A Dynamical Adjustment Approach to Estimating Forced and Internal Variability in the North Atlantic

Douglas Nedza,\textsuperscript{a}  Timothy DelSole,\textsuperscript{a,b}

\textsuperscript{a} Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, Virginia, USA
\textsuperscript{b} Center for Ocean-Land-Atmospheric Sciences, Fairfax, Virginia, USA

Corresponding author: Douglas Nedza, dnedza@gmu.com

Early Online Release: This preliminary version has been accepted for publication in Journal of Climate, may be fully cited, and has been assigned DOI 10.1175/JCLI-D-23-0651.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2024 American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the default AMS reuse license. For information regarding reuse and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuse Licenses).
ABSTRACT: Quantifying the relative contributions of external forcing and internal variability to North Atlantic Sea Surface Temperature (NASST) has important implications for attributing and predicting climate changes around the North Atlantic basin. Many previous methods have approached this problem by estimating the externally forced signal directly, making assumptions about forced variability for which there is no consensus. In this work, the separation of variability is approached in a fundamentally different way that does not specify the forced response’s temporal evolution. We propose a dynamical adjustment method in which the internal, spatially uniform component of NASST is predicted based on patterns of NASST that are orthogonal to the spatially uniform pattern. When applied to preindustrial simulations, the dynamical adjustment demonstrates skill in reconstructing the NASST basin mean variability. Applying the dynamical adjustment to historical simulations demonstrates skill in a majority of climate models, although the skill is reduced relative to preindustrial simulations because external variability partly contaminates the predictors. The dynamical adjustment is compared to several other methods which directly estimate the externally forced signal. We find that dynamical adjustment performs similarly to these comparative methods, despite the fundamentally different prediction method. However, methods based on different principles yield considerably different estimates of external and internal variability.
1. Introduction

Multi-decadal variability over the North Atlantic has significant societal impacts because of its influence on hurricane activity, shifts in the Intertropical Convergence Zone, summer monsoon rainfall over Sahel and India, and summer climate over Europe and North America (Goldenberg et al. 2001; Sutton and Hodson 2005; Knight et al. 2006; Zhang and Delworth 2006). Future projections of global and regional temperatures and associated uncertainties depend critically on whether observed variability is forced or generated internally. Forced variability refers to the response to changes in radiative forcing, which may include the following: changes in land cover or land use; changes in atmospheric concentration of greenhouse gases and anthropogenic aerosols; volcanic and natural aerosols; and changes in solar irradiance. Internal variability refers to variability that appears in the absence of changing external forcing due to dynamical instabilities and interactions within the coupled ocean-land-atmosphere system. There is no consensus about the relative importance of forced and internal variability in NASST. Some studies have suggested that anthropogenic aerosols are a predominant forcing (Booth et al. 2012). Others have attributed NASST variability to the effects of climate change, resulting in atmospheric changes that impact the ocean (He et al. 2022). Others have argued that internal variability within the coupled ocean-atmosphere system, particularly the North Atlantic Oscillation (NAO) and the Atlantic Meridional Overturning Circulation (AMOC), can contribute to decadal-scale NASST variability (Zhang et al. 2013). The relative importance over time and interactions between these different drivers of NASST variability remains an active area of research.

Following Frankignoul et al. (2017), methods for separating forced and internal variability may be classified into two categories. One category relies solely on observations to estimate external variability. For example, the externally forced response could be assumed a smooth function of time that can be represented by a low-order polynomial, with parameters estimated from observations. However, if internal variability generates multi-decadal variability, then regressing a polynomial out of observations may introduce error by overfitting to internal variability.

Another example in this category is to use the globally averaged SST to represent the temporal evolution of externally forced variability. Trenberth and Shea (2006) argue that the externally forced component can be removed by subtracting globally averaged SST from NASST, which assumes that the forced response of the North Atlantic is the same as the global average. Deser and Phillips...
(2021) demonstrate that regressing the global average SST out of grid point data outperforms the subtraction method of Trenberth and Shea (2006), correcting for spatial differences in the forced response, although potentially overfitting to internal variability. Deser and Phillips (2021) point out that this estimate of external variability is unable to capture the effect of external forcings that are unique to the North Atlantic region, such as the regional aerosol forcing implicated by Booth et al. (2012), or regional feedbacks of the climate system to external forcing that operate on a different timescale as global forcing (Armour et al. 2016). More sophisticated approaches in this category rely on assumptions about differences in time scales between the externally forced and internal variability, including locally weighted scatterplot smoothing (LOWESS) and ensemble empirical model decomposition (EEMD) (Cleveland 1979; Wu and Huang 2009). Other time scale based methods include the eigenmode decomposition of linear inverse models, both with and without optimal initial conditions (Frankignoul et al. 2017), and low-frequency component analysis, maximizing the ratio of low frequency to total variability (Wills et al. 2019).

A second category of methods uses physics-based models, called general circulation models (GCMs), that represent Earth’s coupled ocean-atmosphere system. Because these GCMs generate their own internal variability, ensemble techniques are used to estimate the externally forced response. As larger ensemble sizes produce more accurate estimates of a single GCM’s response to external forcings, there has been a recent emphasis on producing externally forced ensembles with tens to hundreds of members that can facilitate detailed analysis on regional and decadal scales (Kay et al. 2015). In addition, comparing large ensembles from several GCMs can provide a robust basis for examining the relative influence of internal variability and model differences in simulating the observed climate (Deser et al. 2020; Jain et al. 2023).

Within the context of separating components of variability, after an ensemble average is calculated, the estimate of the externally forced response can be removed via subtraction or with linear regression (Frankcombe et al. 2015; Steinman et al. 2015). More sophisticated approaches that utilize ensembles of GCM simulations, such as optimal fingerprinting (Allen and Tett 1999) and signal-to-noise maximizing EOF analysis (Ting et al. 2009; Wills et al. 2020) have also been applied. However, there is considerable debate about whether GCMs accurately represent the response to external forcings. In particular, weak observational constraints on aerosol forcings and parameterizations allow for a range of GCM behavior in response to aerosol forcing (Rotstayn
et al. 2015; Hourdin et al. 2017). If the ability of GCMs to produce realistic responses to historical forcings is in doubt, conclusions from methods that rely on a simulated response to external forcing will remain unconvincing.

There is a third category of methods that avoids specifying characteristics of forced variability. Specifically, these methods focus on identifying the internal contribution to observed variability. These methods are called dynamical adjustment, an example of which is the study of Wallace et al. (1995). Wallace et al. (1995) performed dynamical adjustment on observed Northern Hemisphere mean temperature by regressing out a single spatial pattern orthogonal to the hemispheric mean. This pattern, termed cold-ocean-warm-land (COWL), represented dynamically induced changes in the hemispheric mean associated with the configuration of warm and cold air masses. Removing COWL-associated variability produced a residual estimate of external variability exhibiting reduced interannual variability and enhanced autocorrelations, expected characteristics of the externally forced signal. This method avoids specifying the temporal shape or amplitude of externally forced variability, but makes the assumption that external variability is nearly spatially uniform and mostly absent in patterns that are orthogonal to the spatially uniform. Wallace et al. (1995) noted that the COWL pattern may contain fingerprints of external forcing, particularly high latitude continental warming. Other fingerprints, due to aerosols, trace gases, and changes in cloud cover, may possess distinct spatial structures that also project onto COWL. Other dynamical adjustment studies have employed Sea Level Pressure (SLP) to estimate and remove the effect of atmospheric circulations on surface temperature (Hurrell 1996). Using SLP in dynamical adjustment has been found to account for some of the effect of internal variability on surface temperature, reducing the spread of trends between different members of an GCM ensemble (Deser et al. 2014, 2016; Sippel et al. 2019), and trend discrepancies between GCMs and observations (Saffioti et al. 2016; Wallace et al. 2012).

In the context of quantifying variability in NASST, dynamical adjustment is particularly attractive because it avoids making strong assumptions about forced variability for which there is no consensus. In this paper, we develop such an approach. Specifically, we propose a dynamical adjustment method that uses a machine learning model to estimate the internal variability of a spatially uniform pattern. This pattern has a value equal to the basin-mean North Atlantic sea surface temperature (NASST). The physics-based machine learning model estimates the characteristics
of internal variability, particularly co-variability between the basin mean and spatial patterns that are orthogonal to the spatially uniform pattern. This co-variability is leveraged to estimate the internal contribution to basin mean fluctuations on the basis of temporal variance of the orthogonal patterns. SST is selected as the predictor over SLP because while some SST variability might be related to SLP, particularly that driven by the atmosphere, other variations arising from internal ocean dynamics may be unrelated to atmospheric circulation. The proposed method extends the approach of Wallace et al. (1995), where a single orthogonal pattern was derived from observations. Specifically, our method uses a set of orthogonal patterns to estimate the internal variability, rather than relying on a single pattern. To emphasize large scale patterns of spatial variability we select Laplacian eigenfunctions for this study, calculated using the method of DelSole and Tippett (2015). Previous studies have demonstrated that grid point patterns of SST variability can be used to estimate internal variability in the North Atlantic, in particular, internal AMOC variability by DelSole and Nedza (2021) and basin mean NASST in CESM1 by Liu et al. (2023). However, as discussed by DelSole and Nedza (2021), grid point patterns associated with variability can be highly model specific. Our use of Laplacian eigenfunctions follows DelSole and Nedza (2021), who enhanced the interpretability of their predictive model of AMOC variability, without loss of skill, by using them to represent spatial patterns of SST.

For this introductory study on the proposed dynamical adjustment methodology, our goal is to establish that internal variability in the basin mean NASST can be estimated using a linear regression model using Laplacian eigenfunctions as predictors. In contrast to the observational approach of Wallace et al. (1995), the predictive coefficients for our dynamical adjustment are learned in a multi-model set of preindustrial GCM simulations, which contain only internal variability. Training the dynamical adjustment in a multi-model set of GCMs emphasizes behavior that is shared amongst GCMs, providing a robust basis for estimating internal variability in independent datasets. Training and validation of the dynamical adjustment model is conducted using a Leave-One-Model-Out experimental design, whereby the dynamical adjustment is tested using data that was excluded from the training step. After training and evaluating in preindustrial simulations, the dynamical adjustment is applied to forced historical simulations.

The remainder of this paper is organized as follows: Section 2 discusses the data used and its representation using Laplacian eigenfunctions. Section 3 introduces the dynamical adjustment
Table 1. CMIP5 archived global climate models utilized in this study with the length of preindustrial simulation and number of historical simulations.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Preindustrial Years</th>
<th>Historical Ensemble Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>CanESM2</td>
<td>500</td>
<td>5</td>
</tr>
<tr>
<td>CNRM-CM5</td>
<td>500</td>
<td>10</td>
</tr>
<tr>
<td>CSIRO-Mk3-6-0</td>
<td>500</td>
<td>10</td>
</tr>
<tr>
<td>IPSL-CM5A-LR</td>
<td>500</td>
<td>6</td>
</tr>
<tr>
<td>HadGEM2-ES</td>
<td>500</td>
<td>4</td>
</tr>
<tr>
<td>MPI-ESM-LR</td>
<td>500</td>
<td>3</td>
</tr>
<tr>
<td>MPI-ESM-MR</td>
<td>500</td>
<td>3</td>
</tr>
<tr>
<td>MRI-CGCM3</td>
<td>500</td>
<td>3</td>
</tr>
<tr>
<td>GISS-E2-R</td>
<td>500</td>
<td>6</td>
</tr>
<tr>
<td>CCSM4</td>
<td>500</td>
<td>6</td>
</tr>
<tr>
<td>NorESM1-M</td>
<td>500</td>
<td>3</td>
</tr>
<tr>
<td>GFDL-CM3</td>
<td>500</td>
<td>5</td>
</tr>
</tbody>
</table>

method and how it will be evaluated, as well as the methods that are applied for comparison. Section 4 includes the results of applying the dynamical adjustment to independent preindustrial and historical GCM simulations. Results from the comparative methods and the relative performance of the dynamical adjustment is discussed. The coefficients selected for the dynamical adjustment are examined and a reduced parameter dynamical adjustment approach explored. Section 5 contains concluding statements and thoughts on future research directions.

2. Data

The data used in this study is NASST within the domain of 0-60°N and 10-70°W. Data is analyzed from 12 CMIP5 GCMs (Taylor et al. 2012). These 12 GCMs are selected based on availability of a preindustrial simulation at least 500 years in length and a minimum of 3 historical ensemble members, see Table 1. The last 500 years of each preindustrial simulation is selected. All historical ensemble members are included. This provides a combined multi-model preindustrial data set of length 6000 years and 64 historical ensemble members. Models that do not meet the above criteria were excluded from the analysis. MIROC5 was omitted due to a known data issue.

Data is interpolated onto a common 2.5° by 2.5° uniform grid. Annual means are calculated from July to June to include one winter season in each year. Results are not sensitive to annual mean window. Time series for the results shown are smoothed with a 11 year Lanczos-weighted
running mean (Duchon 1979). An 11 year running mean is selected to emphasize the decadal
to multi-decadal time scales that are of particular interest in NASST, but may smooth through
the response to volcanic impulses. Other widths of smoothing window, including 21 year, 5
year, and 1 year (annual means with no additional smoothing) were also studied. Results for other
smoothing windows show quantitative differences, but produce conclusions similar to those shown.
Conclusions were found to be insensitive to the specific smoothing method employed. Boxcar,
Triangular, and Gaussian weighting schemes were also applied with only minor changes in results
and no change in final conclusions.

Preindustrial control runs are used to train the dynamical adjustment model. A third order
polynomial over the 500-year period is removed to mitigate potential effects of model spin up or
drift. The resulting time series has zero time mean. To remove differences in variability across
GCMs, preventing weighting of the dynamical adjustment model toward high variance GCMs,
each preindustrial simulation is normalized such that the variance of the NASST basin mean has
unit variance. The normalization is performed in such a way to maintain the variance spectrum of
NASST.

Methods for separating forced and internal variability will be tested using historical simulations
from the CMIP5 archive. Historical simulations are initialized from random states of a long GCM
integration and then integrated under realistic estimates of external forcing. As such, historical
simulations contain both forced and internal variability. A GCM’s forced variability is estimated
by the ensemble average of historical simulations from that GCM. For this study, the resulting
ensemble mean defines the ‘true’ external variability. The accuracy of an ensemble mean at
representing a GCM’s true response to external variability is a function of ensemble size and is
related to the magnitude of internal variability (Frankcombe et al. 2018; Milinski et al. 2020).
This dependency on ensemble size motivates the exclusion of GCMs with less than 3 historical
ensemble members. The internal variability is calculated as the residual about the ensemble mean.
The time period of July 1860 to June 2005 is selected from each historical simulation. The time
series are centered for each historical ensemble member individually. The variance of historical
simulations is not normalized. In this study, historical simulations are only used for validation.
The dynamical adjustment model is linear, therefore the scale of the predictions is determined by
the scale of the predictors.
The dynamical adjustment is applied to version 5 of the Extended Reconstructed SST (ERSSTv5), a monthly global analysis compiled from observational data with missing data filled in (Huang et al. 2017b,a). In similar fashion to the GCM data, ERSSTv5 is interpolated to the 2.5° by 2.5° uniform grid and projected onto Laplacian eigenfunctions representing the North Atlantic domain, 0-60°N, 10-70°W. For this study we utilize time steps from July 1860 to June 2018. Annual mean time series are calculated using a July to June window, then centered and smoothed with an 11-year running mean.

a. Laplacian Eigenfunctions

Laplacian eigenfunctions are used as the basis set to represent NASST. Laplacian eigenfunctions are solutions of Laplace’s equation over the North Atlantic domain obtained using the algorithm of DelSole and Tippett (2015). These eigenfunctions depend only on the geometry of the domain and hence are independent of data, thereby providing a common, convenient basis set for all GCMs. The eigenfunctions are spatially orthogonal and ordered from largest to smallest length scale. The characteristic length scale of an eigenfunction, \( \theta_j \), is calculated using

\[
\theta_j = \frac{a \pi}{n_j},
\]

where \( n_j \) is the total wavenumber of eigenfunction \( j \) and \( a \) is the radius of the earth (Laprise 1992). The first eigenfunction is spatially uniform with \( n = 0 \), and therefore \( \theta_1 = \infty \). In plots the length scale is assigned a value of 5000km, the approximate width of the North Atlantic basin. The length scale \( \theta_j \) is utilized primarily for plotting purposes and is not a parameter that appears in the dynamical adjustment method. An alternative version of the dynamical adjustment that incorporates these lengths scales is discussed in Section 4f. The first four Laplacian eigenfunctions over the North Atlantic domain are shown in fig. 1.

The time series for the \( j \)’th Laplacian eigenfunction, denoted \( \psi_j(t) \), is obtained by projecting the area-weighted NASST data onto the \( j \)’th Laplacian eigenfunction (the precise procedure is discussed in DelSole and Tippett 2015). The time series for the first eigenfunction, a uniform pattern over the North Atlantic domain, corresponds to the spatial mean of North Atlantic SST and is a common index of North Atlantic variability. This time series will be referred to as the NASST basin mean, without any insinuation about whether its variability is dominated by

Accepted for publication in Journal of Climate. DOI 10.1175/JCLI-D-23-0651.1.
Fig. 1. Laplacian eigenfunctions 1 through 4 for the North Atlantic domain, 0-60°N and 10-70°W. The first Laplacian eigenfunction is spatially uniform and represents the basin mean. Subsequent eigenfunctions have zero spatial mean, are spatially orthogonal to each other, and have decreasing characteristic length scale. The contour breaks are equally spaced, but the global maximum absolute value is irrelevant in this study and therefore suppressed in the figure.

The original data can be recovered from a sufficiently large set of Laplacian eigenfunctions and their associated time series. Reconstructing data from a truncated set of eigenfunctions will only recover the data down to the minimum length scale of the truncated set. We include the first $J = 100$ eigenfunctions, which corresponds to a minimum characteristic length of around 500km.

The temporal variance of each Laplacian eigenfunction is shown in figure 2 for the preindustrial and historical simulations of each GCM and ERSSTv5. The distribution of variance is very similar
Fig. 2. Temporal variance associated with each NASST Laplacian eigenfunction in normalized preindustrial (solid grey) and un-normalized historical (dashed grey) simulations. The offset between preindustrial and historical variances is due to the preindustrial variance distribution being normalized such that the spatially uniform pattern has unit variance. No normalization is applied to the historical data. The arithmetic average across the 12 GCMs’ historical simulations’ is plotted as the lowest black line. The blue line represents the average ensemble mean variance across GCMs. The red line is the average variance of the residuals about the ensemble mean. The green line plots temporal variance of Laplacian eigenfunctions from ERSSTv5. A $k^{-3}$ line is plotted, offset from the variance distribution, as a visual guide and $g(\theta)$, defined in eqn. (10), an empirical function approximating the shape of variance distribution, is plotted for reference, (following DelSole and Nedza 2021). Vertical lines correspond to Laplacian eigenvectors 2, 3, and 4 for visual identification.

Accepted for publication in Journal of Climate. DOI 10.1175/JCLI-D-23-0651.1.
between preindustrial and historical simulations for most spatial scales. However, for the spatially uniform pattern, the first Laplacian eigenfunction, there tends to be much more variance in the historical simulations compared to the preindustrial. This difference is attributed to externally forced variability. The temporal variance approximately follows a $k^{-3}$ distribution across most spatial scales (exempting the largest), as reported in DelSole and Nedza (2021). Temporal variance for Laplacian eigenfunctions derived from ERSSTv5 is plotted in figure 2. The distribution of variance in ERSSTv5 falls within the range of historical simulations, and generally agrees with the average across GCMs. However, we note that the variance of smaller spatial scales in ERSSTv5 does appear slightly lower than for most GCMs. This may be due to the reconstruction method, which employs a finite number of spatially smoothed empirical orthogonal teleconnection patterns (Huang et al. 2017a).

3. Methods

a. Dynamical Adjustment

The dynamical adjustment takes the form of the linear regression model:

$$Y_{1 Internal} = X_{2:100} \beta + \beta_0 + \epsilon. \quad (1)$$

The predictand, $Y_{1 Internal}$, is the internal component of the basin mean NASST, the first Laplacian eigenfunction. The predictors, $X_{2:100}$, are the time series associated Laplacian eigenfunctions 2-100 (the first is omitted from the predictor set since it is the quantity being predicted). The coefficient vector, $\beta$, relates anomalies in the predictor time series to anomalies in the basin mean. $\beta_0$ represents the intercept term and $\epsilon$ the error of the regression model, representing internal variability not captured by the estimated coefficients.

The coefficients are learned in a multi-model set of preindustrial simulations. Preindustrial simulations have no interannual variations in external forcings and do not contain externally forced variability; the coefficients only characterize relationships that occur in internal variability. Training in a multi-model ensemble, the dynamical adjustment is intended to learn relationships that are shared across GCMs. The skill of the dynamical adjustment is evaluated on preindustrial and historical simulations from an independent GCM that is excluded from the training data. This
process is repeated for all 12 GCMs in this study, with a new set of coefficients learned each time. Details on the coefficients are discussed in Section 4f.

The dynamical adjustment can easily be modified to use leading or lagging predictors, or a combination of simultaneous, leading, and/or lagging predictors. For these experiments, a separate set of dynamical adjustment coefficients are learned under the same experimental outline described above. Results are discussed briefly in Section 4a.

**Regularized Regression**

The predictor time series are highly co-varying which can contribute to overfitting in an ordinary least squares (OLS) framework. To mitigate potential issues due to covarying predictors and overfitting, regularized regression is applied when learning the dynamical adjustment coefficients. Regularized regression includes a penalty term when computing estimates of the coefficients, \( \hat{\beta} \), as

\[
\hat{\beta} = \arg\min_{\beta} \left( \left( \psi_1(t) - \beta_0 - \sum_{j \neq 1}^J \psi_j(t) \beta_j \right)^2 + \lambda R(\beta) \right),
\]

where \( \hat{\beta} \) is a vector containing the estimated regression coefficients, \( \psi_j(t) \) are the time series associated with Laplacian eigenfunction \( j \), and the angled brackets represent a mean across all time steps and training GCMs. The penalty term consists of \( \lambda \), the regularization parameter, and \( R(\beta) \), a norm of \( \beta \). Standard choices of this norm include

\[
R(\beta) = \sum_{j \neq 1}^J |\beta_j| \quad \text{Lasso}
\]

\[
R(\beta) = \sum_{j \neq 1}^J \beta_j^2 \quad \text{Ridge}
\]

\[
R(\beta) = \sum_{j \neq 1}^J \left( \alpha |\beta_j| + (1 - \alpha) \beta_j^2 \right) \quad \text{Elastic Net},
\]

where \( \alpha \) is a hyper parameter that controls the application of the elastic net regularization. The elastic net results investigated (not shown) utilized \( \alpha = 0.5 \).
Fig. 3. Mean-squared error (red dotted curve) and standard errors (error bars) of the estimated basin mean NASST as produced from cv.glmnet output. These values are estimated across a training set of 11 preindustrial simulations using a cross validation procedure that excludes one of the training GCMs at a time. This plot corresponds to the preindustrial training sample that excludes CanESM2. The $\lambda$ value that produces the minimum MSE is indicated by the left vertical dotted line. The largest $\lambda$ value that produces a MSE within one standard error of the minimum is indicated by the right vertical dotted line. The numbers along the top margin indicate the number of Laplacian eigenfunctions assigned non-zero coefficients in the lasso model.
The penalty term penalizes based upon the chosen norm of the coefficients. Each norm has a different impact upon coefficients. The norm associated with lasso has the effect of setting some coefficients exactly to zero. Ridge, in contrast, tends to shrink the entire set of coefficients without setting any to zero, while the elastic net norm has an effect of combining the two. We refer the reader to Tibshirani (1996) for more detail on regularized regression. The regularized regression is implemented using the R package, `cv.glmnet` (Friedman et al. 2010). All three norms above were applied with none of them performing systematically better than the others. Therefore, results discussed in this study will be limited to lasso. Lasso is chosen because of the norm’s characteristic behavior of setting some coefficients to zero. By reducing the number of non-zero coefficients, we hope to identify spatial patterns that are most important for prediction. However, due to the highly covarying nature of the NASST Laplacian eigenfunctions, interpreting these spatial patterns may be difficult and similar predictions may be possible even when excluding some predictors.

The regularization parameter, $\lambda$, must be selected for each lasso model. $\lambda$ is selected through a cross validation procedure performed on the training data. For each training set of 11 models, cross validation is performed by training on 10 models and validating upon the excluded model. This procedure is repeated 11 times for each training set, excluding data from a different model each time. The distributions of error for each $\lambda$ value is estimated under this cross validation procedure, an example of which is shown in figure 3. This example corresponds to the training set that excludes CanESM2, including the 11 other models. The cross validation distributions are similar for all training sets.

There are two common criteria for selecting $\lambda$. One criteria is to select the $\lambda$ that minimizes the cross-validated mean square error. The other is the 'one-standard-error' rule, selecting the largest $\lambda$ whose mean square error lies within the standard error range of the minimum. Selecting the 'one-standard-error' $\lambda$ value has the effect of setting more predictor coefficients to zero compared to the minimum $\lambda$ selection. Results are not sensitive to 'one-standard-error' or minimum $\lambda$ value selection. Results discussed in this study correspond to the 'one-standard-error' $\lambda$ selection. A brief discussion of the coefficients selected by the lasso model is included in Section 4f.

b. Comparative Methods

In this section, we discuss methods that will be used for comparative purposes.
1) POLYNOMIAL METHOD

The polynomial method estimates the basin mean NASST external variability by fitting a polynomial to each historical ensemble member individually. In the simplest case, the polynomial is a linear trend. However, the linear assumption has little physical support and tends to perform poorly compared to other methods, motivating the use of a quadratic trend by Frankcombe et al. (2015) and Frankignoul et al. (2017). Hawkins and Sutton (2009) applied a fourth order polynomial. We explore application of polynomials to tenth order, systematically evaluating each order and assessing performance. In this study, the polynomial method is applied to time series of the same length and therefore a given polynomial order has the same capacity for capturing variability in each GCM simulation.

2) LOCALLY WEIGHTED SCATTERPLOT SMOOTHING

Locally weighted scatterplot smoothing, LOWESS, has been suggested as a filter for removing time scales of variability associated with most internally generated modes of oceanic variability (Cheng et al. 2022). LOWESS generates a smoothed time series by taking the value of a polynomial fit at each time step. The value of this polynomial is determined using a weighted regression including surrounding points within a prescribed span (Cleveland 1979). Cheng et al. (2022) suggests that this method is much more successful at isolating low frequency variability, particularly that due to anthropogenic forcing, in observed global mean oceanic heat content compared to a polynomial fit.

Cheng et al. (2022) suggests the use of a 25 year LOWESS span for filtering out time scales associated with most internally generated modes of oceanic variability. We investigated a variety of spans and found a 25 year LOWESS span to perform well across GCMs. Increasing the span width does not improve performance in all GCMs. Local linear fitting demonstrates an improvement over local constant fitting, particularly in the vicinity of end points. Applying a local quadratic fit does not produce a more accurate separation of variability compared to the local linear fit. Results shown correspond to a local linear fit with a 25 year LOWESS span unless otherwise noted.
3) **ENSEMBLE EMPIRICAL MODE DECOMPOSITION**

Empirical Mode Decomposition, EMD, decomposes a given time series, \(x(t)\), into intrinsic mode components, IMFs, of increasing time scales, \(c_j(t)\), and a residual, \(r_n(t)\) (Huang et al. 1998).

\[
x(t) = \sum_{j=1}^{n} c_j(t) + r_n(t),
\]

(2)

EMD is an iterative process of fitting an upper and lower envelope of the data using cubic splines connecting the local maxima and minima of \(x(t)\). The local mean of the upper and lower envelopes is calculated, constituting the first IMF component, \(c_1(t)\). This component is then removed from the data and the process repeated on the residual, \(r_1(t)\), with this process iterated until the final residual, \(r_n(t)\), is monotonic or contains at most one internal extrema. For the complete process and considerations of this method, we refer the reader to (Huang et al. 1998; Huang and Wu 2008).

Ensemble Empirical Mode Decomposition, EEMD, extends from EMD, and produces more robust results less sensitive to the presence of noise. EEMD modifies the above process by adding white noise to the original time series (Wu and Huang 2009). Many independent realizations of white noise are added to the original data and then each decomposed using EMD, producing an ensemble for each of the IMFs as well as the residual. The ensemble for each component is then averaged, removing the influence of the independent realizations of white noise, and producing a more robust estimate of the IMFs than simply applying EMD to the original time series.

EEMD has previously been applied to the study of low frequency components in temperature time series, including the identification of a non-linear secular trend and multi-decadal component that contributes non-trivially to the overall increase in observed global mean temperatures (Wu et al. 2011), and as part of a methodological comparison when estimating the separation of temperature variability into externally forced and internal components (Frankignoul et al. 2017).

In this study, EEMD is implemented using the R package `hht()` produced by Bowman and Lees (2013) with 400 ensemble members and white noise with one-fifth the standard deviation of the data being decomposed, parameter choices consistent with Ji et al. (2014) and Frankignoul et al. (2017). Results shown represent the performance using the residual, \(r_{n-2}(t)\), equivalent to the summation of the final residual term, \(r_n(t)\), and the two lowest frequency IMFs, \(c_n(t)\) and \(c_{n-1}(t)\), to represent the externally forced variability. This summation performs well on average across
GCMs when compared to estimating the external component of variability using the residual alone or summed with any number of successive IMFs.

4) Multi-Model Ensemble Mean

For the multi-model ensemble mean (MMEM) method, an MMEM is calculated from all GCMs except the GCM in which the method is being evaluated. More precisely, one GCM is withheld, an ensemble mean is computed for each of the 11 remaining GCMs, and then the mean of the 11 ensemble means is used for the MMEM. We emphasize that instead of computing a mean over all historical runs, the MMEM is a mean of ensemble means, which equally weights each GCM. The validation GCM is excluded from the MMEM such that no information from that GCM is included. The MMEM is either subtracted or regressed from each ensemble member of the validation GCM. The regression step rescales the MMEM to correct for differences in magnitude of externally forced response. These approaches are referred to as ‘differenced’ or ‘scaled’ MMEM methods as described by Frankcombe et al. (2015).

GCMs within this study exhibit a range of responses to external forcing, likely related to individual GCMs’ aerosol effective radiative forcing (Rotstayn et al. 2015). The MMEM represents the average behavior across GCMs and will not represent responses unique to a single or subset of GCMs. Although a MMEM could be computed from a subset of GCMs, the question arises as to which subset should be selected for application to independent data. There is no consensus on the best criterion for selecting GCMs in this situation. We proceed under the premise that each GCM is equally likely to be representative of the real system and therefore do not group or weight the GCMs when calculating the MMEM.

c. Validation Measure

The skill of methods in this study will be evaluated using normalized mean squared error (NMSE). Let $\psi_1^{PI}$ denote the basin mean NASST time series for a preindustrial simulation and $I(\psi_1^{PI})$ denote a prediction function of internal variability. The error of the prediction model at each time step is defined as,

$$e = \psi_1^{PI} - I(\psi_1^{PI}).$$  \hspace{1cm} (3)
The mean-squared error of $I(\psi^P_1)$ is defined to be,

$$\text{MSE} = \langle (e)^2 \rangle,$$  \hspace{1cm} \hspace{1cm} (4)

where the angle brackets denote the mean across all time steps in the validation GCM. The normalized mean-squared-error is then defined as,

$$\text{NMSE} = \frac{\text{MSE}}{\langle (\psi^P_1)^2 \rangle}.$$  \hspace{1cm} \hspace{1cm} (5)

For a perfect prediction, $I(\psi^P_1) = \psi^P_1$, MSE = 0, and NMSE = 0. For a prediction based on the climatological mean, in this case, $I(\psi^P_1) = 0$, MSE = $\langle (\psi^P_1)^2 \rangle$ and NMSE = 1. NMSE = 1 is considered the threshold for skill with NMSE ≥ 1 considered to represent no skill. Conveniently, NMSE can be converted to the fraction of explained variability by 1 - NMSE. The standard error of MSE is,

$$\text{SE}_{\text{MSE}} = \sqrt{\frac{\langle (e^2 - \text{MSE})^2 \rangle}{(N-1)}},$$  \hspace{1cm} \hspace{1cm} (6)

where $N$ is the length of the validation data. SE_{MSE} can then be normalized to produce the normalized standard error, SE_{NMSE}, that applies to NMSE,

$$\text{SE}_{\text{NMSE}} = \frac{\text{SE}_{\text{MSE}}}{\langle (\psi^P_1)^2 \rangle}.$$  \hspace{1cm} \hspace{1cm} (7)

The estimate of internal variability can also be evaluated by calculating the correlation coefficient, defined as:

$$R_{\psi^P_1, I(\psi^P_1)} = \frac{\text{cov}[\psi^P_1, I(\psi^P_1)]}{\sqrt{\text{var}[\psi^P_1] \text{var}[I(\psi^P_1)]}}.$$  \hspace{1cm} \hspace{1cm} (8)

The fraction of variability in $\psi^P_1$ that $I(\psi^P_1)$ represents, after rescaling as $\psi^P_1 = \beta I(\psi^P_1) + c$ equals the correlation coefficient squared, $R^2$. Importantly, it is noted that correlation forgives amplitude errors, up to a single scaling factor, in the estimate of internal variability.
HISTORICAL SIMULATIONS

For historical simulations, validating a method is not as straightforward because the simulations contain both forced and internal variability. If $\psi_1^H$ represents the time series of the basin mean NASST in a historical simulation, it can be decomposed into a forced, $f_1$, and internal, $i_1$, component:

$$\psi_1^H = f_1 + i_1.$$  \hspace{1cm} (9)

In this study, $f_1$ is estimated by each GCM’s ensemble mean and $i_1$ by each ensemble member’s residual about the ensemble mean. To calculate the NMSE and $\text{SE}_{\text{NMSE}}$ of the dynamical adjustment, equations 3 through 7 are still appropriate with $\psi_1^{PI}$ replaced with $i_1$ and $I(\psi_1^{PI})$ replaced by $I(i_1)$.

To compare dynamical adjustment to other methods in historical simulations, it should be recognized that the methods for separating forced and internal variability predict different things. Specifically, dynamical adjustment predicts the internal component of variability whereas the other methods, the polynomial and MMEM, predict the forced component. A prediction of internal variability, $I(i_1)$, can be transformed into a prediction for the forced component by computing the residual $\psi_1^H - I(i_1)$. Here, we emphasize that any error in the estimate of internal variability, $I(i_1)$, will be included in the residual estimate of forced variability. Similarly, any prediction of forced variability can be transformed into a prediction of internal variability. For this case, any error in the prediction of forced variability will be included in the residual prediction of internal variability. Employing such transformations, all predictions can be expressed uniformly either as predictions of internal variability or as predictions of forced variability, regardless of the original form of the predictions. Importantly, the mean square error of the prediction is the same for both representations: the error of a prediction of internal variability, $(\psi_1^H - f_1) - I(i_1)$, is the negative of the error of a prediction of forced variability, $f_1 - (\psi_1^H - I(i_1))$. However, $\text{NMSE}$ can differ between the two representations due to differences in the target, which determines the appropriate normalization constant. Specifically, if the target is internal variability, then the appropriate normalization constant is the variance of internal variability ($i_1$), whereas if the target is forced variability, the appropriate normalization constant is the variance of the forced part ($f_1$).
The resulting NMSEs differ by a scaling factor equal to the ratio of variances of forced and internal variability.

Depending on the choice of normalizations, a given prediction may be deemed skillful under one normalization but not the other (i.e., the prediction may have NMSE < 1 under one normalization and NMSE > 1 under the other). We propose that the simplest way to proceed is to evaluate both skill measures, and define skill as the prediction producing a NMSE less than unity under both normalizations. For this study, normalizing by internal variability produces a more strict threshold of skill. This is a consequence of the fact that forced variance in the historical runs, with an 11-year running mean applied, generally exceeds or is equal to the internal variance in the GCMs in this study. Accordingly, we opt to normalize by internal variability, in which case a prediction that is skillful under this measure is skillful under both measures. This choice has the further advantage that it enables skill comparisons between historical and control simulations; normalizing by forced variance is not an option for control simulations because the variance of forced variability vanishes for these runs.

4. Results

a. Preindustrial Validation

We now show the result of applying the dynamical adjustment to estimate variability in independent preindustrial simulations. Figure 4 shows the preindustrial NASST basin mean time series for each GCM along with the dynamical adjustment estimate. Explained variability calculated using NMSE ranges from a low of 27% (NorESM1-M) to a high of 67% (CSIRO-Mk3-6-0, IPSL-CM5A-LR). We interpret this range of explained variability as the dynamical adjustment being unable to represent the diversity of GCMs’ internal variability. We carried out experiments that involved training and validating both within individual GCMs and across multiple GCMs. However, the outcomes of single-model trained dynamical adjustment were highly model-specific and presented challenges in summarizing them cohesively without a clear, overarching hypothesis. Consequently, we opted to narrow our focus to the multi-model trained dynamical adjustment. This focus is supported by better performance of the multi-model trained dynamical adjustment compared to single-model trained dynamical adjustment in almost every independent validation (not shown).
Fig. 4. NASST basin mean anomaly time series of the last 500 years of each GCM’s preindustrial simulation (black). Each time series has a third order polynomial removed and has been normalized to unit variance. The dynamical adjustment estimate is plotted in blue. Time series are labeled by their respective model along the right vertical axis. Each model name is followed by \(R^2\), 1-NMSE), which quantifies the fraction of variance explained by dynamical adjustment.

Explained variability calculated using correlation and NMSE are similar for all GCMs, with the explained variability estimated by \(R^2\) being equal or higher due to the metric’s forgiveness.
of amplitude errors. The influence of this amplitude correction is clear in HadGEM3-ES, where the explained variability between metrics differs by 14% due to consistent underestimation of amplitude.

In GCMs where the dynamical adjustment estimate performs well, for example CSIRO-Mk3-6-0 and IPSL-CM5A-LR, low frequency variability is well estimated, although there is a tendency to underestimate peaks and amplitude. In GCMs where the dynamical estimate performs poorly, for example CCSM4 and NorESM1-M, some low frequency variability is estimated correctly with other fluctuations being misrepresented in direction. The accuracy of higher frequency variations depends on GCM. The \( \text{NMSE} \pm \text{SE}_{\text{NMSE}} \) of applying dynamical adjustment to independent preindustrial simulations are summarized in figure 6. The dynamical adjustment is skillful, producing a \( \text{NMSE} \leq 1 \), in every GCM.

In separate analysis, the dynamical adjustment approach was applied using predictors that lead the predictand by 1, 2, and 11 years (not shown). For 1 and 2 year leading predictors, the dynamical adjustment performs similarly to the simultaneous predictors, with a small degradation of skill based on lead time. For the 11 year leading predictors, where the predictors and predictand have non-overlapping running mean windows, the \( \text{NMSE} \pm \text{SE}_{\text{NMSE}} \) bars include the value of one for the majority of GCMs, implying that the dynamical adjustment has no skill. Due to their similar display of skill, or lack of skill for 11 year leading predictors, dynamical adjustment predictions using only leading predictors will not be discussed for application on historical simulations in this study.

In separate analysis, the dynamical adjustment was applied using a combination of simultaneous and 1, 2, or 11 year leading predictors (not shown). Using any combination of predictors performs as well as simultaneous predictors only. This is not surprising as the 1 and 2 year leading predictors are unlikely to contribute much additional information compared to the simultaneous only due to the lead being well within the much wider window of smoothing (11 year). Due to the similar performance, using combinations of predictors will not be discussed further.

b. Historical Validation

The dynamical adjustment, trained in preindustrial simulations, is applied to the historical simulations of the GCM withheld from the multi-model training set. No modification to the
Fig. 5. Ensemble mean of NASST basin mean corresponding to 1860-2005 for each GCM (black). Estimates of external variability for the first two ensemble members from each GCM, including the dynamically adjusted time series (blue), scaled MMEM (green), and dynamically adjusted time series when removing the ensemble mean from the predictors (red) are plotted. The time series are not rescaled, y-axes are not consistent for each GCM.

predictor time series from the historical simulations is performed and the dynamical adjustment estimate is not influenced by the validation GCM’s ensemble size. No re-tuning of the dynamical
Fig. 6. Normalized mean squared error (NMSE) ± the associated standard error (SE_{NMSE}) of the dynamical adjustment applied on independent preindustrial simulations (black), historical simulations (blue), and historical simulations when removing the ensemble mean from predictor time series (red). Dynamical adjustment is plotted as pairs of bars of the same color, with the left bar representing the performance of the full lasso model provided with 99 Laplacian eigenfunctions, and the right bar, denoted with an ‘x’ at the center, representing a reduced regression model with Laplacian eigenfunctions 2, 3, and 4 as predictors. NMSE ± SE_{NMSE} is shown for LOWESS (brown), and scaled MMEM (green) applied to historical simulations. Results are organized in columns of validation model, named along the bottom axis, with the historical ensemble size in parenthesis.
adjustment model is performed. The dynamical adjustment estimate of external variability in basin mean NASST, obtained by subtracting the predicted internal component from the total, for the first two ensemble members of each GCM is shown in figure 5. The estimate of external variability from dynamical adjustment generally fits the shape of the true external variability, appearing to effectively account for the presence of decadal to multi-decadal internal variability in most cases. However, dynamical adjustment sometimes fails to correctly or completely remove decadal internal variability. This is expected based on the preindustrial results shown in figure 4 where the phase or magnitude of variability is not always perfectly represented. Some examples of this include CSIRO-Mk3-6-0, NorESM1-M, and GFDL-CM3. In each of these cases, for at least one of the plotted estimates, the dynamical adjustment has clear errors in the magnitude of decadal variability in the external estimate, representing a failure to properly account for the internal contribution during that time. The NMSE ± SE of the dynamical adjustment in historical simulations is summarized in figure 6. For 9 models, the NMSE is below one indicating a skillful prediction. For 3 GCMs, the NMSE is above one. For these 3 GCMs, the dynamical adjustment is not considered skillful. The estimate in each historical ensemble has a higher NMSE, and therefore less skill, than the preindustrial validation for the same GCM.

The application of dynamical adjustment to historical simulations assumes that over the North Atlantic, the externally forced response is confined to the spatially uniform pattern and does not project onto the other Laplacian eigenfunctions. This assumption is unlikely to be accurate, particularly considering the presence of non-spatially uniform aerosol forcing in the North Atlantic region and the observed warming hole. Consequently, the time series associated with Laplacian eigenfunctions other than the spatially uniform pattern likely contain a combination of internal and external variability. The dynamical adjustment model is linear, therefore the estimate of basin mean NASST in historical simulations could be separated into two components, one based on the internal variability in the predictor time series and one based on the external variability. However, there is no basis for the external variability-based component of the dynamical adjustment estimate to represent the internal variability in basin mean NASST. This suggests that the degradation of skill when applying the dynamical adjustment to historical simulations is due to the presence of external variability in the predictor time series, contrary to the original premise that predictors be pure internal variability. This suggests that the degradation of skill when applying the dynamical
adjustment to historical simulations is due to the presence of external variability in the predictor time series, contrary to the idea that predictors be pure internal variability. To investigate this hypothesis further, we examine the components of variance in each Laplacian eigenfunction, which can be done in an ensemble framework. Figure 2 plots the total variance of historical simulations, averaged across the 12 GCMs, for each Laplacian eigenfunction. This total variance is decomposed into contributions from the externally forced component, represented by the historical ensemble means, and the internal variability, calculated using the residuals about the ensemble means and corrected to account for the finite ensemble sizes. For the first Laplacian eigenfunction, the basin mean, the contribution of external variability dominates that of internal variability. For all other Laplacian eigenfunctions, the internal variability dominates the total. This demonstrates that on average the external variability is nearly an order of magnitude smaller than internal variability in the predictor time series, supporting an essential assumption of dynamical adjustment. However, considering the average across all GCMs disguises important GCM differences. Separate analysis reveals that external variability is present in spatial patterns other than the basin mean, but the particular spatial patterns in which this occurs and the magnitude of the external contribution is GCM dependent (not shown). GCM-to-GCM differences in the presence of externally forced variability could be related to varying representations of aerosol forcing response (Rotstayn et al. 2015) or a change of the physical system in response to one or several external forcings. A complete review of GCM-to-GCM differences for each of the 100 Laplacian eigenfunctions and attribution to individual mechanisms is beyond the scope of this study and therefore is not discussed further here. Nevertheless, these results confirm the presence of external variability in patterns orthogonal to the uniform and it is plausible that this presence degrades the dynamical adjustment’s estimate of internal variability.

To test if the degradation in skill in historical simulations is due to the presence of forced variability, forced variability is removed from each predictor time series by subtracting that GCM’s ensemble mean calculated for each predictor prior to applying the dynamical adjustment. These results can be compared to the previous application of dynamical adjustment. The dynamical adjustment model is not retrained or tuned for this application. The dynamical adjustment estimate of basin mean NASST external variability when removing external variability in the predictor time series is plotted in figure 5. In general, this approach appears as a small improvement over the
base dynamical adjustment. The NMSE of this approach can be compared in figure 6, where the results are summarized. For every GCM, the NMSE for this application of dynamical adjustment lie below the previous results and below one. We conclude that removing the ensemble mean from the predictor time series when applying dynamical adjustment improves the skill in every GCM, even in the three models for which dynamical adjustment had previously not demonstrated skill. For 8 GCMs, these results overlap with the preindustrial results. For these GCMs, the ability for the dynamical adjustment to estimate internal variability in the basin mean NASST is similar to the performance in preindustrial simulations. The near recovery of preindustrial levels of skill is suggestive of some level of consistency in the internal variability between preindustrial and historical simulations as well as the additive nature of variability. Further, this suggests that external variability in the predictors was responsible for the degradation between preindustrial and historical validations. For the other 4 GCMs, although removing external variability from the predictors does improve the performance in historical simulations, the performance does not reach the preindustrial validation. This suggests that there may be another source of error in these historical simulations. We have been unable to identify the source of these discrepancies in this study.

c. Results of the Comparative Methods

The polynomial method is applied to individual ensemble members to estimate the externally forced variability in basin mean NASST. Polynomial orders are applied from one, a linear trend, up to tenth order. The associated NMSE are shown for orders one through eight in figure 7. In each column, the skill of each polynomial order is plotted with increasing order from left to right. That is, the left most black bar represents the NMSE of a first order polynomial, the second, a red bar, represents the second order polynomial and so on. In general, the performance of the polynomial method improves with increasing order, plateauing between order 4 and 6. Few of the polynomials produce a NMSE less than one indicating that the estimate of internal variability based on a residual from any order polynomial tends to have no skill. The best polynomial order depends on the GCM and may be related to timescales of internal variability and the relative magnitude of forced and internal variability.
Results for the application of LOWESS and EEMD are also shown in figure 7. LOWESS outperforms the standard polynomial and EEMD in this application, demonstrating skill in all but two GCMs. Compared to the polynomial and EEMD methods, LOWESS performs as well as or better in all GCMs. EEMD performs similarly to polynomials across GCMs. In general, the polynomial, LOWESS, and EEMD methods suffer from the same underlying issue. Each of these methods, in application, separate different timescales of variability. Separating timescales can be very effective at identifying external and internal components of variability if these components have distinct timescales. In NASST, there is considerable debate over the relative contribution of external and internal variability on decadal to multi-decadal timescales. Methods that only rely on time scale separation when identifying these components may therefore find it difficult to correctly identify relative contributions.

d. Results of Ensemble Mean Method

The scaled MMEM estimate of external variability is plotted in figure 5 for the first two ensemble members of each GCM. The application of a single scaling step corrects for amplitude across the entire time period and produces a good approximation of the externally forced variability. However, externally forced variations within the study period, on decadal to multi-decadal time scales, are not always well captured by the scaled MMEM. For example, in the MPI models the MMEM overestimates the decadal fluctuations of the externally forced variability. In contrast, the MMEM underestimates such variability in CSIRO-Mk3-6-0 and GFDL-CM3. The MMEM may be fundamentally incapable of accurately capturing external variability on decadal time scales for every GCM. The MMEM is constructed using many GCMs, averaging the magnitude and timing of externally forced decadal variability. As a consequence, a single MMEM is not able to capture the diversity of GCM responses, particularly due to the large range of aerosol radiative forcing (Rotstayn et al. 2015). The NMSE of using a scaled MMEM to separate variability in basin mean NASST is shown in figure 6. The scaled MMEM is skillful in 7 out of 12 models. In the other 5 models, the scaled MMEM lacks skill, with a NMSE greater than or equal to 1. We do not show the differenced MMEM approach because it is skillful in only 3 models and is generally outperformed by the scaled approach. This improvement of the scaling approach over the differenced agrees with previous studies (Frankcombe et al. 2015).
**Fig. 7.** NMSE ± SE\(_{\text{NMSE}}\) of polynomials applied to basin mean NASST for each historical ensemble member. Results are organized by column for each GCM, named along the bottom axis with the historical ensemble size in parenthesis. The first order polynomial, a linear trend, is shown as the black, left most, bar in each column. The second order polynomial is shown as the red, second from left, bar in each column. Subsequent polynomial orders, up to order 8, are plotted in each column from left to right. NMSE for LOWESS (brown) and EEMD (pink) are also shown.
Having discussed each methods’ results, we compare the dynamical adjustment, applied to historical simulations without modification, to other methods here. The NMSE of each method is shown in figure 6, where skill between two methods is considered indistinguishable if the standard error bars overlap. A method is considered the best if it has the lowest NMSE and does not overlap with any other method. Comparing NMSE demonstrates that no method is consistently best across the GCMs used in this study. The dynamical adjustment is the best performing method in 4 of 12 GCMs and ties a comparative method in 2 additional GCMs. We interpret this to demonstrate that dynamical adjustment, as proposed in this paper, has similar potential as other methods when applied to separating basin mean NASST variability. This conclusion differs only slightly when different filtering schemes are used. For example, using Boxcar filter weights, the dynamical adjustment is the best performing method in 5 GCMs, and ties a comparative method in 2 GCMs.

Applying dynamical adjustment to predictors after removing the ensemble mean gives the best performance in 8 of 12 GCMs. In 3 of the remaining models, this application of dynamical adjustment is tied with a comparative method for best performance. In the remaining model, the scaled MMEM is the best method. This relative comparison may differ slightly using different filtering schemes. Using Boxcar filter weights, this modified application of dynamical adjustment gives the best performance in 10 of 12 GCMs, with the scaled MMEM being the more skillful in the two remaining GCMs. This modified application of dynamical adjustment appears as the best overall method, however, it cannot be applied to observations as the true external variability in observed predictors is not known.

To gain insight into the how the dynamical adjustment works, we examine the coefficients selected by lasso. These coefficients are shown as the bottom set of lines in figure 8. Each line corresponds to a different set of preindustrial training data, labeled by which GCM was excluded from the training. The coefficients learned from the different training samples are very consistent, with predictor selection and coefficient values being particularly consistent for lower order Laplacian eigenfunctions, which have the largest characteristic length scales. This consistency is expected due to the fact that the training sets are nearly the same; two training sets contain 10 of the same
Fig. 8. $\hat{\beta}$ for the dynamical adjustment trained in multi-model sets of preindustrial simulations corresponding to the ‘one-standard-error’ $\lambda$ selection. $\hat{\beta}$ are labelled by the GCM that was excluded from the training set. The lower set of coefficients are those used to produce the results shown in this study and are trained using lasso regularization. The upper set of coefficients, offset by a value of 1, are trained on the same data, using lasso regularization that incorporates a length scale based weighting. The short black lines represent the coefficients for the reduced regression model, offset by a value of 1. The green dots are the average coefficients across reduced regression models. The coefficients are organized using a logarithmic x-axis that emphasizes the Laplacian eigenfunctions with the largest length scales.
GCMs and only differ by the 11th. Higher order eigenfunctions appear in some iterations of the dynamical adjustment lasso model, but the predictor selection does not show the same robustness as for lower order eigenfunctions.

In general, the largest spatial scales have one hundred times more temporal variance than the smallest spatial scales (see fig. 2). This leads us to question if the small-scale predictors are really important to dynamical adjustment. The software package \texttt{cv.glmnet} standardizes the predictor time series when training the dynamical adjustment lasso model. This means that all predictors are penalized the same and no information about the length scales or variance distribution is utilized when selecting coefficient values. To emphasize large scale Laplacian predictors, we modify lasso to penalize small-scale predictors more strongly than large-scale predictors. To do this, we follow DelSole and Nedza (2021) and apply a penalization of the form, \( w_j = 1/g(\theta_j) \), where

\[
g(\theta_j) = \frac{8}{(4000/\theta_j)^3 + 7}.
\]

and \( \theta_j \) is the length scale of the \( j^{\text{th}} \) Laplacian eigenfunction. \( g(\theta) \) approximates the shape of the variance distribution of NASST (plotted in fig. 2).

The scale selective penalization has only a small effect on the NMSE of the dynamical adjustment (not shown). Overall, the NMSE is very similar with some validation GCMs showing a small improvement and others a small degradation, with no change in conclusions. The scale selective penalization does influence the coefficient selection, shown as the upper set of lines in figure 8, with a marked elimination of higher order Laplacian eigenfunctions assigned non-zero coefficients.

The skillful performance at estimating internal variability with the scale-weighted lasso model suggests that predictive skill can be mostly attributed to large scale patterns. Skill does not depend on smaller scale features that may be more GCM dependent. Although the length scale weighting has removed small length scale predictors without a loss in predictive skill, the number of non-zero coefficients continues to make interpretation of the model difficult. To address the possibility of a skillful prediction using a very reduced set of Laplacian eigenfunctions, we create a simple regression model with only Laplacian eigenfunctions 2, 3, and 4 as predictors (those shown in figure 1), representing the largest scale patterns orthogonal to the uniform pattern. The coefficients for this regression model, trained on the same preindustrial data, are shown in figure 8 by the set
of black lines overlaid the scale-selective dynamical adjustment coefficients. Similar to the lasso models, we see that the coefficients are consistent across the different training sets. However, the magnitude of the coefficients are slightly different than the lasso models. Using coefficients averaged across the reduced regression models, a single reduced regression model can be written as

$$\hat{Y}_{Internal}^1 = -0.62\psi_2 - 0.63\psi_3 - 0.27\psi_4.$$  \hspace{1cm} (11)

The NMSE of these simple regression models is included in figure 6. For some GCMs, in both preindustrial, historical, and historical when removing the ensemble mean from the predictor time series, this reduced regression model performs about as well as the original lasso model. For other GCMs, this reduced regression model does not perform as well as the full lasso model. The fact that much of the predictive skill can be produced using a small predictor set of the largest spatial patterns of NASST variability may be useful in constraining future work.

**g. Application to Observations**

We now examine the result of applying the above methods to ERSSTv5. We emphasize that the true external and internal variability is not known in observational data, and therefore these results should be interpreted cautiously.

The dynamical adjustment is applied to ERSSTv5 data without modification. As discussed in section 4f, the dynamical adjustment models trained on the different sets of 11 GCMs are very consistent. The estimated internal variability plotted in figure 9b represents the averaged coefficients amongst these dynamical adjustment models. The reduced regression model, equation 11, is also applied to ERSSTv5. The dynamical adjustment estimates of external variability, calculated by subtracting the estimate of internal variability from the total, are plotted in figure 9a. Polynomials, LOWESS, EEMD, and MMEM methods as described previously are also applied to ERSSTv5 data, directly estimating the external component of variability. A selection of these external estimates are plotted in figure 9a, and their residual estimates of internal variability in figure 9b. The MMEM is applied only to time steps that overlap between CMIP5 and ERSSTv5 and does not cover the entire time series. Correlations are calculated between the different internal variability estimates and listed in table 2. These correlations illustrate that estimates of internal variability vary by method and parameter choices. The dynamical adjustment estimate of internal variability tends to
Fig. 9. Estimates of external and internal components of ERSSTv5 basin mean NASST. The top panel shows results of estimating external variability using LOWESS with a 25 year (brown) and 50 year (purple) span; EEMD with the residual summed with one (gold) and two (pink) intrinsic modes and MMEM (green). For reference, the ERSSTv5 basin mean NASST, which includes both external and internal components, is indicated by the black line. The bottom panel shows results of estimating internal variability using dynamical adjustment (blue). Each of the above estimates is subtracted from the ERSSTv5 basin mean NASST to determine its external or internal counterpart, which is then plotted in the corresponding alternate panel.
Table 2. Correlation values calculated between estimates of internal variability in ERSSTv5. Among these correlations are the correlations between the curves shown in figure 9b. The method applied is listed along the left and top sides. The method names on the left are followed by the standard deviation of the estimated internal variability in parenthesis. Polynomial estimates (Poly) are named by the order of polynomial applied. LOWESS estimates (LOW) are named by the span of the smoothing. EEMD estimates are named by the number of summed components; that is EEMD1 is the residual only, EEMD2 is the residual summed with the lowest frequency intrinsic mode. 'Dynam' refers to the full dynamical adjustment model and 'Reduce' refers to equation 11. Correlations are calculated for methods applied to 1865-2000 to align with the length of the MMEM. The upper right corner is suppressed due to symmetry of the correlation values.

<table>
<thead>
<tr>
<th></th>
<th>Poly1</th>
<th>Poly2</th>
<th>Poly3</th>
<th>LOW25</th>
<th>LOW50</th>
<th>EEMD1</th>
<th>EEMD2</th>
<th>EEMD3</th>
<th>MMEM</th>
<th>Dynam</th>
<th>Reduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly1 (0.19)</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poly2 (0.18)</td>
<td>0.93</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poly3 (0.16)</td>
<td>0.86</td>
<td>0.92</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOW25 (0.07)</td>
<td>0.57</td>
<td>0.61</td>
<td>0.64</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOW50 (0.1)</td>
<td>0.76</td>
<td>0.82</td>
<td>0.89</td>
<td>0.89</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EEMD1 (0.18)</td>
<td>0.94</td>
<td>0.92</td>
<td>0.90</td>
<td>0.60</td>
<td>0.79</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EEMD2 (0.16)</td>
<td>0.93</td>
<td>0.90</td>
<td>0.89</td>
<td>0.64</td>
<td>0.81</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EEMD3 (0.08)</td>
<td>0.47</td>
<td>0.52</td>
<td>0.56</td>
<td>0.97</td>
<td>0.86</td>
<td>0.49</td>
<td>0.53</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMEM (0.16)</td>
<td>0.91</td>
<td>0.89</td>
<td>0.79</td>
<td>0.55</td>
<td>0.72</td>
<td>0.90</td>
<td>0.91</td>
<td>0.45</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynam (0.07)</td>
<td>0.17</td>
<td>0.25</td>
<td>0.29</td>
<td>0.22</td>
<td>0.21</td>
<td>0.45</td>
<td>0.44</td>
<td>0.11</td>
<td>0.26</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Reduce (0.07)</td>
<td>0.12</td>
<td>0.18</td>
<td>0.24</td>
<td>0.06</td>
<td>0.07</td>
<td>0.34</td>
<td>0.31</td>
<td>-0.04</td>
<td>0.07</td>
<td>0.89</td>
<td>1.00</td>
</tr>
</tbody>
</table>

be uncorrelated with other estimates. Standard deviations of each internal variability estimate are calculated and shown following the method name in table 2. The magnitude of standard deviations varies between methods. Methods that estimate external variability, such polynomials and MMEM, tend to have larger standard deviations. The dynamical adjustment estimates have smaller standard deviations, similar in magnitude to LOWESS (25 year span) and EEMD (calculated using two intrinsic modes added to the residual). We refrain from interpreting differences in the estimates of observed variability as neither dynamical adjustment nor the other methods could separate forced and internal variability in historical simulations with consistent skill.

5. Conclusion

This study introduces a new approach to dynamical adjustment. Specifically, the internal component of the basin mean NASST is estimated based on a set of spatial patterns of NASST orthogonal to a uniform pattern. In this implementation, the spatial patterns are Laplacian eigenfunctions,
calculated for the North Atlantic basin using the method of DelSole and Tippett (2015). After projecting NASST onto Laplacian eigenfunctions, time series are smoothed using an 11 year running mean. While the conclusions of this study are not sensitive to the choice of averaging window, the 11 year running mean was chosen to demonstrate that this method is capable of representing variability on time scales longer than interannual variations in atmospheric circulation patterns.

This implementation of dynamical adjustment takes the form of linear regression, with regularization applied to control overfitting. The linear coefficients relating the Laplacian eigenfunctions to the basin mean NASST are learned in a multi-model set of preindustrial simulations which contain only internal variability. The predictive skill of the dynamical adjustment is evaluated in independent preindustrial simulations, where the dynamical adjustment model was skillful for every GCM although the degree of skill depended on the validation GCM. Some amount of this spread in validation skill is due to differences in GCM internal variability that are not well captured when training in a multi-model data set. Future research could utilize this method of representing internal variability to study differences between GCMs, including identifying what features are responsible for predictive skill and analyzing whether differences in identified features could be related to model biases or deficiencies.

The dynamical adjustment was applied to historical simulations without retraining. In contrast to preindustrial simulations, where all variability can be attributed to the coupled atmosphere-ocean system, there are two components of variability, internal and external, in historical simulations. Either the external or internal variability can be viewed as the target in this study, therefore, either could be used as the normalizing variance when calculating NMSE. Normalization by the internal variability is a more strict threshold of skill for the GCMs evaluated in this study, and is chosen as the metric for skill in historical simulations. Similar to the preindustrial validation, the dynamical adjustment is applied to historical simulations from a GCM excluded from the training set. Dynamical adjustment applied to historical simulations is found to have skill in 9 out of 12 GCMs. However, dynamical adjustment performs worse in all historical simulations compared to preindustrial simulations and demonstrates no skill in 3 of 12 GCMs. This degradation in skill is attributed to the presence of externally forced variability in the predictor time series, as demonstrated by the return of skill when dynamical adjustment is applied to predictor time series in which the external variability was removed (e.g., by subtracting out the ensemble mean).
Thus, while the dynamical adjustment approach avoids specifying the temporal shape or amplitude of externally forced variability in the basin mean NASST, it does make the slightly incorrect assumption that external variability does not project onto the spatial patterns of NASST orthogonal to the basin mean. Further study could employ large ensembles and single forcing runs to investigate whether the projection of external variability onto the patterns orthogonal to the basin mean and the consequent degradation of dynamical adjustment skill can be attributed to specific external forcings (Deser et al. 2020).

The dynamical adjustment is compared to several methods that have been previously used to separate forced and internal variability. These methods are evaluated on the same GCMs, allowing a comparison of their performance. It is found that these methods are not skillful in all GCMs and perform similarly, on average, to the proposed dynamical adjustment when applied to historical simulations. However, the dynamical adjustment applied with the removal of external variability from the predictor time series, in general, outperforms other methods across GCMs, being the most skillful method in 8 of 12 GCMs, and tying other methods in 3 additional GCMs. This outperformance emphasizes the potential of this dynamical adjustment method at estimating internal variability in basin mean NASST.

The dynamical adjustment is applied to the observation analysis product ERSSTv5 (Huang et al. 2017b). The dynamical adjustment estimate of internal variability exhibits a relatively small standard deviation and is not highly correlated with estimates produced by other methods. Since neither the dynamical adjustment nor any other method could separate forced and internal variability in historical simulations with consistent skill, we refrain from interpreting differences in the estimated internal variability as it is not clear whether an application to observations, where the true internal variability is not known, should be interpreted as accurate. Nevertheless, it is clear that methods based on different principles yield considerably different estimates of external and internal variability.

Future studies will explore observationally constrained modifications to the dynamical adjustment in search of a consistently skillful estimate in historical simulations, thereby facilitating a more meaningful application to observations. The ultimate goal of this research is to produce an accurate estimate of internal variability over the observed period. We see various opportunities for improving the dynamical adjustment. Recall that the removal of known external variability in the
predictor time series (by subtracting the ensemble mean) greatly enhanced the performance of the
dynamical adjustment. Thus, one strategy for improving the dynamical adjustment would be the
estimation and removal of external variability in the predictor time series. Another strategy is to
include additional variables into the model that may better characterize internal variability in the
North Atlantic, or may have the attractive property of less, or easier to estimate, external variability.
Variables of particular interest include integrated oceanic heat content, sea level height, and sea
surface salinity. Finally, another strategy would be to train the dynamical adjustment in historical
simulations, where external variability would be present during the training step. In doing so, any
perceived issues in translating relationships from preindustrial simulations to historically forced
simulations could be avoided.
Acknowledgments. This research was supported primarily by the National Oceanic and Atmospheric Administration (NA20OAR4310401). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

Data availability statement. The data used in this paper are publicly available from the CMIP5 archive https://esgf-node.llnl.gov/search/cmip5/.

References


Huang, B., and Coauthors, 2017b: Noaa extended reconstructed sea surface temperature (ersst), version 5. https://doi.org/10.7289/V5T72FNM.


42

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-23-0651.1.


Accepted for publication in Journal of Climate. DOI 10.1175/JCLI-D-23-0651.1.


