LIM outperforms persistence for all three Atlantic indices (Figs. 2e–g). The performance of our O-LIM in the tropical Atlantic is lower than the results reported by PM98. One reason for this may be the relatively low signal-to-noise ratio (SNR) in the tropical Atlantic during the last few decades (Prigent et al. 2020), which are part of our validation period but not theirs. Furthermore, PM98 use the EOF-reconstructed SST for verification, whereas we use the actual observations. Last, it is possible that they did not perform linear detrending, as suggested by the high persistence of their predictands: for the ATL3, for example, they obtain 0.5 at 6-month lead, compared to our 0.2 [see also Ding et al. (2019) for the influence of linear trends on skill]. A recent study by Li et al. (2020) used a LIM based on tropical Atlantic SST only and found that it was clearly outperformed by persistence in the equatorial Atlantic.

The performance of the O-LIM in the Niño-3.4 region (170°–120°W, 5°S–5°N) clearly rises above that of the persistence forecast, particularly at longer lead times (Figs. 2d,h). Throughout the forecast period, the ACC of the O-LIM remains above 0.5, which is often considered a threshold for the usefulness of predictions.

The LIM clearly performs better in the tropical Pacific than in the tropical Atlantic, which is partly explained by the higher predictability of ENSO vis-à-vis the Atlantic variability patterns. In addition to that, however, the truncation of PCs leads to an uneven spatial distribution of explained observed variance; that is, the 17 PCs that are chosen for our reduced state space explain more variance in the Pacific than in the Atlantic, due to the variability being concentrated in the Pacific. We have therefore recalculated the LIM using SST in the tropical Atlantic only (30°S–30°N over the width of the basin) and name this O-LIM_TATL. This LIM (blue lines in Fig. 2) performs better than the global-tropics LIM (green

lines in Fig. 2), particularly in the ATL3 and STA regions, where variability is, to some extent, independent of ENSO (Chang et al. 2006a; Lübbecke and McPhaden 2012; Tokinaga et al. 2019). Rather than optimizing performance for any particular basin, our goal in this section is to examine the general performance of the LIM technique. We will therefore focus on LIMs constructed from global tropical SSTs in this section (simply denoted as “LIM”). In section 4, where we examine the link between bias and skill, we will also briefly discuss results from M-LIM_TATL.

b. LIMs constructed from GCMs

Having confirmed that our O-LIM performs reasonably well, we evaluate the ability of each M-LIM to predict the variability of its parent GCM. This is done using a jackknifing procedure similar to the one in section 3a but with a 50-yr validation period. To construct a metric representative of the M-LIM ensemble, we form the ensemble average over each metric (ACC or RMSE). For ACC, this is done using the Fisher z -transformation.

The global pattern of ACC at lead time 6 (Fig. 3a) shows similarity with that of the O-LIM (Fig. 1a), with the highest values in the central tropical Pacific. One difference is that the M-LIMs generally have higher skill than their observation-based counterpart (evident when comparing Figs. 1a and 3a). The global distribution of RMSE (Figs. 3c,d) shows that the M-LIM ensemble outperforms persistence almost everywhere, with particularly large differences in the equatorial Pacific.

Focusing on our four regions of interest (ATL3, NTA, STA, and Niño-3.4), we find that M-LIMs are able to predict model data better than persistence at most lead times (Fig. 4), especially those longer than 3 months. Generally, the metrics are quite comparable with those of the O-LIM, except for the STA, where the M-LIMs are better at predicting themselves than the O-LIM.

Overall, the results suggest that the M-LIMs are quite successful at predicting the variability of their parent GCM. Including SSH into the state vector leads to slightly improved performance in the tropical Pacific and slight deterioration in the tropical Atlantic (see the supplemental material).

c. GCM-derived LIMs predicting observations

While LIMs are typically trained and validated on observational data it is straightforward to train a LIM on model output and use it to predict observations. One would expect that, due to variability errors in GCMs, a LIM trained on GCM output will perform worse than a LIM trained on observations, and the extent to which systematic GCM errors affect the ability of the corresponding LIM to predict observations will be the focus of section 4. Here we only validate the technique and compare the ensemble mean of M-LIM skills to the skill of the O-LIM.

Since variability in the piControl simulations is completely independent from that of the observations we use the entire integration period to train the M-LIMs. To obtain an exact comparison between M-LIM and O-LIM performance, we predict the same moving 14-yr periods as in section 3a. Applying the M-LIMs to the observational data requires projecting the

FIG. 2. Skill of O-LIM (green line), O-LIM_TATL (O-LIM built from tropical Atlantic SST only; blue line), and the persistence forecast (black line) as a function of lead time for selected indices. All metrics are unstratified by season. The individual panels show (a) ATL3 (20°W – 0° , 3°S – 3°N) ACC, (b) NTA (40° – 10°W , 10° – 20°N) ACC, (c) STA (20°W – 20°E , 25° – 5°S) ACC, (d) Niño-3.4 (170° – 120°W , 5°S – 5°N) ACC, (e) ATL3 RMSE, (f) NTA RMSE, (g) STA RMSE, and (h) Niño-3.4 RMSE.

FIG. 3. As in Fig. 1, but for the ensemble mean of LIMs calculated from CMIP5 piControl simulations (M-LIMs) using cross-validation with a 50-yr validation window. The ensemble mean of the ACC was calculated using a Fisher z -transformation.

observed state onto the model PCs that were used to construct the individual M-LIMs.

The global map of ACC at lead time 6 suggests that the skill of the M-LIMs is comparable to that of the O-LIM in most regions (Figs. 5a,b) and exceeds the O-LIM in the southern tropical Atlantic. RMSE performance of the two predictions is also relatively similar (Figs. 5c,d), although the M-LIMs perform better in the central equatorial Pacific.

For the ATL3 index, the average ACC of the M-LIMs is somewhat lower than that of the O-LIM. For Niño-3.4 (Fig. 6d)

the M-LIM is comparable to the O-LIM at lead times 1–6 but falls behind at longer lead times. For the NTA and STA (Figs. 6b,c), the M-LIM average outperforms the O-LIM at most lead times. The RMSE tends to be similar between the M-LIM average and the O-LIM, though the latter tends to perform slightly better (Figs. 6e–h). Since performance varies across the M-LIM ensemble, there are some M-LIMs that consistently outperform the O-LIM in all metrics (not shown).

The results suggest that, somewhat surprisingly, LIMs constructed purely from model output can outperform a LIM

FIG. 4. As in Fig. 2, but for the skill of the M-LIMs in Fig. 3.

FIG. 5. As in Fig. 3, but for the M-LIM predicting observed SST for the period 1948–2017. No cross-validation is used here but rather the M-LIMs are trained on 100% of the available data from the piControl simulations.

constructed from observations. We have tested whether this is related to the longer training period of the M-LIMs. For this purpose, we build each M-LIM from a moving 50-yr training period (M-LIM_50yr; dashed green line in Fig. 6). This consistently decreases performance, suggesting that the length of the training period is indeed an important factor for the success of M-LIMs. A similar dependence of performance on training period was found by Ding et al. (2018) in their analog model. While the drop in performance is consistent, it is not large

enough to make the ACC of M-LIM_50yr inferior to that of the O-LIM in the STA.

In summary, the results suggest that the variability patterns generated by some of the GCMs are sufficiently realistic to serve as a surrogate for observations, at least in the context of this relatively simple statistical technique. Other recent studies have also reported successful prediction of observations with statistical models trained on GCM output (Ding et al. 2018; Ham et al. 2019).

FIG. 6. As in Fig. 4, but for the skill of the M-LIMs in Fig. 5 for predicting observed indices for the period 1948–2017 (green lines). The dashed green lines show the skill of M-LIMs trained on 50 years only. The skill of the O-LIM is repeated to facilitate comparison (gray line).

TABLE 3. List of the CHFP seasonal forecast models used in this study. The second column shows the corresponding models in the CMIP5 archive. The Canadian forecast model (CCCma-CanCM4) differs from its CMIP5 counterpart (CanESM2) because it does not have atmospheric and oceanic carbon cycle components. The initialization months and forecast periods of the CHFP models are listed in the third and fourth columns, respectively.

CHFP model	Corresponding CMIP5 model	Initialization months	Forecast period
CCCma-CanCM4	CanESM2	Feb, May, Aug, Nov	1979–2008
MIROC5	MIROC5	Feb, May, Aug, Nov	1979–2011
MPI-ESM-LR	MPI-ESM-LR	May, Nov	1982–2011
JMAMRI-CGCM3	MRI-CGCM3	Feb, May, Aug, Nov	1979–2010

d. Comparison of M-LIMs with the corresponding GCM forecasts

An important question is whether the skill of the M-LIMs is indicative of the skill of their parent GCMs when those are run in forecast mode. A strong link is desirable for our purposes because our ultimate goal is to draw conclusions about the skill of GCM prediction systems. To address this question, we use the four GCMs that have both piControl simulations and seasonal predictions available (CanESM2, MIROC5, MPI-ESM-LR, MRI-CGCM3; see Table 3). Of those four models, the MIROC5 uses an initialization technique in which the observed anomalies are added to the climatology of the model. This technique, called anomaly initialization, largely avoids model drift during the forecast, which is thought to be beneficial for decadal predictions (Mochizuki et al. 2010) but not commonly used in seasonal prediction, although some results suggest it may increase skill there as well (Mulholland et al. 2015). The other three models use conventional full-field initialization.

The CHFP seasonal prediction archive provides hindcasts initialized from various calendar months and varying hindcast periods. See Table 3 for the details of each model. For each GCM, we calculate the skill for each initialization month and lead time. These calculations are performed on our four indices of interest. To obtain a fair comparison we match these forecasts with the corresponding LIM, so that each skill data point of the GCM hindcast has a corresponding LIM data point. The results are displayed in a scatterplot, where each point corresponds to the skill for one particular initialization and lead month (Fig. 7).

Generally, the ACC of the LIMs and their parent GCMs is quite close (Fig. 7a). The latter tend to be somewhat higher (points below the 1-to-1 line), however, which is particularly evident for the Niño-3.4 index. The correlations across all data points for a given model are relatively high (Table 4), ranging from 0.58 (MPI-ESM-LR) to 0.93 (MRI-CGCM3). It could be argued that, in some regions, the relatively high correlation between M-LIM and GCM skill is due to both of them being close to the predictability limit. The Niño-3.4 index, however, indicates that even when the skill of the M-LIM is systematically lower, there is a strong relation with the GCM skill: calculating correlations for the Niño-3.4 forecasts gives values between 0.84 (CanESM2) and 0.93 (MRI-CGCM3).

For RMSE, the correlations between GCM and LIM predictions are generally higher (Fig. 7b), with values ranging from 0.84 (MPI-ESM-LR) to 0.89 (CanESM2 and MRI-CGCM3).

Again, the GCM predictions tend to outperform the LIM predictions, particularly for the Niño-3.4 index.

The fact that LIM predictions can sometimes outperform those of their parent GCMs may seem surprising because the former are built from just one output variable, namely SST, and are based on a relatively simple linear technique. On the other hand, the LIM does not have to deal with problems arising from model initialization, including model drift, which may give it some advantage.

e. POP analysis

As explained in section 2, POPs are the eigenvectors of the linear operator and can be understood as the normal modes of the system. Since the propagator matrix can be expressed in terms of these eigenvectors (e.g., Vimont 2012), POPs offer a way of analyzing the variability patterns that contribute most to the prediction skill of the LIM. However, since the POPs are generally not orthogonal to each other, there is no straightforward way to calculate their explained variance as in the case of PCs. Thus, the predictive power of each POP has to be assessed by performing predictions using that POP only. This can be done by setting all other eigenvalues to 0 in the diagonal matrix that results from the eigenvector decomposition. This ensures that the contribution of those POPs to the prediction vanishes [Vimont (2012) used a similar technique but set the damping time scale to 0].

Our interest is to examine whether the POPs with higher prediction skill are also more realistic. We thus select, for each model, the POP with the highest prediction skill for the Niño-3.4 and ATL3 indices. After ranking these models, we select the ones with the highest, median, and lowest skill and plot the corresponding POPs in Fig. 8, together with the best POP from observations.

The real part of the observational POP shows the familiar ENSO pattern, with highest loadings in the eastern equatorial Pacific (Fig. 8a). The imaginary component, which can be considered as the developing phase of the event, shows similar equatorial confinement but is more evenly spread across the equator. The oscillation period of the POP is about 4 years and thus well within the spectral range of ENSO. The patterns look similar to some previous POP analyses (e.g., Kleeman and Moore 1999; Gehne et al. 2014), although other studies using different analysis techniques have suggested that the imaginary POP should be more pronounced in the northern tropical Pacific and be related to the Pacific meridional mode (PMM; Chiang and Vimont 2004). Potential reasons for these differences

FIG. 7. Skill of select M-LIMs vs skill of the corresponding GCM run in forecast mode in the CHFP data archive. Skills are calculated for four indices (Niño-3.4 in green, STA in pink, NTA in blue, ATL3 in red) using all available lead times and initialization months. The skill metrics shown are (a) ACC and (b) RMSE. All dots above the diagonal 1-to-1 line indicate M-LIMs outperforming the O-LIM for ACC and vice versa for RMSE.

TABLE 4. Correlation coefficients of the scatterplots shown in Fig. 7a (top 5 rows) and Fig. 7b (bottom 5 rows). The correlation coefficients are calculated for all four indices separately (ATL3, NTA, STA, and Niño-3.4), and for all indices combined.

Index	Model			
	CanESM2	MIROC5	MPI-ESM-LR	MRI-CGCM3
ACC ATL3	0.656	0.685	0.620	0.916
ACC NTA	0.881	0.424	0.009	0.783
ACC STA	0.775	0.608	0.608	0.938
ACC Niño-3.4	0.843	0.858	0.913	0.933
ACC all	0.827	0.704	0.582	0.930
RMSE ATL3	0.954	0.628	0.777	0.943
RMSE NTA	0.967	0.825	0.621	0.912
RMSE STA	0.900	0.797	0.444	0.945
RMSE Niño-3.4	0.902	0.802	0.868	0.921
RMSE all	0.891	0.869	0.835	0.894

to previous analyses are manifold, including data source, analysis period, analysis region (global tropics, in our case), variables used, and the choice of covariance lag (τ_0).

The M-LIM with the most successful Niño-3.4 prediction, based on the IPSL-CM5B-LR, displays a structure that is very similar to that of the observations, though slightly shifted toward the center of the basin. Additionally, it shows some subtropical anomalies in the imaginary POP. At ~ 3 years, the oscillation period is shorter than that of the observations but still within the range of ENSO behavior.

The CanESM2, with its Niño-3.4 skill in the medium range, displays some obvious deficiencies in the spatial pattern as loadings are too high in the western tropical Pacific. This is related to a common GCM bias in which the thermocline is too shallow in the west, leading to excessive variability. Excessive westward extension of the equatorial cold tongue may also contribute to this. The oscillation period is below 3 years.

Finally, the M-LIM with the poorest Niño-3.4 prediction skill, the ACCESS1.0, has a real component that is confined to the eastern equatorial Pacific, while the imaginary component has negative loadings in the central equatorial Pacific. Furthermore, both real and imaginary components have loadings whose sign is opposite to the reference POP. The 40-month oscillation period is in the right range but the 12-month damping period is rather short. Overall, the comparison of the three models suggests an impact of tropical Pacific variability errors on ENSO prediction skill.

For the ATL3 index, the observation-derived POP with the highest skill is the same as for Niño-3.4 (Fig. 8b; note that the sign has been reversed to facilitate comparison with the other models). The peak phase (shading in Fig. 8b) shows positive loadings in the eastern equatorial Atlantic, while the developing phase shows positive loadings in the South Atlantic, consistent with some previous studies (e.g., Nnamchi et al. 2016). This is accompanied by La Niña conditions in the equatorial Pacific, which is also consistent with observations (Chang et al. 2006a; Lübbecke and McPhaden 2012; Tokinaga et al. 2019).

The M-LIM with the highest ATL3 prediction skill, CCSM4, shows a realistic full-fledged AZM pattern, although the positive loadings in the northeastern tropical Atlantic have the opposite sign as in observations. The oscillation period is

somewhat long (~ 6 years) and the damping much stronger than in the observations.

The middle-of-the-road M-LIM, GFDL CM3, features a purely damped pattern (imaginary component equal to zero), although the spatial pattern of its real POP bears relatively good resemblance to the AZM. Finally, the worst-performing M-LIM, NorESM1-ME, shows a reasonably realistic pattern on the equator, but unrealistically high loadings in the western equatorial Atlantic. The damping time is very short damping time (~ 5 months), suggesting that this POP too mostly represents a heavily damped mode.

Overall, the results suggest that, in the equatorial Atlantic, there is no clear relation between the skill of a POP and the realism of its spatial pattern and oscillation period. The reason why this link may be weaker for the equatorial Atlantic becomes clearer when examining the 10 POPs with the highest prediction skill (Fig. 9). For the Niño-3.4 region, there is a steep drop from POP 2 to POP 3, in both the observations and most models (POPs, when oscillatory, come in pairs, so that the first two POPs represent the first mode). Thus, the first mode carries the bulk of the predictive potential, with the other modes only incrementally adding to this. In the ATL3, on the other hand, the separation between the first and subsequent modes is much weaker, particularly in the M-LIMs but also in the O-LIM. This indicates that no individual POP is representative of the skill of a given LIM. Rather, the skill emerges from the combined effect of many POPs, though other studies argue that this should be the case for ENSO as well (Penland and Magorian 1993; Penland and Sardeshmukh 1995).

4. Link between systematic errors and skill

In section 3e we have attempted to relate the prediction skill obtained from individual POPs to their similarity with observations. This forms part of the wider problem of relating prediction skill to errors in variability patterns. In the present section we deepen this analysis by introducing another measure of variability error and by relating both variability and mean state errors to prediction skill.

As an additional measure of how close simulated variability is to observations, we calculate lagged regression patterns of the Niño-3.4 and ATL3 indices with tropical

FIG. 8. Real (shading) and imaginary (contours) components of principal oscillation patterns (POPs) calculated for O-LIM and three M-LIMs. Units are arbitrary. All POPs have been scaled according to the method described in Gallagher et al. (1991); see text for details. The damping time and oscillation period (months) are indicated above each panel. The POPs shown for each LIM are the ones with the highest skill in the (a) Niño-3.4 and (b) ATL3 region. The M-LIM for each region are chosen as the LIMs with rank 1 (best), 18 (medium), and 35 (worst) in the intermodel comparison.

SSTs. The rationale for choosing this particular measure is that it is closely related to the covariance calculations that form the basis of the LIMs. This represents a more intuitive measure of variability errors than the POPs, while at the same time linking to the prediction skill of LIMs.

Analogously to section 3e, we only show regression patterns for the observations and three models, the one with the highest, median, and lowest performance metric. The metric used here is the RMSE of the regression coefficient averaged over the tropical Pacific (Fig. 10a) and tropical Atlantic (Fig. 10b). The lags chosen for the regression are 7 months for the Pacific and 3 months for the Atlantic, which reflects the different time scales of the phenomena of interest. The Niño-3.4 regression pattern for the observations (Fig. 10a, top row) shows high

positive values in the central and eastern tropical Pacific. Additionally, it shows evidence for a PMM-like precursor (e.g., Penland and Sardeshmukh 1995; Chang et al. 2007) extending from the California coast toward the equator. The model with the most similar regression pattern is CESM1-CAM5, although the high values along the equator extend a little too far west, and the precursor signature is weak. The median model, ACCESS1.3, has too high values in the eastern equatorial Pacific. Finally, the weakest model in terms of the regression metric, MIROC-ESM-CHEM, shows poor equatorial confinement, with high values extending far off the equator.

For the ATL3 regression, the observations show high values in the eastern equatorial Atlantic that extend poleward along the southwest African coast (Fig. 10b, top row). Further

FIG. 9. ACC of forecasts using the 10 most skillful POPs in the (a) Niño-3.4 and (b) ATL3 region. The thick black line shows O-LIM; the thin blue lines show all 35 M-LIMs.

positive values are found in the central southern tropical Atlantic. The ACCESS1.0 model captures this pattern very well, though the signal is more confined to the equator and the southwest African coast. The median model, MIROC4h, also has a relatively realistic pattern in the equatorial region but excessive values in the southwestern tropical Atlantic. Finally, GFDL-ESM2M shows very weak positive regression coefficients along the equator and southwest African coast, while producing spurious negative values just north of the equator.

We first examine the global tropics for a link between mean state biases and prediction skill. The metrics used for bias and skill are the annual mean of the absolute SST bias and the RMSE at lead time 7 months, respectively, averaged over the global tropics. The multimodel scatterplot of these quantities (Fig. 11a) suggests that there is no systematic relation between them as the intermodel correlation is 0.01.

We also test the simplest metric for variability error, the bias of the SST standard deviation, for its relation to global tropical RMSE (Fig. 11b). The correlation coefficient is 0.34, which suggests that, counterintuitively, higher variability errors are linked to higher prediction skill. The correlation, however, is not significant at the 95% level.

As the global metrics show little link between bias and skill, we move on to the regression patterns described at the beginning of this section. To obtain a representative variability

metric for each M-LIM, we calculate the RMSE of its regression pattern relative to the observations (see Fig. 10) and take the area average. For the Niño-3.4 index, we take the regression patterns at lag 9 months. This value is scattered against the Niño-3.4 ACC for predictions initialized in February at lead time 9 months (Fig. 12a). The lag/lead time of 9 months is chosen because models exhibit a wide range of performances at this interval. For shorter lead times, most M-LIMs perform equally well, whereas for longer lead times most perform equally poorly. For similar reasons, February is chosen as the initialization month to test performance across the persistence barrier. We have experimented with different settings but found that they tend to yield lower intermodel correlations (this is also true for the other panels in Fig. 12).

The scatterplot (Fig. 12a) suggests that models with a larger error in the regression pattern also have lower ACC in the Niño-3.4 prediction. At -0.49 the correlation is relatively strong and significant at the 95% level. This high value, however, depends to some extent on two extreme values (labeled “1” and “2” in Fig. 12a). Once these are removed the correlation drops to -0.32 , just below the 95% level.

We also test RMSE as a measure for prediction skill (Fig. 12c). This suggests a slightly stronger relation with a correlation coefficient of 0.52 (since RMSE decreases with skill, this indicates a relation between variability errors and skill that is of the same sign as in Fig. 12a). The same two outliers as in Fig. 12a contribute substantially to this correlation.

For the equatorial Atlantic, we choose a 3-month lag for the SST regression patterns with the ATL3 and 3-month lead time for the ATL3 ACC. Here we focus on March initializations to test the models’ ability to predict AZM events, which typically peak in JJA. The two quantities are negatively related (Fig. 12b), with a correlation coefficient of -0.17 . At 0.05, the correlation for the RMSE skill metric (Fig. 12d) is even lower. We note that the intermodel correlations for ACC and RMSE are equally weak when M-LIM_TATL is used (not shown). Li et al. (2020) found a consistent improvement of the skill of their LIM when they constructed it from a less biased high-resolution version of their GCM. The low-resolution version of their GCM, however, had a bias that was more severe than in typical CMIP5 GCMs, which may explain the robust performance gain.

In addition to point correlations, we can examine spatial correlation patterns of SST bias and skill along the model dimension. The goal is to analyze how the skill in the equatorial Pacific and Atlantic is related to SST errors across the tropics. Choosing the same forecast skill measures as in Figs. 12a and 12b (i.e., ACC for the Niño-3.4 and ATL3 at lead months 9 and 3, respectively), we correlate these with the absolute SST bias at each grid point along the model dimension. For the Niño-3.4 region (Fig. 13a) the spatial correlation pattern shows significant values in the western tropical Pacific, with negative correlations as low as -0.5 . Weaker negative correlations are found in the southern tropical Atlantic and southern Indian Ocean. These negative correlations suggest that larger absolute biases are associated with lower ENSO prediction skill. Since SST biases in those regions are typically negative (Fig. 13c), this means that models with cooler SST bias have lower ENSO prediction skill. In the tropical Pacific, this may be

FIG. 10. Regression of SST anomalies on (a) the Niño-3.4 index leading by 7 months and (b) the ATL3 index leading by 3 months. The rows show the observations (NCEP–NCAR reanalysis) and three selected CMIP5 models. The latter were chosen as the models with rank 1 (best), 18 (medium), and 35 (worst) in terms of their similarity with observed regression pattern (top row; see text for details).

associated with the cold SST anomalies suppressing deep convection and coupled air–sea feedbacks. These findings are consistent with those of [Ding et al. \(2020\)](#).

For the ATL3 index ([Fig. 13b](#)), there are areas of both negative and positive correlations in the tropical Atlantic, although none of them is statistically significant. Positive correlations are located in the northern tropical Atlantic, while negative correlations are found in the southern tropical Atlantic. Comparison with the mean bias of the model ensemble ([Fig. 13c](#)) suggests that a cold bias in the northern tropical Atlantic is associated with higher AZM prediction skill, while warm biases in the southeastern

tropical Atlantic are associated with lower AZM prediction skill. The former is rather counterintuitive and not easy to interpret. A possible explanation is that the cold NTA bias enhances the trade winds and thereby strengthens the Bjerknes feedback. The latter could be interpreted as an overly deep thermocline (which is associated with warm SST biases) leading to reduced SST variability and prediction skill. Examining the bias of SST standard deviation ([Fig. 13f](#)) appears to confirm that models with warm SST biases in the region have lower variability. This link of variability errors to prediction skill is also hinted at by an intermodel correlation analogous to that in [Fig. 13b](#), which uses the bias of SST standard

FIG. 11. Multimodel scatterplots of quantities averaged over the tropical oceans (30°S–30°N). (a) Climatological annual mean SST bias vs the RMSE of the corresponding M-LIM at lead month 7. (b) Climatological annual mean bias of the standard deviation of SST vs the RMSE of the corresponding M-LIM at lead month 7. The correlation of the multimodel scatter is noted in the lower right.

deviation instead of the bias of SST itself (Fig. 13e). There are weakly negative correlations in the equatorial Atlantic, which, considering the generally underestimated variability (Fig. 13f), suggests that models with stronger variability have better prediction skill. Examining a similar plot for the Niño-3.4 index, (Fig. 13d) shows only a few patches of borderline significant correlation.

5. Summary and conclusions

a. Summary

We have constructed linear inverse models (LIMs) from both observations and model output of SST. Our goals were to establish a baseline for tropical SST predictions; to examine the linkages between mean state bias, variability error, and prediction skill; and to examine to the extent to which the skill of a LIM derived from the output of a particular GCM is indicative of that GCM's skill when run in forecast mode. The major findings of our study are as follows:

- 1) The LIM constructed from observed SST performs better than persistence in most regions. In the tropical Pacific, the ACC for the Niño-3.4 index is above 0.5 up to a lead time of 12 months. In the southern tropical Atlantic, the LIM performs worse than persistence, although skill matches performance when the LIM is constructed from tropical Atlantic SSTs only.

- 2) On average, LIMs generated from free-running GCM simulations (M-LIMs) are comparable to the LIM generated from observations, and some GCM-derived LIMs are more skillful. While the GCM output suffers from some variability errors, these are often not severe. Thus, the long record available for LIM construction may outweigh the model errors, as it mitigates the overfitting problem.
- 3) The GCM-derived LIMs are competitive with some full-fledged GCM prediction systems, particularly in the tropical Atlantic. It remains an open question whether this narrow performance gap is due to current GCM deficiencies or whether it is due to both forecast techniques approaching the predictability limit.
- 4) The LIMs generated from GCM SST output provide a fairly good measure for the skill achieved by their parent GCMs when run in forecast mode. Thus, the SST variability generated by the model provides sufficient information to estimate GCM skill. Furthermore, in some cases the skill of the LIM is actually superior to that of its parent GCM. This indicates that the GCM predictions may suffer from initialization problems (initialization shock etc.), which are not an issue for the LIM.
- 5) There is a weak-to-moderate relation between model mean state error and prediction skill derived from M-LIMs in some regions. There is some evidence for this in the eastern equatorial Atlantic, where an erroneously deep thermo-

FIG. 12. Multimodel scatterplots of variability errors in each CMIP5 model vs the skill of the corresponding M-LIM. Two regions [(left) Niño-3.4 and (right) ATL3] and two skill metrics [(a),(b) ACC and (c),(d) RMSE] are considered. The variability error is calculated by taking the difference between regression patterns obtained from observations and model simulations. See text for details. The intermodel correlation is shown in the lower right.

cline reduces SST variability, possibly due to the weakened Bjerknes feedback, which in turn affects prediction skill. In the equatorial Pacific, skill appears to be linked to cold SST biases in the western tropical Pacific, consistent with the results of Ding et al. (2020). Such cold SST biases may reduce the strength of air–sea coupling in the region, which can affect the simulated variability. There is, however, no clear link between SST variability errors and the skill of the LIMs. Results from other regions not presented here show even less evidence for a link between systematic model errors and prediction skill.

Observed variability patterns in the tropical oceans involve not only SSTs but also subsurface temperatures through the Bjerknes feedback. One might therefore expect that including information on subsurface ocean temperatures into the LIMs could lead to a stronger intermodel relation between biases and prediction skill. This, however, is not borne out by our analysis of LIMs built from combined SST and SSH. Our

analysis shows that despite the fact that forecast skill of LIMs is generally improved by including SSH, particularly at long lead times, which is consistent with previous studies (Newman et al. 2011; Newman and Sardeshmukh 2017), there is little evidence of improvement in the relationship between LIM forecast skills and model biases (see the supplemental material for more details). This suggests that the main findings from SST-only-based LIM analysis are not changed by the inclusion of subsurface ocean temperature information.

b. Conclusions

One of the central questions of the current study is whether mean state and variability errors are a major impediment to prediction skill. This question is particularly pertinent to the tropical Atlantic, where mean state biases are typically large and skill is relatively low. While we do find some evidence for a link between the two, it appears that the relation is not very strong, with some GCM-derived LIMs achieving good skill in

FIG. 13. Intermodel correlation of (a) Niño-3.4 ACC and SST bias, (b) ATL3 ACC and SST bias, (d) Niño-3.4 ACC and bias of standard deviation of SST, and (e) ATL3 ACC and bias of standard deviation of SST. (c) The annual mean SST bias (shading; K) and its standard deviation across the model ensemble (contours; K). (f) The annual mean bias of the standard deviation of SST (shading; K) and its standard deviation across the model ensemble (contours; K).

the equatorial Atlantic despite their parent model having severe mean state SST biases. In those models, the simulated variability patterns are typically quite realistic as well. The fact that the skill of GCM-derived LIMs is highly correlated with the skill of their parent GCM in forecast mode also indicates that the link between variability errors and prediction skill is relatively solid. Thus, for the equatorial Atlantic, it appears to be the link between mean state and variability errors that is weak. In other regions, additionally the link between variability errors and prediction skill appears to be less prominent.

The tentative answer to our central question is therefore that mean state errors are not the major stumbling block for skillful prediction skill for the current generation of climate models, which is consistent with previous work (Richter et al. 2018). This answer, however, cannot be definitive because we have not examined actual GCM predictions.

If model errors are not key for poor prediction skill in the tropical Atlantic, which factors are? Two possibilities are inherent predictability limits and initialization issues. Regarding the latter, it is of note that some LIM-derived GCMs tend to outperform their parent GCMs in the equatorial and southern tropical Atlantic. This suggests that improved initialization methods may also improve prediction skill in those models.

Acknowledgments. The authors thank the three anonymous reviewers for their constructive comments. Thanks to Shoichiro Kido for his assistance with performing the tau test. We acknowledge the WCRP's Working Group on Coupled Modelling, which is responsible for CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison which provides coordinating support and led development of software infrastructure for CMIP, and the climate modeling groups for making available their model output. We acknowledge the WCRP/CLIVAR Working Group on Seasonal to Interannual Prediction (WGSIP) for establishing the Climate-system Historical Forecast Project (CHFP), and the Centro de Investigaciones del Mar y la Atmosfera (CIMA) for providing the model output <http://chfps.cima.fcen.uba.ar/>. We also thank the data providers for making the model output available through CHFP. Ingo Richter was partially supported by the Japan Society for the Promotion of Science, KAKENHI Grant 18H01281. P.C. and X.L. acknowledge the support of NSF Grant AGS-1462127 and of the International Laboratory for High-Resolution Earth System Prediction (IHESP).

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>.
- Amaya, D. J., M. J. DeFlorio, A. J. Miller, and S.-P. Xie, 2017: WES feedback and the Atlantic meridional mode: Observations and CMIP5 comparisons. *Climate Dyn.*, **49**, 1665–1679, <https://doi.org/10.1007/s00382-016-3411-1>.
- Barnston, A. G., M. K. Tippet, M. Ranganathan, and M. L. L'Heureux, 2019: Deterministic skill of ENSO predictions from the North American multimodel ensemble. *Climate Dyn.*, **53**, 7215–7234, <https://doi.org/10.1007/s00382-017-3603-3>.
- Bjerknes, J., 1969: Atmospheric teleconnections from the equatorial Pacific. *Mon. Wea. Rev.*, **97**, 163–172, [https://doi.org/10.1175/1520-0493\(1969\)097<0163:ATFTEP>2.3.CO;2](https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2).
- Capotondi, A., and P. D. Sardeshmukh, 2015: Optimal precursors of different types of ENSO events. *Geophys. Res. Lett.*, **42**, 9952–9960, <https://doi.org/10.1002/2015GL066171>.
- Carton, J. A., and B. Huang, 1994: Warm events in the tropical Atlantic. *J. Phys. Oceanogr.*, **24**, 888–903, [https://doi.org/10.1175/1520-0485\(1994\)024<0888:WEITTA>2.0.CO;2](https://doi.org/10.1175/1520-0485(1994)024<0888:WEITTA>2.0.CO;2).
- Chang, P., L. Ji, and H. Li, 1997: A decadal climate variation in the tropical Atlantic Ocean from thermodynamic air–sea interactions. *Nature*, **385**, 516–518, <https://doi.org/10.1038/385516a0>.
- , —, and R. Saravanan, 2001: A hybrid coupled model study of tropical Atlantic variability. *J. Climate*, **14**, 361–390, [https://doi.org/10.1175/1520-0442\(2001\)013<0361:AHCMO>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)013<0361:AHCMO>2.0.CO;2).
- , Y. Fang, R. Saravanan, L. Ji, and H. Seidel, 2006a: The cause of the fragile relationship between the Pacific El Niño and the Atlantic Niño. *Nature*, **443**, 324–328, <https://doi.org/10.1038/nature05053>.
- , and Coauthors, 2006b: Climate fluctuations of tropical coupled systems—The role of ocean dynamics. *J. Climate*, **19**, 5122–5174, <https://doi.org/10.1175/JCLI3903.1>.
- , L. Zhang, R. Saravanan, D. J. Vimont, J. C. H. Chiang, L. Ji, H. Seidel, and M. K. Tippet, 2007: Pacific meridional mode and El Niño–Southern Oscillation. *Geophys. Res. Lett.*, **34**, L16608, <https://doi.org/10.1029/2007GL030302>.
- Chiang, J. C. H., and D. J. Vimont, 2004: Analogous Pacific and Atlantic meridional modes of the tropical atmosphere–ocean variability. *J. Climate*, **17**, 4143–4158, <https://doi.org/10.1175/JCLI4953.1>.
- Chiodi, A. M., and D. E. Harrison, 2017: Observed El Niño SSTA development and the effects of easterly and westerly wind events in 2014/15. *J. Climate*, **30**, 1505–1519, <https://doi.org/10.1175/JCLI-D-16-0385.1>.
- Davey, M. K., and Coauthors, 2002: STOIC: A study of coupled model climatology and variability in tropical ocean regions. *Climate Dyn.*, **18**, 403–420, <https://doi.org/10.1007/s00382-001-0188-6>.
- DelSole, T., and J. Shukla, 2010: Model fidelity versus skill in seasonal forecasting. *J. Climate*, **23**, 4794–4806, <https://doi.org/10.1175/2010JCLI3164.1>.
- Ding, H., N. Keenlyside, M. Latif, W. Park, and S. Wahl, 2015: The impact of mean state errors on equatorial Atlantic interannual variability in a climate model. *J. Geophys. Res. Oceans*, **120**, 1133–1151, <https://doi.org/10.1002/2014JC010384>.
- , M. Newman, M. A. Alexander, and A. T. Wittenberg, 2018: Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs. *J. Climate*, **31**, 5437–5459, <https://doi.org/10.1175/JCLI-D-17-0661.1>.
- , —, —, and —, 2019: Diagnosing secular variations in retrospective ENSO seasonal forecast skill using CMIP5 model-analogs. *Geophys. Res. Lett.*, **46**, 1721–1730, <https://doi.org/10.1029/2018GL080598>.
- , —, —, and —, 2020: Relating CMIP5 model biases to seasonal forecast skill in the tropical Pacific. *Geophys. Res. Lett.*, **47**, e2019GL086765, <https://doi.org/10.1029/2019GL086765>.
- Gallagher, F., H. von Storch, R. Schnur, and G. Hannoschöck, 1991: The POP manual. Deutsches KlimaRechenZentrum Tech. Rep. 1, 64 pp.
- Gehne, M., R. Kleeman, and K. E. Trenberth, 2014: Irregularity and decadal variation in ENSO: A simplified model based on principal oscillation patterns. *Climate Dyn.*, **43**, 3327–3350, <https://doi.org/10.1007/s00382-014-2108-6>.

- Gualdi, S., A. Alessandri, and A. Navarra, 2005: Impact of atmospheric horizontal resolution on El Niño Southern Oscillation forecasts. *Tellus*, **57A**, 357–374, <https://doi.org/10.3402/tellusa.v57i3.14662>.
- Ham, Y., J. Kim, and J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568–572, <https://doi.org/10.1038/s41586-019-1559-7>.
- Hurrell, J. W., 2001: The North Atlantic oscillation. *Science*, **291**, 603–605, <https://doi.org/10.1126/science.1058761>.
- Imada, Y., H. Tatebe, M. Watanabe, M. Ishii, and M. Kimoto, 2016: South Pacific influence on the termination of El Niño in 2014. *Sci. Rep.*, **6**, 30341, <https://doi.org/10.1038/srep30341>.
- Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, [https://doi.org/10.1175/1520-0477\(1996\)077<0437:TNYRP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2).
- Kirtman, B., and A. Pirani, 2009: The state of the art of seasonal prediction: Outcomes and recommendations from the First World Climate Research Program Workshop on Seasonal Prediction. *Bull. Amer. Meteor. Soc.*, **90**, 455–458, <https://doi.org/10.1175/2008BAMS2707.1>.
- , and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, <https://doi.org/10.1175/BAMS-D-12-00050.1>.
- Kleeman, R., and A. M. Moore, 1999: A new method for determining the reliability of dynamical ENSO predictions. *Mon. Wea. Rev.*, **127**, 694–705, [https://doi.org/10.1175/1520-0493\(1999\)127<0694:ANMFDT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<0694:ANMFDT>2.0.CO;2).
- Lee, J.-Y., and Coauthors, 2010: How are seasonal prediction skills related to models' performance on mean state and annual cycle?. *Climate Dyn.*, **35**, 267–283, <https://doi.org/10.1007/s00382-010-0857-4>.
- Li, X., M. H. Bordbar, M. Latif, W. Park, and J. Harlaß, 2020: Monthly to seasonal prediction of tropical Atlantic sea surface temperature with statistical models constructed from observations and data from the Kiel Climate Model. *Climate Dyn.*, **54**, 1829–1850, <https://doi.org/10.1007/s00382-020-05140-6>.
- Lübbecke, J. F., and M. J. McPhaden, 2012: On the inconsistent relationship between Pacific and Atlantic Niños. *J. Climate*, **25**, 4294–4303, <https://doi.org/10.1175/JCLI-D-11-00553.1>.
- , B. Rodríguez-Fonseca, I. Richter, M. Martín-Rey, T. Losada, I. Polo, and N. Keenlyside, 2018: Equatorial Atlantic variability—Modes, mechanisms, and global teleconnections. *Wiley Interdiscip. Rev.: Climate Change*, **9**, e527, <https://doi.org/10.1002/WCC.527>.
- Luo, J.-J., S. Masson, S. K. Behera, and T. Yamagata, 2008: Extended ENSO predictions using a fully coupled ocean–atmosphere model. *J. Climate*, **21**, 84–93, <https://doi.org/10.1175/2007JCLI1412.1>.
- Magnusson, L., M. Alonso-Balmaseda, S. Corti, F. Molteni, and T. Stockdale, 2013: Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. *Climate Dyn.*, **41**, 2393–2409, <https://doi.org/10.1007/s00382-012-1599-2>.
- Manganello, J. V., and B. Huang, 2009: The influence of systematic errors in the Southeast Pacific on ENSO variability and prediction in a coupled GCM. *Climate Dyn.*, **32**, 1015–1034, <https://doi.org/10.1007/s00382-008-0407-5>.
- McGregor, S., M. F. Stuecker, J. B. Kajtar, M. H. England, and M. Collins, 2018: Model tropical Atlantic biases underpin diminished Pacific decadal variability. *Nat. Climate Change*, **8**, 493–498, <https://doi.org/10.1038/s41558-018-0163-4>.
- McPhaden, M. J., 2015: Playing hide and seek with El Niño. *Nat. Climate Change*, **5**, 791–795, <https://doi.org/10.1038/nclimate2775>.
- Merle, J., 1980: Annual and interannual variability of temperature in the eastern equatorial Atlantic Ocean—Hypothesis of an Atlantic El Niño. *Oceanol. Acta*, **3**, 209–220.
- Mochizuki, T., and Coauthors, 2010: Pacific decadal oscillation hindcasts relevant to near-term climate prediction. *Proc. Natl. Acad. Sci. USA*, **107**, 1833–1837, <https://doi.org/10.1073/pnas.0906531107>.
- Mulholland, D. P., P. Laloyaux, K. Haines, and M. A. Balmaseda, 2015: Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts. *Mon. Wea. Rev.*, **143**, 4631–4644, <https://doi.org/10.1175/MWR-D-15-0076.1>.
- Newman, M., and P. D. Sardeshmukh, 2017: Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures? *Geophys. Res. Lett.*, **44**, 8520–8529, <https://doi.org/10.1002/2017GL074088>.
- , M. A. Alexander, and J. D. Scott, 2011: An empirical model of tropical ocean dynamics. *Climate Dyn.*, **37**, 1823–1841, <https://doi.org/10.1007/s00382-011-1034-0>.
- Nnamchi, H. C., J. Li, F. Kucharski, I.-S. Kang, N. S. Keenlyside, P. Chang, and R. Farneti, 2016: An equatorial–extratropical dipole structure of the Atlantic Niño. *J. Climate*, **29**, 7295–7311, <https://doi.org/10.1175/JCLI-D-15-0894.1>.
- Penland, C., 1989: Random forcing and forecasting using principal oscillation pattern analysis. *Mon. Wea. Rev.*, **117**, 2165–2185, [https://doi.org/10.1175/1520-0493\(1989\)117<2165:RFAFUP>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<2165:RFAFUP>2.0.CO;2).
- , 2019: The Nyquist issue in linear inverse modeling. *Mon. Wea. Rev.*, **147**, 1341–1349, <https://doi.org/10.1175/MWR-D-18-0104.1>.
- , and T. Magorian, 1993: Prediction of Niño-3 sea surface temperatures using linear inverse modeling. *J. Climate*, **6**, 1067–1076, [https://doi.org/10.1175/1520-0442\(1993\)006<1067:PONSST>2.0.CO;2](https://doi.org/10.1175/1520-0442(1993)006<1067:PONSST>2.0.CO;2).
- , and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J. Climate*, **8**, 1999–2024, [https://doi.org/10.1175/1520-0442\(1995\)008<1999:TOGOTS>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2).
- , and L. Matrosova, 1998: Prediction of tropical Atlantic sea surface temperatures using linear inverse modeling. *J. Climate*, **11**, 483–496, [https://doi.org/10.1175/1520-0442\(1998\)011<0483:POTASS>2.0.CO;2](https://doi.org/10.1175/1520-0442(1998)011<0483:POTASS>2.0.CO;2).
- Prigent, A., J. Lübbecke, T. Bayr, M. Latif, and C. Wengel, 2020: Weakened SST variability in the tropical Atlantic Ocean since 2000. *Climate Dyn.*, **54**, 2731–2744, <https://doi.org/10.1007/s00382-020-05138-0>.
- Richter, I., and S.-P. Xie, 2008: On the origin of equatorial Atlantic biases in coupled general circulation models. *Climate Dyn.*, **31**, 587–598, <https://doi.org/10.1007/s00382-008-0364-z>.
- , and T. Doi, 2019: Estimating the role of SST in atmospheric surface wind variability over the tropical Atlantic and Pacific. *J. Climate*, **32**, 3899–3915, <https://doi.org/10.1175/JCLI-D-18-0468.1>.
- , and H. Tokinaga, 2020: The Atlantic Niño: Dynamics, thermodynamics, and teleconnections. *Tropical and Extratropical Air–Sea Interactions*, S. K. Behera, Ed., Elsevier, 171–205, <https://doi.org/10.1016/B978-0-12-818156-0.00008-3>.
- , S. K. Behera, Y. Masumoto, B. Taguchi, H. Sasaki, and T. Yamagata, 2013: Multiple causes of interannual sea surface temperature variability in the equatorial Atlantic Ocean. *Nat. Geosci.*, **6**, 43–47, <https://doi.org/10.1038/ngeo1660>.
- , S.-P. Xie, S. K. Behera, T. Doi, and Y. Masumoto, 2014a: Equatorial Atlantic variability and its relation to mean state biases in CMIP5. *Climate Dyn.*, **42**, 171–188, <https://doi.org/10.1007/s00382-012-1624-5>.

- , S. K. Behera, T. Doi, B. Taguchi, Y. Masumoto, and S.-P. Xie, 2014b: What controls equatorial Atlantic winds in boreal spring? *Climate Dyn.*, **43**, 3091–3104, <https://doi.org/10.1007/s00382-014-2170-0>.
- , T. Doi, S. K. Behera, and N. Keenlyside, 2018: On the link between mean state biases and prediction skill in the tropics: An atmospheric perspective. *Climate Dyn.*, **50**, 3355–3374, <https://doi.org/10.1007/s00382-017-3809-4>.
- Servain, J., I. Wainer, J. P. McCreary, and A. Dessier, 1999: Relationship between the equatorial and meridional modes of climatic variability in the tropical Atlantic. *Geophys. Res. Lett.*, **26**, 485–488, <https://doi.org/10.1029/1999GL900014>.
- Sheshadri, A., and R. A. Plumb, 2017: Propagating annular modes: Empirical orthogonal functions, principal oscillation patterns, and time scales. *J. Atmos. Sci.*, **74**, 1345–1361, <https://doi.org/10.1175/JAS-D-16-0291.1>.
- Stockdale, T. N., M. A. Balmaseda, and A. Vidard, 2006: Tropical Atlantic SST prediction with coupled ocean–atmosphere GCMs. *J. Climate*, **19**, 6047–6061, <https://doi.org/10.1175/JCLI3947.1>.
- Tang, B., 1995: Periods of linear development of the ENSO cycle and POP forecast experiments. *J. Climate*, **8**, 682–691, [https://doi.org/10.1175/1520-0442\(1995\)008<0682:POLDOT>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<0682:POLDOT>2.0.CO;2).
- Tokinaga, H., and S.-P. Xie, 2011: Weakening of the equatorial Atlantic cold tongue over the past six decades. *Nat. Geosci.*, **4**, 222–226, <https://doi.org/10.1038/ngeo1078>.
- , I. Richter, and Y. Kosaka, 2019: ENSO influence on the Atlantic Niño, revisited: Multi-year versus single-year ENSO events. *J. Climate*, **32**, 4585–4600, <https://doi.org/10.1175/JCLI-D-18-0683.1>.
- Vimont, D. J., 2012: Analysis of the Atlantic meridional mode using linear inverse modeling: Seasonality and regional influences. *J. Climate*, **25**, 1194–1212, <https://doi.org/10.1175/JCLI-D-11-00012.1>.
- von Storch, H., G. Bürger, R. Schnur, and J.-S. von Storch, 1995: Principal oscillation patterns. A review. *J. Climate*, **8**, 377–400, [https://doi.org/10.1175/1520-0442\(1995\)008<0377:POPAP>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<0377:POPAP>2.0.CO;2).
- Xie, S.-P., and S. G. H. Philander, 1994: A coupled ocean–atmosphere model of relevance to the ITCZ in the eastern Pacific. *Tellus*, **46A**, 340–350, <https://doi.org/10.3402/tellusa.v46i4.15484>.
- Zebiak, S. E., 1993: Air–sea interaction in the equatorial Atlantic region. *J. Climate*, **6**, 1567–1586, [https://doi.org/10.1175/1520-0442\(1993\)006<1567:AIITEA>2.0.CO;2](https://doi.org/10.1175/1520-0442(1993)006<1567:AIITEA>2.0.CO;2).
- Zhu, J., A. Kumar, B. Huang, M. A. Balmaseda, Z.-Z. Hu, L. Marx, and J. L. Kinter III, 2016: The role of off-equatorial surface temperature anomalies in the 2014 El Niño prediction. *Sci. Rep.*, **6**, 19677, <https://doi.org/10.1038/srep19677>.