Forced Component Estimation Statistical Method Intercomparison Project (ForceSMIP)

Robert C. J. Wills[a], Clara Deser[b], Karen A. McKinnon[c], Adam Phillips[b], Stephen Po-Chedley[d], Sebastian Sippel[e], Anna L. Merrifield[a], Constantin Bône[f], Céline Bonfils[d], Gustau Camps-Valls[g], Stephen Cropper[c], Charlotte Connolly[h], Shiheng Duan[d], Homer Durand[g], Alexander Feigin[i], M. A. Fernandez[h], Guillaume Gastineau[f], Andrei Gavrilov[i,g], Emily Gordon[j], Moritz Günther[k], Maren Höver[l,a], Sergey Kravtsov[m], Yan-Ning Kuo[n], Justin Lien[o], Gavin D. Madakumbura[c], Nathan Mankovich[g], Matthew Newman[p], Jamin Rader[h], Jia-Rui Shi[q], Sang-Ik Shin[p,r], Gherardo Varando[s]

[a] *ETH Zurich, Zurich, Switzerland*

[b] *National Center for Atmospheric Research, Boulder, Colorado*

[c] *University of California Los Angeles, Los Angeles, California*

[d] *Lawrence Livermore National Laboratory, Livermore, California*

[e] *Leipzig University, Leipzig, Germany*

[f] *UMR LOCEAN, IPSL, Sorbonne Université, IRD, CNRS, MNHN, Paris, France*

[g] *Image Processing Laboratory, University of Valencia, Valencia, Spain*

[h] *Colorado State University, Fort Collins, Colorado*

[i] *Gaponov-Grekhov Institute of Applied Physics, Russian Academy of Sciences, Nizhny Novgorod, Russia*

[j] *Stanford University, Stanford, California*

[k] *Max Planck Institute for Meteorology, Hamburg, Germany*

[l] *Oxford University, Oxford, UK*

[m] *University of Wisconsin-Milwaukee, Milwaukee, Wisconsin*

[n] *Cornell University, Ithaca, New York*

[o] *Tohoku University, Sendai, Japan*

[p] *NOAA/Physical Sciences Laboratory, Boulder, Colorado*

[q] *New York University, New York City, New York*

[r] *CIRES, University of Colorado Boulder, Boulder, Colorado*

[s] *Image Processing Laboratory, University of Valencia, Valencia, Spain*

Corresponding author: Robert Jnglin Wills, r.jnglinwills@usys.ethz.ch

ABSTRACT: Anthropogenic climate change is unfolding rapidly, yet its regional manifestation can be obscured by internal variability. A primary goal of climate science is to identify the externally forced climate response from amongst the noise of internal variability. Separating the forced response from internal variability can be addressed in climate models by using a large ensemble to average over different possible realizations of internal variability. However, with only one realization of the real world, it is a major challenge to isolate the forced response in observations. In the Forced Component Estimation Statistical Method Intercomparison Project (ForceSMIP), contributors used existing and newly developed statistical and machine learning methods to estimate the forced response over 1950-2022 within individual realizations of the climate system. Participants used neural networks, linear inverse models, fingerprinting methods, and low-frequency component analysis, among other approaches. These methods were trained on large ensemble from multiple climate models and then applied to observations. Here we evaluate method performance within large ensembles and investigate the estimates of the forced response in observations. Our results show that many different types of methods are skillful for estimating the forced response, though the relative skill of individual methods varies depending on the variable and evaluation metric. Methods with comparable skill in models can give a wide range of estimates of the forced response in observations, illustrating the epistemic uncertainty in forced response estimates. ForceSMIP gives new insights into the forced response in observations, its uncertainty, and methods for its estimation.

SIGNIFICANCE STATEMENT: The ForceSMIP project aims to reduce uncertainty in estimates of the climate response to anthropogenic and other external forcing and to evaluate statistical and machine learning methods designed to estimate the forced response from individual realizations of the climate system. New and existing statistical and machine learning methods are evaluated within climate models, for which the forced response is known. Applying these methods to observations gives an estimate of the real-world forced response. The observational forced response estimate agrees with climate models on the large-scale features but also shows discrepancies that give insights into responses that may not be simulated well by climate models. In some regions with large internal variability, such as the North Atlantic ocean, it remains difficult to determine the relative contributions of anthropogenic forcing and internal variability to historical changes.

1. Introduction

Climate variability and change is composed of forced and unforced components. The forced component of climate change, or forced response, includes all spatiotemporal changes in climate in response to external forcing. Here we consider the net response to forcing from greenhouse gasses, anthropogenic aerosols, land-use change, stratospheric ozone, and natural forcing (e.g., volcanic sulfur emissions and solar variability). The unforced component is due to internal variability of the climate system, for example due to modes of climate variability such as the El Niño-Southern Oscillation (ENSO), Atlantic multi-decadal variability (AMV), and the North Atlantic Oscillation (NAO). In some regions or variables that are prone to large internal variability, the unforced component can be comparable in magnitude to or larger than the forced component, even in multi-decadal trends (Deser et al. 2012, 2014; Lehner et al. 2020). Accurate estimation of the forced and unforced components of regional climate change is critical for the attribution of historical climate changes and the characterization and understanding of climate variability and extremes.

In climate models, the forced component can be isolated using large ensembles, where the same climate model is run many times with the same forcing but differences in initial conditions, leading to differences in the phasing of internal variability. For a climate measure of interest, the ensemble mean of a large ensemble gives an estimate of the forced response, with larger ensembles needed for variables with lower signal-to-noise ratio (Milinski et al. 2020). Assuming linear additivity of the forced and unforced components, the difference of an individual realization from the ensemble

mean gives the contribution of internal variability. An example is shown for 1980-2022 SST trends from a single member of the ACCESS-ESM1-5 large ensemble in Fig. 1, where the full trend (Fig. 1a) is separated into forced and unforced components (Fig. 1b and c, respectively) based on the ensemble mean. Large ensembles are now a widespread tool used for climate change attribution, climate projections, and studies of climate variability and extremes (Deser et al. 2020). However, there is only a single realization of the actual climate system, and it is therefore substantially harder to separate observed climate change into forced and unforced components, which is critical for evaluating climate models and understanding discrepancies between models and observations (Wills et al. 2022; Blackport and Fyfe 2022; Simpson et al. 2025).

Individual studies have used one or more statistical methods to estimate the forced response in observations for various applications. For example, separating the forced and unforced component of AMV and the associated Sahel rainfall variability has received particular attention (Ting et al. 2009; Booth et al. 2012; Zhang et al. 2013; Frankcombe et al. 2015; Bellucci et al. 2017; Frankignoul et al. 2017; Wills et al. 2020; Qin et al. 2020; Latif et al. 2022; He et al. 2023). By using different methods to estimate the forced response, these studies have reached widely differing conclusions ranging from the AMV is mostly forced (Booth et al. 2012; Wills et al. 2020; He et al. 2023) to the AMV is mostly internal variability (Zhang et al. 2013; Ting et al. 2009; Qin et al. 2020; Latif et al. 2022), although many of these studies acknowledge the uncertainty in this conclusion. There are also a range of conclusions on the forced and unforced contributions to the multi-decadal modulation of the global warming rate (DelSole et al. 2011; Dai et al. 2015; Stolpe et al. 2017; Kravtsov et al. 2018) and multi-decadal changes in the Pacific SST pattern (Olonscheck et al. 2020; Wills et al. 2022; Seager et al. 2022; Rugenstein et al. 2023) and the Aleutian low (Smith et al. 2016; Oudar et al. 2018), among other climate indices. All of these questions would benefit from a systematic comparison of methods for estimating the forced response in observations, and this is what the Forced Component Estimation Statistical Method Intercomparison Project (ForceSMIP) aims to do.

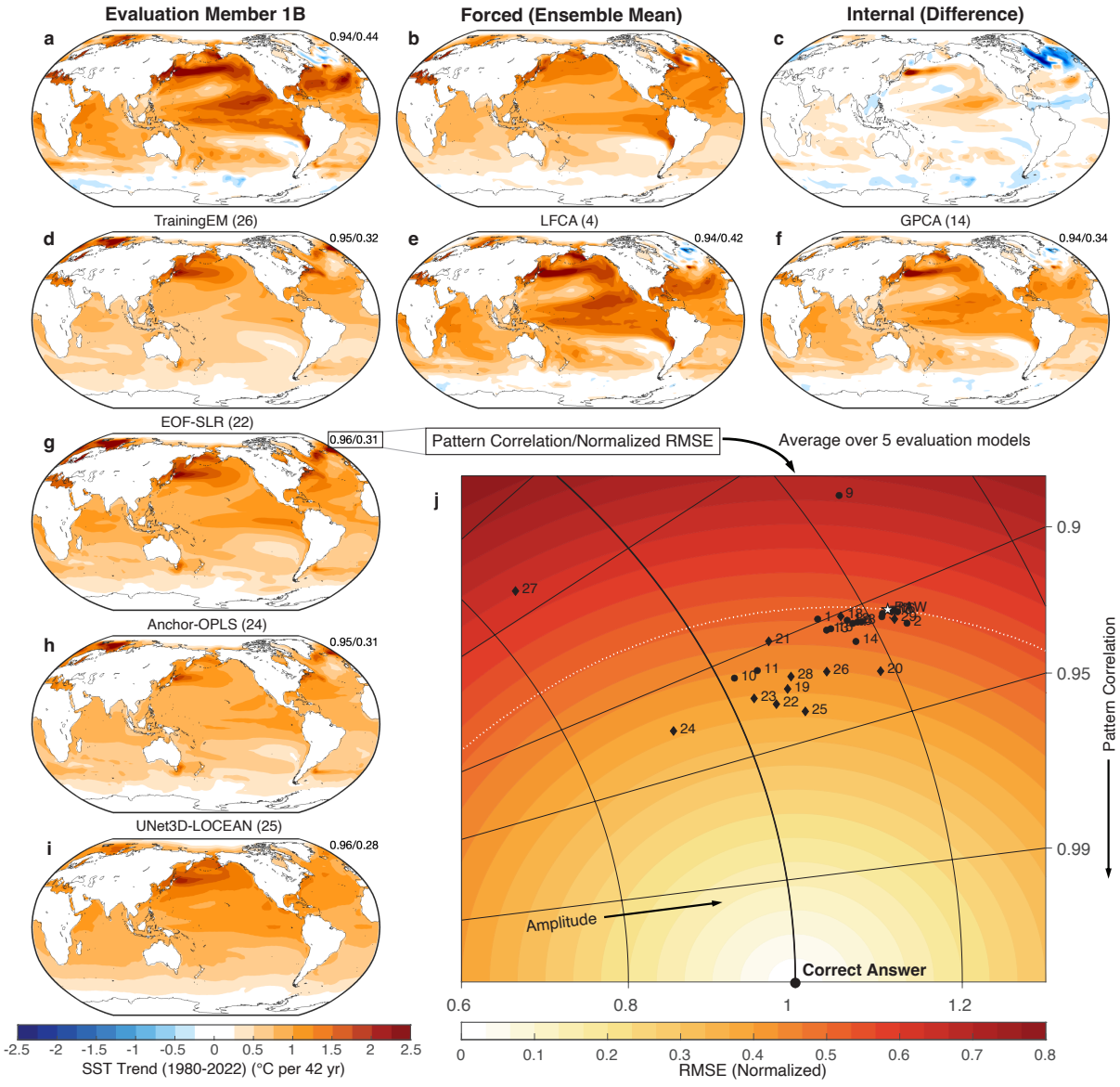


FIG. 1. Illustration of selected methods and how they are evaluated in ForceSMIP using climate model large ensembles. ForceSMIP participants generated a forced response estimate for each of 10 unlabeled evaluation members. While the forced response estimate includes spatiotemporal variations across 8 variables over 1950-2022, here each panel shows 1980-2022 annual-mean SST trends: (a) A single evaluation member (1B) from a large ensemble, which after the submissions was revealed to be from ACCESS-ESM1-5. (b) The "correct answer" is thus estimated from the 40-member ensemble mean of ACCESS-ESM1-5. (c) The internal variability contribution to the trend in (a) is computed as (a) - (b). (d) The TrainingEM method is rescaled from the ensemble mean of the training models and does not use information from ACCESS-ESM1-5 other than the global-mean surface temperature trend. It is a reference method meant to illustrate the forced response that would be estimated from a multi-model ensemble mean. (e)-(i) Forced response estimates from selected ForceSMIP methods, with names and numbers in the titles corresponding to those in Table 1. (j) Taylor diagram showing root mean square error (RMSE) normalized by the root-mean-square-amplitude of the ensemble mean (colors), the root-mean-square-amplitude normalized by the root-mean-square-amplitude of the ensemble mean, i.e., $\sigma_i/\sigma_{\mathrm{REF}}$ (black arcs), and the uncentered pattern correlation r_i (black rays). Each method is shown as a symbol with numbers corresponding to those in Table 1; diamonds show methods that use pattern information from the training models; circles show methods that do not. The raw data (a) is shown as a white star, and the dashed white line shows $\delta\mathrm{RMSE}_i/\mathrm{RMSE}_{\mathrm{RAW}} = \delta r_i / r_{\mathrm{RAW}}$. The skill metrics are averaged over the 5 "unseen model" evaluation members as explained in the text.

TABLE 1. Statistical and machine learning methods for forced response estimation submitted to ForceSMIP Tier 1. Included is information about the institutions involved in developing the methods, a rough categorization of the method type (NN = neural network), whether the method uses pattern information from the training models, whether the method is applied to multiple field variables at once (e.g., using the SST forced response to inform the precipitation forced response), and the number of tunable parameters in the method (i.e., parameters which can be influenced by the training models; reported by the method contributor). Methods are ordered by the number of tunable parameters, and this numbering is used throughout the text and figures.

#	Name	Institution(s)	Type of Method	Pattern Information	Multi-Field	N Parameters
1	RegGMST	NCAR	Regression	No	Yes	0
2	4th-Order-Polynomial	N/A	Reference	No	No	0
3	10yr-Lowpass	N/A	Reference	No	No	0
4	LFCA	ETHZ	LFCA	No	No	2
5	LFCA-2	ETHZ	LFCA	No	No	2
6	MF-LFCA	ETHZ	LFCA	No	Yes	2
7	MF-LFCA-2	ETHZ	LFCA	No	Yes	2
8	LIMnMCA	Cornell, Tohoku	LIM	No	Yes	2
9	ICA-lowpass	MPI-M	Other	No	No	3
10	LIMopt	ETHZ	LIM	No	No	3
11	LIMopt-filter	ETHZ	LIM	No	No	4
12	Colored-LIMnMCA	Cornell, Tohoku	LIM	No	Yes	5
13	DMDc	Valencia	LIM	No	No	75
14	GPCA	Valencia	Causal Inference	No	No	88
15	GPCA-DA	Valencia	Causal Inference	No	Yes	89
16	RegGMST-LENSem	NCAR	Regression	No	Yes	876
17	MLR-Forcing	LLNL	Regression	No	Yes	1.1e4
18	SNMP-OF	ETHZ	Fingerprinting	Yes	Yes	1.0e4
19	AllFinger	LLNL, WHOI, UCLA	Fingerprinting	Yes	No	1.0e4
20	MonthFinger	LLNL, WHO, UCLA	Fingerprinting	Yes	No	1.2e5
21	3DUNet-Fingerprinters	UCLA, LLNL, WHOI	NN	Yes	No	5.4e5
22	EOF-SLR	IAP, Milwaukee	Fingerprinting	Yes	No	O(1e6)
23	LDM-SLR	IAP, Milwaukee	Fingerprinting	Yes	No	O(1e6)
24	Anchor-OPLS	Valencia	Regression	Yes	No	2.1e6
25	UNet3D-LOCEAN	LOCEAN	NN	Yes	Yes	2.7e6
26	TrainingEM	N/A	Reference	Yes	Yes	9.1e6
27	RandomForest	UCLA	Random Forest	Yes	No	1.0e7
28	EncoderDecoder	CSU	NN	Yes	No	2.3e7
29	EnsFMP	ETHZ	Fingerprinting	Yes	Yes	4.5e7
30	ANN-Fingerprinters	LLNL	NN	Yes	No	1e16

Large ensembles provide a perfect-model testbed for methods that estimate the forced response from individual ensemble members, because their ensemble mean gives a good estimate of the true forced response in that model. This has been the approach of several previous studies, which have developed statistical or machine learning (StatML) methods to estimate the forced response in single realizations, evaluated them using large ensembles, and then applied them to observations (Deser et al. 2014; Frankcombe et al. 2015; Frankignoul et al. 2017; Sippel et al. 2019; Wills et al. 2020; Bône et al. 2024; Rader et al. 2025). However, these studies have generally focused on one or two methods compared to some simple reference methods, and there has been no broader systematic intercomparison of methods. Furthermore, these studies have primarily targeted surface temperature and/or precipitation, and it is not clear how well the methods used generalize to other climate variables. ForceSMIP aims to systematically compare various StatML methods for forced response estimation across multiple variables in a common framework. Here we both assess which methods are skillful within the large ensemble testbed and assess the spread of estimated forced responses in observations.

The rest of the paper is organized as follows. In Section 2, we present the ForceSMIP framework and the climate model large ensemble and observational datasets used. In Section 3, we describe the 30 StatML methods that have been submitted to ForceSMIP. In Section 4, we evaluate the skill of methods for the spatial patterns of long-term trends across multiple variables, grid-scale spatiotemporal variability, and the temporal evolution of selected climate indices. In Section 5, we show examples of the forced responses in observations based on the most skillful methods. Finally, in Section 6, we draw conclusions and discuss implications, potential applications, and future directions.

2. ForceSMIP Framework and Data

The overarching idea of ForceSMIP is that community contributors develop and train StatML methods to estimate the forced response from single ensemble members and then apply them to model-based evaluation data and observations. The methods are then evaluated based on their forced response estimates in the model-based evaluation data, where the true forced response is known. Finally, the observational forced response estimates can be compared across methods that have proven skillful in the model testbed.

In order to train their methods, contributors were provided with data from 5 climate model large ensembles (Table 2). The identity of these ***training models*** was revealed to the participants. Data from all ensemble members of the historical and future scenario simulations was provided for 8 climate variables over 1850-2100: sea-surface temperature (SST), 2-meter air temperature (T2m), precipitation (PR), sea-level pressure (SLP), monthly-maximum of daily precipitation (monmaxpr), monthly-maximum of daily-maximum temperature (monmaxtasmax), monthly-minimum of daily-minimum temperature (monmintasmin), and zonal-mean atmospheric temperature (zmTa). The first four variables were taken from monthly outputs of tos, tas, pr, and psl, respectively, using the naming conventions of CMIP6 (Eyring et al. 2016). The remaining four variables were processed from daily output of pr, tasmax, and tasmin and monthly output of ta, respectively. All variables were interpolated to a common 2.5° grid following Brunner et al. (2020). Four of the variables were then additionally processed with CDO (Schulzweida 2023) commands to make derived variables: daily pr with monmax to make monmaxpr, daily tasmax with monmax to make monmaxtasmax, daily tasmin with monmin to make monmintasmin, and montly ta with zonmean to make zmTa, where monmax takes a monthly maximum, monmin takes a monthly minimum, and zonmean takes a zonal mean. After this processing, all variables have two spatial dimensions (lat and pressure for zmTa; lat and lon for all others) and monthly time resolution.

After developing and training their methods, the contributors submitted: (1) descriptions and basic information about their methods, (2) their method code, and (3) output from application of their method to estimate the forced response across all variables in 10 evaluation members over the period 1950-2022. For the purposes of ForceSMIP, we use a broad definition of the ***forced response***: it includes all spatiotemporal variations in the ensemble mean, thus including climate variations due to natural climate forcings (e.g., volcanic eruptions and solar variability) and anthropogenic influences (e.g., anthropogenic emissions of greenhouse gases and aerosols). Contributors therefore had to submit forced response estimates for all variables at monthly time resolution for all points on the 2.5° analysis grid. Nevertheless, much of the discussion in the hackathon that preceded the method submission focused on 1950-2022 trends or 1980-2022 trends, and many participants focused on skill metrics like the pattern correlation and root-mean-square error (RMSE) in long-term linear trends, as shown in Figs. 1 and 2. These figures will be discussed in more detail in Section 4, but the overall idea is that by applying StatML to a single ensemble member (for which

TABLE 2. Large ensemble and observational data used in ForceSMIP. The first 5 models are the training models and the next 5 models are "unseen models", which are the source of the evaluation members 1B, 1D, 1E, 1G, and 1J used for method evaluation in this paper. Evaluation member 1I is the observational data. "Total Members" indicates the number of members used to compute the ensemble mean, with the number in parenthesis indicating the number of future scenario members if it is different than the number of historical simulation members. CESM2 members are those with smoothed biomass burning (Rodgers et al. 2021). Note that due to data problems for zmTa in some members of EC-Earth3, only 13 (51) of the total ensemble members were used to compute the ensemble mean for this variable.

Model	Evaluation Member	Total Members	Future Scenario	Reference
CanESM5	1C (r20i1p2f1)	25	SSP585	Swart et al. (2019)
CESM2	1F (LE 1281.019)	50	SSP370	Rodgers et al. (2021)
MIROC6	1H (r11i1p1f1)	50	SSP585	Tatebe et al. (2019)
MIROC-ES2L	N/A	30	SSP245	Hajima et al. (2020)
MPI-ESM1-2-LR	1A (r23i1p1f1)	30	SSP585	Olonscheck et al. (2023)
ACCESS-ESM1-5	1B (r10i1p1f1)	40	SSP585	Ziehn et al. (2020)
EC-Earth3	1D (r6i1p1f1)	18 (58)	SSP585	Wyser et al. (2021)
GFDL-SPEAR-MED	1E (r3i1p1f1)	30	SSP585	Delworth et al. (2020)
IPSL-CM6A-LR	1G (r3i1p1f1)	33 (11)	SSP245	Boucher et al. (2020)
NorCPM1	1J (r4i1p1f1)	30	SSP245	Bethke et al. (2021)
ERA5/ERSST5	1I	1	N/A	Hersbach et al. (2020); Huang et al. (2017)

the trends over 1980-2022 are shown in Figs. 1a and 2b), the forced response estimates submitted by ForceSMIP contributors (Figs. 1d-i and 2d-i) should approximate as closely as possible the ensemble mean of the corresponding large ensemble (Figs. 1b and 2b) by removing the internal variability (Figs. 1c and 2c). The 1980-2022 trends shown here are just one way in which the spatiotemporally resolved forced response estimates are evaluated in Section 4.

The ***evaluation members*** are individual ensemble members of 9 different climate models (Table 2) and observational and reanalysis data that has been processed equivalently to the model data. All evaluation members had the metadata removed so that it was not possible to determine which dataset they came from. Only two of the ForceSMIP organizers (C. Deser and A. Phillips) knew the identity of these evaluation members. Of the 9 model-based evaluation members, 5 were from ***unseen models*** that were not among the training models. The method evaluation in Section 4 will primarily rely on these 5 unseen-model evaluation members. The forced response estimates for

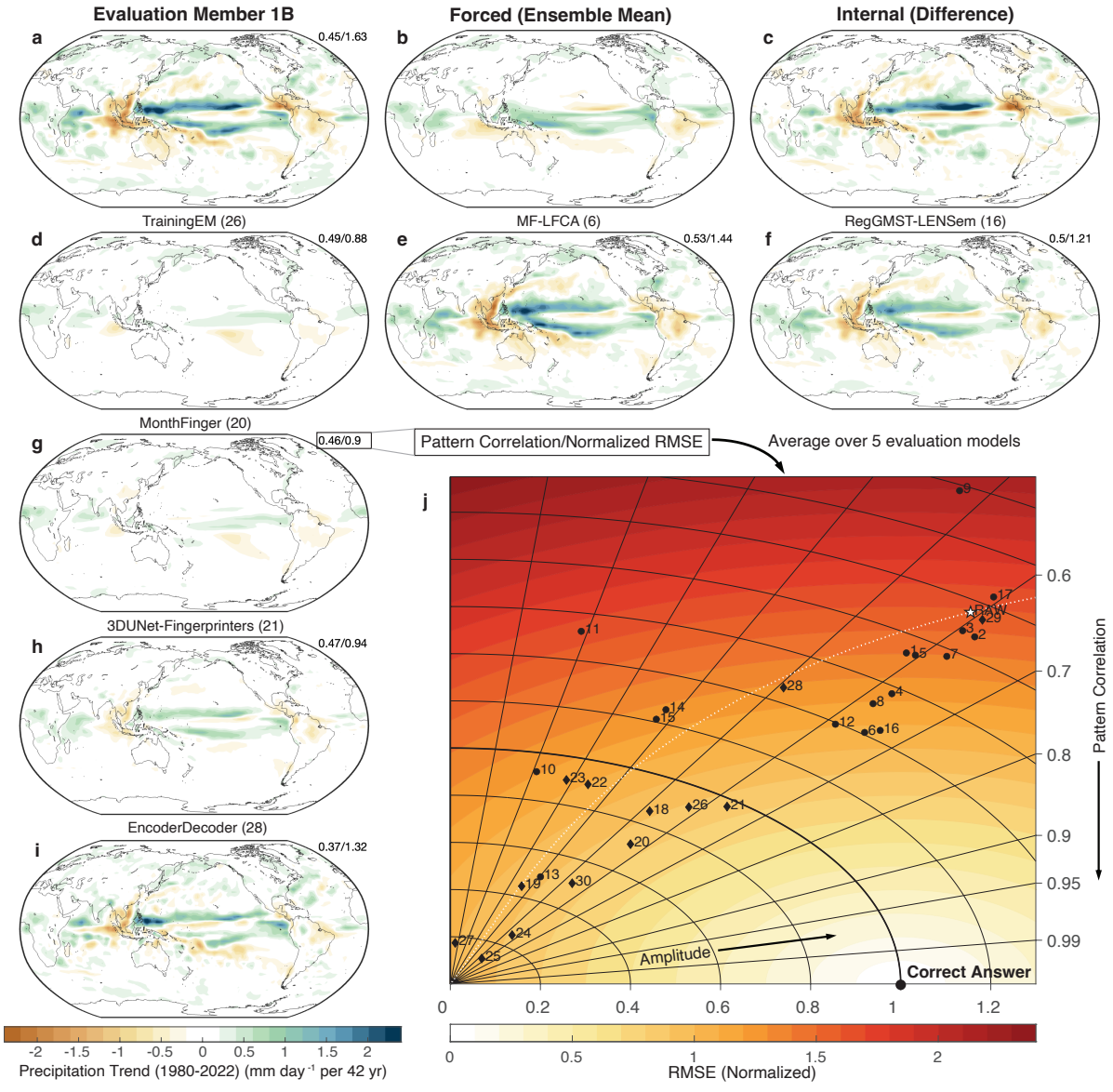


FIG. 2. Same as Fig. 1, but for precipitation.

the evaluation members will be evaluated against the ensemble means computed over all available ensemble members. Note that for two models (EC-Earth3 and IPSL-CM6A-LR), there are a different number of historical and future scenario members, and in these cases the ensemble mean is computed separately in the historical and scenario simulations and then concatenated. The ensemble mean against which methods are evaluated will still have some internal variability in it, which may lead to a slight underestimation of method skill.

Data from observations and reanalysis are included as one of the evaluation members (1I), so that the methods can be evaluated and applied to observations in a single round of forced response submissions. This initial round of "Tier 1" ForceSMIP submissions focuses on 1950-2022, which was chosen based on the availability of reanalysis data over this period. As such, all "observational" data in Tier 1 except SST is actually from ERA5 reanalysis (Hersbach et al. 2020). Daily tasmax, tasmin, and pr were computed from ERA5 hourly 2-meter temperature and rainfall data, and the other variables were computed from monthly ERA5 data. SST is from the NOAA Extended Reconstructed SST version 5 (ERSST5; Huang et al. (2017)). Accordingly, the observational forced response estimates from ForceSMIP Tier 1 will be subject to any biases present in the ERA5 and ERSST5 datasets. Subsequent Tiers of ForceSMIP will focus on different time periods, 1900-2023 and 1979-2023, on which different sets of observational data are available.

While the training data, evaluation methods, and forced response estimates are all available at monthly temporal resolution, all analysis in this paper focuses on annual or seasonal averages (for SST, T2m, PR, SLP, and zmTA), annual maximums, and annual minimums. The annual maximum of monmaxtasmax is called TXx, the annual maximum of monmaxpr is called Rx1day, and the annual minimum of monmintasmin is called TNn, following standard conventions in the study of extreme events (Zhang et al. 2011).

3. Statistical and Machine Learning Methods for Forced Response Estimation

Thirty StatML methods were submitted to this first tier of the ForceSMIP project. They comprise a diverse mix of approaches including linear regression on global-mean temperature or forcing timeseries, low-frequency component analysis (LFCA), linear dynamical mode methods such as linear inverse models (LIMs), linear fingerprinting methods, and neural networks or other machine learning (ML) methods (Table 1). This includes both established methods (e.g., LFCA; Wills et al. (2020), LIMopt; Frankignoul et al. (2017), and regression on global-mean surface temperature; Ting et al. (2009); Deser and Phillips (2021)) and methods newly created for ForceSMIP. The development of many of these methods began at a hackathon held at NCAR and ETH Zurich in August 2023. These methods are briefly summarized in the following subsections, with key details listed in Table 1. In Table 1 and throughout the text, methods are ordered by their number of tunable parameters, which range from 0 to O(10^7) or higher. More detailed information about all

methods can be found in the Supplementary Material, and code for all methods can be found at `https://github.com/ForceSMIP/tier1-methods` (a persistent identifier will be issued upon publication).

a. Linear regression on global-mean temperature or forcing timeseries: RegGMST, RegGMST-LENSem, MLR-Forcing

Many studies of internal variability, including ENSO, AMV, and the Pacific Decadal Oscillation (PDO), remove anomalies associated with global-mean sea-surface temperature (GMSST) or global-mean surface temperature (GMST) changes when defining indices of this variability (Trenberth and Shea 2006; Ting et al. 2009; Frankignoul et al. 2017; Deser and Phillips 2021). Underlying these approaches is an implicit estimation of the forced response based on GMSST or GMST, under the assumption that those globally aggregated metrics are good proxies of the forced response. In ForceSMIP, two methods, RegGMST and RegGMST-LENSem, estimate the forced response by regressing each field onto a timeseries of GMST and combining that regression pattern with the GMST timeseries. RegGMST uses regression on GMST from the target evaluation member and RegGMST-LENSem uses regression on the ensemble-mean GMST from the 50 CESM2 large ensemble members (Deser and Phillips 2023b).

A similar approach is to regress each field onto timeseries representing important external forcings or internal variability. The method MLR-forcing uses a multiple-linear-regression approach to regress each field onto regional aerosol forcing timeseries and timeseries representing the response to various forcings (including greenhouse gasses, volcanic emissions, and solar forcing) and on detrended Niño3.4 indices, estimating the forced response as the components associated with the forcing timeseries.

b. Low-frequency component analysis: LFCA, LFCA-2, MF-LFCA, MF-LFCA-2, ICA-lowpass

Low-frequency component analysis (LFCA) is a method to objectively identify the slowest evolving spatial patterns in a dataset, using linear discriminant analysis applied to principal components to find patterns that maximize the ratio of low-frequency to total variance (Schneider and Held 2001; Wills et al. 2018, 2020). It has been used both to study decadal climate variability (e.g., Wills et al. 2019) and to separate forced and unforced components of climate change (Wills et al.

2020). Its usage as a method to separate forced and unforced components is based on the understanding that the forced response evolves on a longer timescale than most internal variability, i.e., it is using timescale separation to separate forced and unforced components. The application of LFCA in ForceSMIP follows Wills et al. (2020), using a 10-year lowpass filter and including 1 or 2 low-frequency patterns in the forced response estimate (methods LFCA and LFCA-2, respectively). Additionally, the methods MF-LFCA and MF-LFCA-2 apply the same method to two variables at a time by combining each field with SST, or in the case of SST, combining it with T2m, with each field normalized by the trace of its covariance matrix.

While not a form of LFCA, the ICA-lowpass method uses independent component analysis (Hyvärinen and Oja 2000), which similarly finds linear combinations of a chosen set of principal components that maximize a variance criterion, in this case the statistical independence of the principal components. ICA-lowpass applies independent component analysis to lowpass filtered data and identifies the forced pattern based on its spatial uniformity, under the assumption that the spatial scales of forced climate change are larger than those of internal variability.

c. Linear dynamical mode methods: LIMopt, LIMopt-filter, LIMnMCA, Colored-LIMnMCA, DMDc, GPCA, GPCA-DA

Linear dynamical mode methods aim to describe the spatiotemporal variability in a dataset by a set of linear dynamical equations, which determine the evolution of a field from one timestep to the next. The specific case of the Linear Inverse Model (LIM), where the evolution operator is determined from lagged covariance information, is widely used in climate science (Penland and Sardeshmukh 1995; Alexander et al. 2008). The concept of a least damped mode of a LIM was introduced by Penland and Sardeshmukh (1995) and has been used to separate the ENSO-related or forced variations in a dataset (Compo and Sardeshmukh 2010; Solomon and Newman 2012; Frankignoul et al. 2017; Xu et al. 2022). For ForceSMIP, the LIMopt and LIMopt-filter methods apply the method LIM optimal perturbation pattern and LIM optimal perturbation filter method of Frankignoul et al. (2017) (see also Wills et al. 2020). The LIMnMCA and ColoredLIMnMCA methods combined a similar approach applied to SST with a maximum covariance analysis to find the covariations between SST and the other ForceSMIP variables, an extra step which we will show made it much more successful than other linear dynamical mode methods for non-temperature

variables (i.e., PR, SLP, and Rx1day). ColoredLIMnMCA differs from LIMnMCA by the use of a LIM for colored Gaussian noise (Lien et al. 2025).

The DMDc is similar in approach to LIMopt, but with a generalization of LIM to include a linear forcing component (Proctor et al. 2016). Similarly, GPCA and GPCA-DA are based on the representation of the data as a combination of an autoregressive process and a forced response, where the forced response is estimated by the "direct Granger effect" of the exogenous forcing signal, and are an extension of the method presented in Varando et al. (2022). Like MLR-Forcing, these methods employ additional forcing timeseries. Compared to GPCA, GPCA-DA additionally uses empirical orthogonal functions (EOFs) of SLP to control against the internal variability they may represent.

d. Linear fingerprinting methods: AllFinger, MonthFinger, SNMP-OF, EOF-SLR, LDM-SLR, Anchor-OPLS, EnsFMP

Broadly speaking, linear fingerprinting methods use model-based forced response patterns as an initial guess of the forced response and then estimate the contribution of this pattern to the observations (or an individual ensemble member treated like observations). While traditional uses of fingerprinting for detection and attribution generally aim to find a timeseries indicating the amplitude of the forced response pattern compared to internal variability, the fingerprinting methods in ForceSMIP additionally combine that timeseries with an estimate of the forced pattern.

AllFinger and MonthFinger are derived from pattern-based fingerprint analyses (Hasselmann 1979; Santer et al. 2023), where the forced pattern fingerprint is obtained by averaging across models and extracting the leading EOF (amplifying the signal and reducing the noise). Observations—or individual model realizations—are projected onto the fingerprint to create a pseudo-PC time series, measuring the similarity between the fingerprint and the target's time-varying patterns. The predicted trend map is reconstructed using the forced pattern fingerprint and the pseudo-PCs.

EOF-SLR and LDM-SLR methods first estimate each model's forced response components (timeseries) in a basis of spatial patterns given by either ensemble EOF or linear dynamic mode (LDM) decomposition (Gavrilov et al. 2020, 2024) of multi-model ensemble simulations. Then a set of optimal fingerprinting patterns is trained to deduce the forced response from a single

realization in this ensemble. These patterns are constructed to be robust to model uncertainty within the training ensemble, and can thus be applied to the unseen data.

Anchor-OPLS is a generalization of the anchor regression framework for fingerprint extraction introduced by Sippel et al. (2021), where forced responses are predicted at every grid point and orthonormalised partial least squares (OPLS) is used instead of ordinary least squares.

SNMP-OF is a combination of signal-to-noise maximizing pattern (SNMP) analysis (Ting et al. 2009; Wills et al. 2020) with optimal fingerprinting (Hegerl et al. 1996); it finds SNMPs from the training models and then projects their optimal fingerprint onto observations, finally recomputing a pattern from regression of observations onto the resulting signal-to-noise maximizing timeseries. EnsFMP attempts to combine the two steps into one by applying SNMP analysis to numerous combinations of model ensemble members and observations. Unlike the other fingerprinting methods in ForceSMIP, these two methods recompute a forced response pattern within observations, and they thus stick closer to the raw data.

e. Machine learning methods: 3DUNet-Fingerprinters, UNet3D-LOCEAN, RandomForest, EncoderDecoder, ANN-Fingerprinters

ML contributions to ForceSMIP include one based on a recently developed method (UNet3D-LOCEAN; Bône et al. 2024), and four methods newly developed for ForceSMIP, including one that has recently been used to attribute the record-high 2023 SST (EncoderDecoder; Rader et al. 2025). Architectures used include a type of convolutional neural network called a U-Net (3DUNet-Fingerprinters, UNet3D-LOCEAN), Encoder-Decoder neural networks (EncoderDecoder, ANN-Fingerprinters), and random forests (RandomForest). Two of the ML methods learn to remove the internal variability (UNet3D-LOCEAN, EncoderDecoder), and the other three learn to estimate the forced response (3DUNet-Fingerprinters, ANN-Fingerprinters, RandomForest). ANN-Fingerprinters additionally uses the year as one of the inputs. The ML methods used in this study vary in complexity (e.g., *N* Parameters in Table 2) and employ different parameter tuning and training strategies. Interestingly, the U-Nets trained on the internal variability and the forced component exhibit different strengths across variables (Section 4).

f. Reference methods: 4th-Order-Polynomial, 10yr-Lowpass, TrainingEM

In addition to the methods submitted to ForceSMIP, we compare against 3 reference methods, which involve minimal processing of either the raw data or the training-data ensemble mean. Two of the reference methods are simple methods to remove high-frequency noise in the raw data. 4th-Order-Polynomial estimates the forced response as a 4th-order-polynomial fit to timeseries of each variable at each grid point. It has been used to estimate the forced response in a seminal paper by Hawkins and Sutton (2009) and later tested in large ensembles by Lehner et al. (2020). 10-yr-Lowpass estimates the forced response as all variability left after application of a 10-yr Lanczos lowpass filter.

While the first two reference methods are based entirely on the data within the single realization of interest, the third reference method, TrainingEM, represents an opposite extreme where most information is taken from the training data. TrainingEM simply takes the multi-model ensemble mean of the 5 training models as the forced response estimate and rescales it by a constant so that it has the same GMST trend over 1950-2022 as the single realization of interest. This is similar to the scaling method introduced by Steinman et al. (2015) and evaluated by Frankcombe et al. (2015). TrainingEM thus represents a type of null hypothesis where climate models have a perfect estimate of the forced response, up to a rescaling based on differences in climate sensitivity.

4. Method Evaluation

In order to evaluate the skill of the ForceSMIP methods in isolating the forced response in individual realizations of the climate system, we focus on their skill in determining the forced response in the 5 unseen climate models (i.e., those not in the training dataset) from a single member of their large ensembles. However, the results are not systematically different in the 4 evaluation members that were part of the training data (Fig. S1). The forced response estimates include monthly values globally for 1950-2022, so there are many metrics on which they could be evaluated. We will focus here on skill in estimating long-term forced trends, the grid-scale temporal evolution of the forced response, and the forced response in an illustrative set of large-scale climate indices.

a. Long-term trends

Our method for evaluating method skill in isolating the forced component of long-term trends can be visualized in Figs. 1 and 2, showing estimates of forced 1980-2022 annual-mean SST and PR trends from a single evaluation member. The forced trend estimate from each method (panels d-l) is compared against the true forced response, as estimated by the ensemble mean of the corresponding large ensemble (panel b). For comparison, we also show how well the linear trend in the raw data from the evaluation member approximates the true forced response (panel a), which is a reference point we expect methods to improve upon. The difference between the full trend in the raw data and the ensemble-mean forced trend is the contribution of internal variability (panel c), which the methods aim to remove.

We quantify the skill of each method's estimate of the forced trend pattern $\mathbf{f}_i$ compared to the true forced trend pattern $\mathbf{f}_0$ in terms of:

1. the uncentered pattern correlation, or cosine similarity, $r_i = \langle \mathbf{f}_i, \mathbf{f}_0 \rangle \, \|\mathbf{f}_i\|^{-1} \|\mathbf{f}_0\|^{-1}$, where $\langle \cdot, \cdot \rangle$ indicates an area-weighted inner product, $\| \cdot \| = p^{-1} \sqrt{\langle \cdot, \cdot \rangle}$ indicates an area-weighted inner-product norm, and p is the total number of grid cells,

2. $\mathrm{RMSE}_i = \|\mathbf{f}_i - \mathbf{f}_0\|$ normalized by the amplitude of the true forced trend pattern $\sigma_0 = \|\mathbf{f}_0\|$, where each method's normalized RMSE is hereafter referred to as nRMSE_i, and

3. the amplitude ratio of the predicted and true forced trend patterns (σ_i / σ_0).

The root mean square over the 5 unseen-model evaluation members of each method's nRMSE_i and forced trend pattern amplitude $\sigma_i = \|\mathbf{f}_i\|$ is plotted on a Taylor diagram (Figs. 1j and 2j). The colored shading shows nRMSE_i, the curved black arcs show contours of the amplitude ratio of the predicted and true forced trend patterns (σ_i / σ_0), and the black rays show contours of the uncentered pattern correlation r_i. Because these three metrics are inter-related, the uncentered pattern correlation r_i shown for each method in the Taylor diagrams is determined from the other two variables by:

$$r_i = \frac{\sigma_i^2 + \sigma_0^2 - \mathrm{RMSE}_i^2}{2\sigma_i \sigma_0} = \frac{1 + (\sigma_i/\sigma_0)^2 - \mathrm{nRMSE}_i^2}{2\sigma_i/\sigma_0}. \tag{1}$$

This equation is exact when applied to a single evaluation member but is approximate when applied to the averages over 5 members in the Taylor diagrams. Note that the Taylor diagrams in this paper

do not show the full quadrant, as is traditional (Taylor 2001). Rather, they zoom in on the regions where the points are. Our variant on the Taylor diagram is partially inspired by the "solar diagram" of Wadoux et al. (2022), however, in our case the quantitative information remains the same as in a traditional Taylor diagram.

One noteworthy observation from Figs. 1 and 2 is that methods that do not use pattern information from the training models (methods 1-17; shown with circular symbols in the Taylor diagrams; hereafter simple methods) estimate forced trends that look more like the raw trend from the evaluation member (Fig. 1e-f, cf. Fig. 1a; 2e-f, cf. Fig. 2a). On the other hand, methods that use pattern information from the training models (methods 18-30; shown with diamond symbols in the Taylor diagrams) estimate forced trends that look more like the ensemble-mean of the training models (Fig. 1g-i, cf. Fig. 1d; 2g-i, cf. Fig. 2d). This is especially true for SST, and we suspect that the reason for more diversity in forced precipitation trends is that not all training models have the same forced precipitation response. Methods that use pattern information generally perform better in terms of nRMSE than the methods that do not, but they will be more influenced by any systematic biases in the training models, and they do not perform as well in terms of pattern correlation for precipitation.

The Taylor diagrams for 1980-2022 trends in all 8 variables are shown in Fig. 3 and 4. For all variables, the majority of ForceSMIP methods are skillful, where we consider a method skillful if $\delta\mathrm{RMSE}_i/\mathrm{RMSE}_{\mathrm{RAW}} < \delta r_i / r_{\mathrm{RAW}}$, i.e., if the fractional reduction (improvement) in RMSE compared to the raw data is greater than any fractional reduction (deterioration) in pattern correlation (below the white lines in Fig. 3 and 4). Hence, a skillful method is required to reduce $RMSE_{\mathrm{i}}$ compared to $RMSE_{\mathrm{RAW}}$, while at the same time not deteriorating the pattern correlation too strongly. This definition of "skillfulness" thus implements the trade-off seen for some variables, such as precipitation, where a reduction in RMSE may be compensated by a deterioration of pattern correlation. Skill for SST, T2m, TXx, and TNn are similar in an absolute sense, with nRMSE_i between 0.3 and 0.6 (i.e., 30-60% errors). However, there is more improvement compared to the raw data for TNn than for the other three surface-temperature variables, due to the larger signal-to-noise ratio of TNn changes (not shown). The most skillful methods are generally similar across the 4 surface-temperature variables (i.e., methods 22, 23, 24, 25). There also tends to be a cluster of simple methods with modest but systematic improvement compared to the raw data.

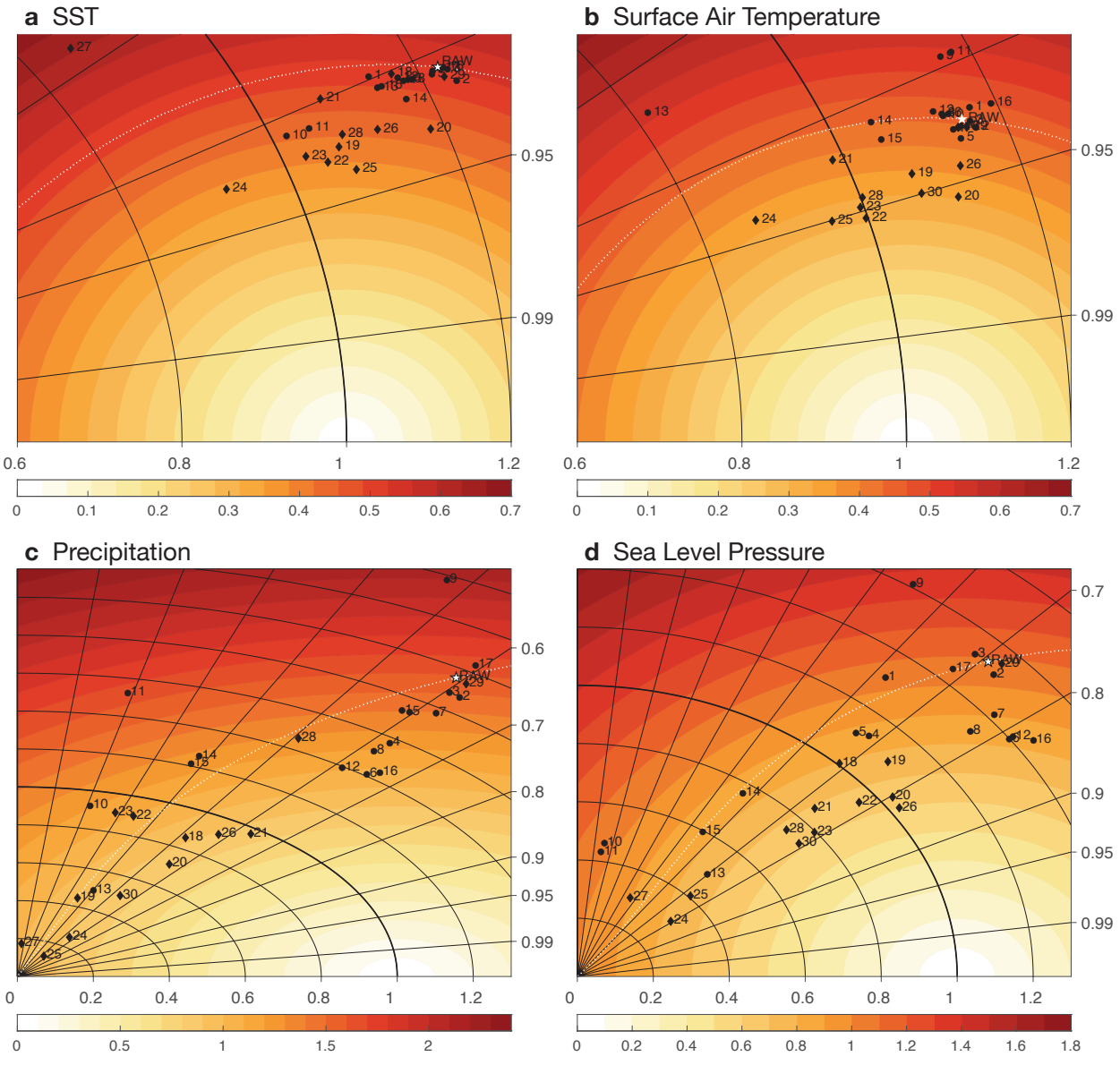


FIG. 3. Taylor diagram of method skill for 1980-2022 trends in (a) SST, (b) surface air temperature, (c) precipitation, and (d) sea level pressure. Colors, lines, and symbols as described in Fig. 1. Outlier methods excluded from the plots are: (a) 9, 30; (b) 27; (c) none; (d) none.

The absolute skill of the methods for trends in PR, SLP, and Rx1day is lower than for the four surface-temperature variables (Figs. 3c,d, 4c; cf. Figs. 3a,b, 4a,b). However, the improvement in nRMSE compared to the raw data is much larger for these variables. This occurs because there is a larger internal variability contribution to the 1980-2022 trends in these variables, and simply

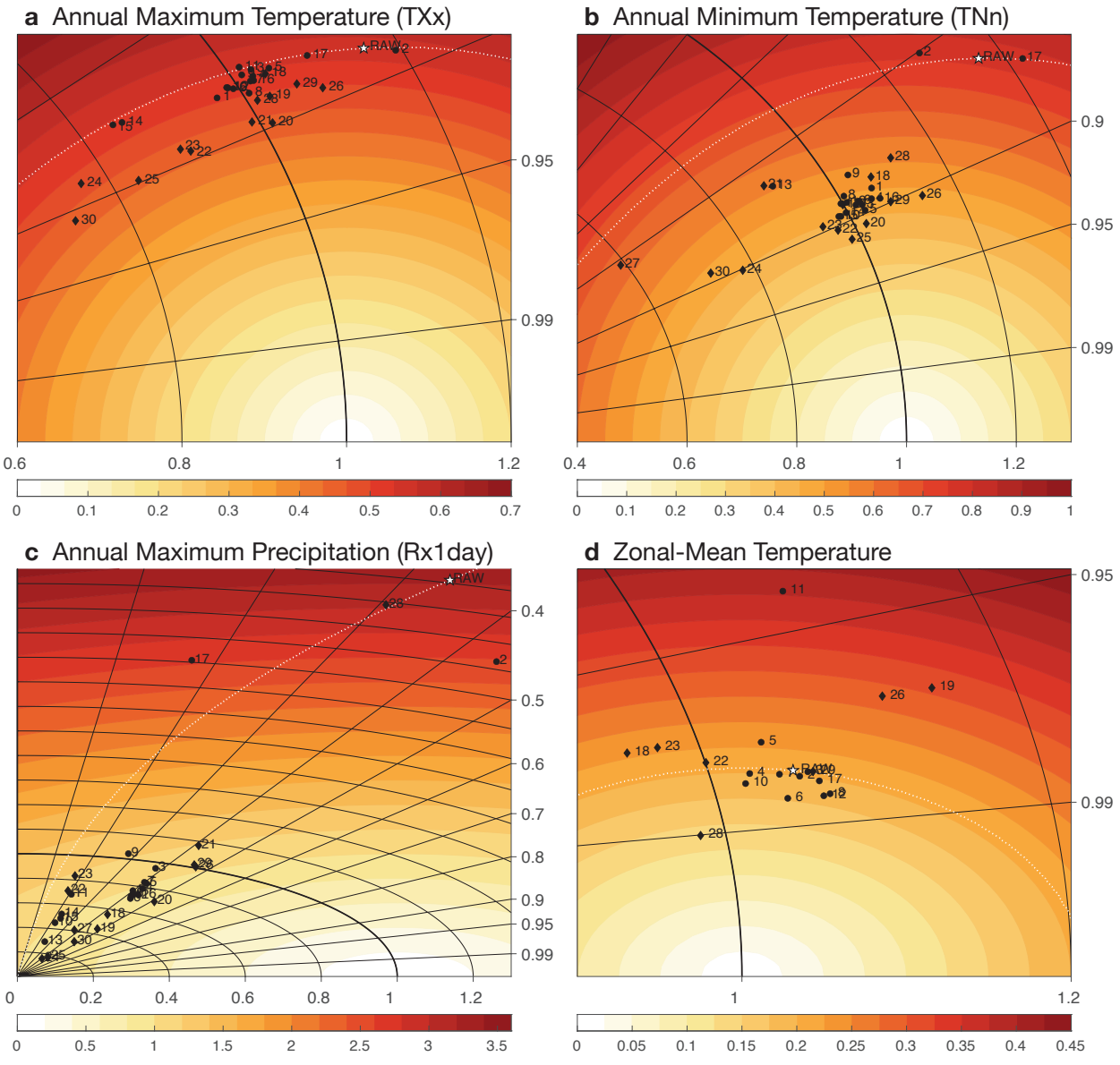


Fig. 4. Taylor diagram of method skill for 1980-2022 trends in (a) annual maximum daily maximum temperature (TXx), (b) annual minimum daily minimum temperature (TNn), (c) annual maximum daily precipitation (Rx1day), and (d) zonal-mean atmospheric temperature (zmTa). black lines, and symbols as described in Fig. 1. Outlier methods excluded from the plots are: (a) 13, 27; (b) none; (c) none; (d) 9, 14, 15, 25. Note additionally that methods 1, 13, 16, 20, 21, 24, 27, and 30 did not estimate the forced response in zmTa.

reducing the amplitude of the raw data would reduced nRMSE. Some of the ML methods (e.g., 25, 27) and one of the fingerprinting methods (24) even take the extreme approach of reducing

the estimated forced response amplitude to near zero for these variables, which does nevertheless reduce nRMSE. The ability to improve nRMSE simply by reducing the amplitude of the estimated forced trend pattern means that we should also pay attention to pattern correlation, which is not influenced by the amplitude. Several of the simple methods consistently improve pattern correlation across these variables (e.g., 6, 7, 8, 12, 16), as does one neural network method (20). Of all variables, annual-mean precipitation (PR) shows the largest number of methods that reduce the pattern correlation compared to the raw data, illustrating the difficulty in isolate the forced response for this variable.

The skill for zonal-mean atmospheric temperature (zmTa) trends is an interesting case, because here the trend in the raw data is already such a skillful estimate of the forced response ($\text{nRMSE}_{\text{RAW}} < 0.25$) that only about half the methods can improve the skill further for this variable.

Here, we have focused on 1980-2022 trends, due in part to recent literature about SST trends over this time period (e.g., Wills et al. 2022). However, we also evaluated skill for other time periods, and the skill for 1950-2022 and 2000-2022 trends in SST are compared to the skill for 1980-2022 trends in Fig. S2. Methods generally show comparable absolute skill across the three time periods, however this represents a much larger improvement compared to the raw data for the short-term trends (2000-2022). This shows that the ForceSMIP methods have even more added value for short-term trends.

To more easily compare across methods and variables, Fig. 5 shows a scorecard for the two main skill metrics, nRMSE_i and uncentered pattern correlation r_i. $1 - \text{nRMSE}_i$ is shown in place of nRMSE_i so that increased skill is positive in both panels. No single method stands out as most skillful across all variables. While the fingerprinting and ML methods that use pattern information from the training models (i.e., methods 18-30) generally stand out in terms of nRMSE, they tend to have lower pattern correlation than simple methods (especially methods 1-8, 12, and 16). The too low amplitude of some ML estimates is not apparent here, so it is important to keep in mind the Taylor diagrams as well (cf. Figs. 3 and 4). There are a number of methods that have problems with specific variables despite skill in other variables. One more general problem is the failure of dynamical mode methods (e.g., 10, 11, 13, 14, 15) applied directly to variables such as PR, SLP, and Rx1day that do not have the monthly or longer autocorrelation that generally underlies these methods. An apparently successful workaround is to apply the dynamical mode method to SST or

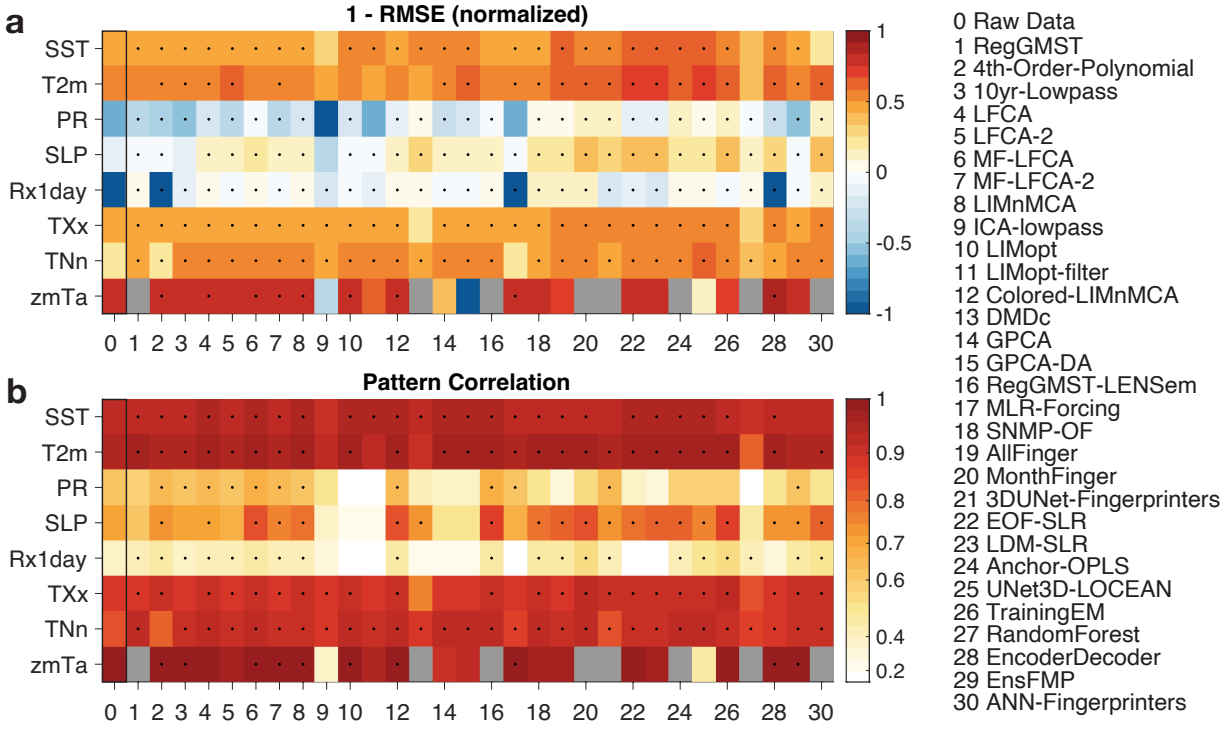


FIG. 5. Skill summary scorecards for all methods' skill in 1980-2022 trends in all variables: (a) one minus the normalized RMSE, normalized by the amplitude of the forced response, as in the Taylor diagrams; (b) the uncentered pattern correlation. The root mean square nRMSE and average uncentered pattern correlation are computed over the 5 "unseen model" evaluation members. Grey indicates that the method did not include a forced response estimate for zmTa. Stippling indicates metrics where the ForceSMIP method gives a more skillful forced trend estimate than the raw data, where the skill of estimating the forced trend by the raw data is shown on the left hand side for reference. Note that values less than -1 in (a) are cropped and the colorbar in (b) increases linearly with the square of the correlation.

another variable with large autocorrelation and then to use the covariance with other variables to get the forced response in the other variables, as was done by methods 8 and 12. Methods that stand out in terms of consistency, with a consistent skill improvement relative to the raw data (stippling in Fig. 5), are 7, 20, 24, 26, and 29, which includes the TrainingEM reference method (26). The absolute skill of methods varies based on which evaluation member they are applied to (Fig. S1), but the relative skill of methods relative to one another stays roughly the same across evaluation members.

b. Spatiotemporal variability and large-scale climate indices

The long-term trends are only one way to evaluate the forced response estimates from the ForceSMIP methods, which include full spatiotemporal variability over 1950-2022. In this section we consider their skill for the spatiotemporal variability in the forced response, both at the grid scale and in selected large-scale climate indices.

We first synthesize the ForceSMIP methods' skill for grid-scale annual-mean spatiotemporal variability. Figure 6a shows 1 – nRMSE, where nRMSE is the global-mean RMSE in the grid-

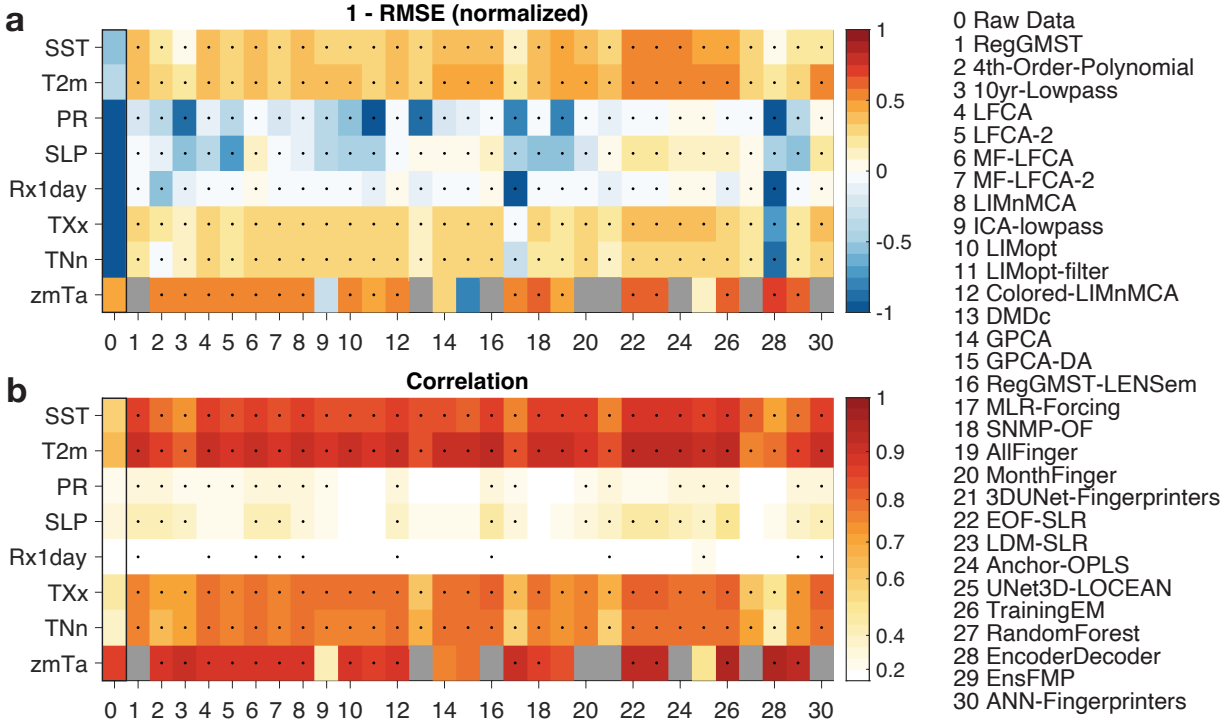
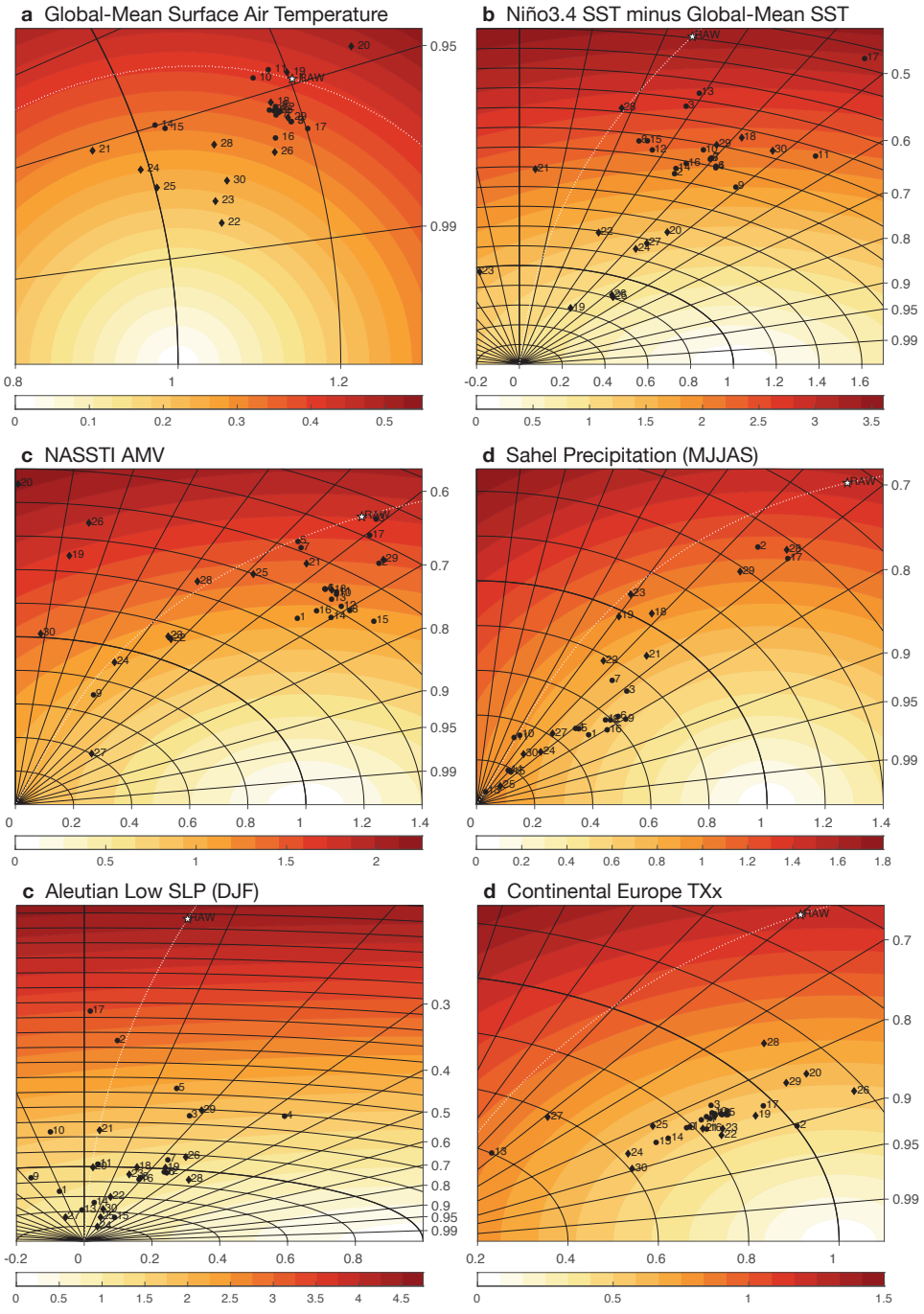


FIG. 6. Skill summary scorecards for all methods' globally averaged skill in 10-yr running-mean grid-point variability in all variables: (a) one minus the normalized RMSE, normalized by the amplitude of the forced response; (b) square-root of global-mean correlation squared. The root mean square nRMSE and average correlation are computed over the 5 "unseen model" evaluation members. Grey indicates that the method did not include a forced response estimate for zmTa. Stippling indicates metrics where the ForceSMIP method has more skill than the raw data, where the skill of estimating the forced response by the raw data is shown on the left hand side for reference. Note that values less than −1 in (a) are cropped and the colorbar in (b) increases linearly with the square of the correlation.

scale forced response estimate normalized by the global-mean root-mean-square amplitude of the true forced response, estimated by the ensemble mean of the corresponding large ensemble. Figure 6b shows the global-mean grid-point correlation of the forced response estimate and the corresponding true forced response (ensemble mean). The absolute skill in both of these skill metrics is less than the absolute skill in long-term trends (cf. Fig. 5), however, the skill added by the ForceSMIP methods compared to the raw data is larger, and there is more widespread stippling, indicating improvement relative to the raw data. All methods show consistent improvement relative to the raw data across all variables in nRMSE, with a few exceptions in zmTa. Methods 1, 6-8, 12, 16, 21, 25, 29, and 30 additionally show improvement relative to the raw data across all variables (except zmTa) in correlation. The skill of methods relative to one another is overall quite similar for the spatiotemporal variability as for the long-term trends.

To evaluate the ForceSMIP methods' skill for large-scale climate indices, we choose 6 example indices: (1) Annual-mean global-mean surface air temperature (GMST), (2) annual-mean Niño3.4 SST minus global-mean SST (GMSST), (3) the North Atlantic SST index (NASSTI) of the AMV, i.e., annual-mean SST averaged over 0-60°N, 0-80°W minus the global mean, (4) Sahel monsoon precipitation in MJJAS, averaged over 10-20°N, 20°W-10°E, (5) DJF Aleutian low SLP averaged over 30-65°N, 160°E-140°W, and (6) TXx averaged over Continental Europe (land in 40-55°N, 0-40°E). A 10-yr running-mean is applied to indices 2-5 to filter out some of the high-frequency noise, which would otherwise persist even in the ensemble mean of a large ensemble.

The skill of the ForceSMIP methods for these six large-scale indices is shown in Fig. 7. In general, there are larger and more systematic nRMSE reductions compared to the raw data than for the long-term trends in the corresponding variables (cf. Figs. 3 and 4). While there is improvement in the correlation skill compared to the raw data for almost all methods in GMST and Continental Europe TXx, there is more varied correlation skill across methods in the other four indices. However, for each index, there is a subset of methods that are substantially improving skill in terms of both nRMSE and correlation. Methods that consistently add skill compared to the raw data across all indices (3-8, 12, 14-16, 18, 22, 24, 25, and 29) include a wide range of method types, including both simple and complex methods.



533 FIG. 7. Taylor diagram showing skill for annual-mean temporal variability of climate indices: (a) GMST,
534 (b) 10-year running-mean Niño3.4 SST minus global mean SST, (c) 10-year running-mean NASSTI SST minus
535 global mean SST, (d) 10-year running-mean MJJAS Sahel precipitation, (e) 10-year running-mean DJF Aleutian
536 Low SLP, and (f) continental Europe (40-55°N, 0-40°W) TXx. Colors, lines, and symbols as described in Fig.
537 1, except with pattern nRMSE and pattern correlation replaced with nRMSE and correlation in these indices.
538 Outlier methods excluded from the plots are: (a) 13, 27, (b) 1, (c) none, (d) 20, 26, (e) none, (f) none.

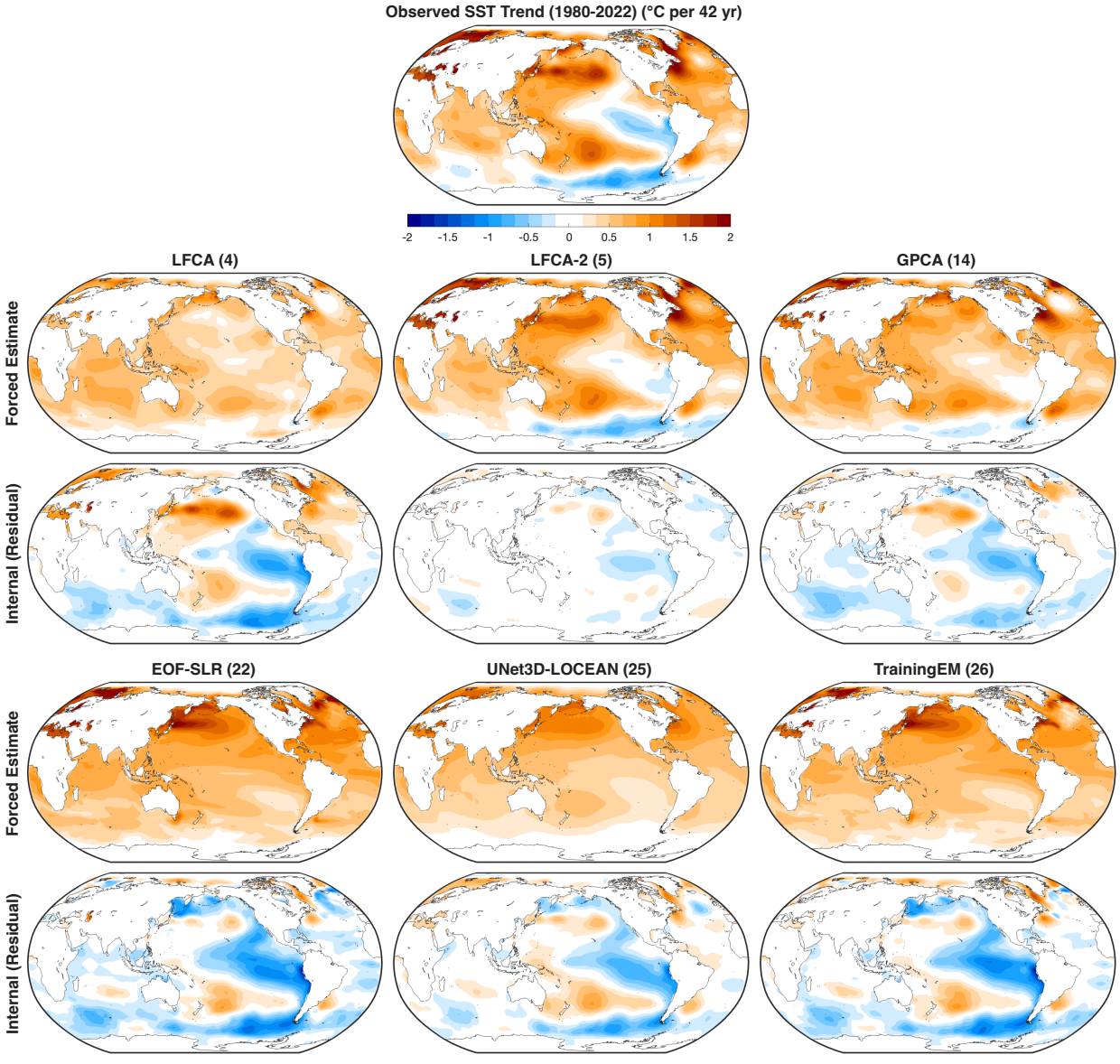
5. Estimating the Forced Response in Observations

The underlying motivation for comparing StatML methods within ForceSMIP is to improve estimates of the forced response in observations. Now, armed with knowledge about which methods are skillful for which variables and metrics, we are ready to estimate the forced response in observations.

Each ForceSMIP method was applied to observational/reanalysis data in the same way it was applied to the evaluation members used for method evaluation in the previous section. Our goal in this section is to provide some examples of the observational forced responses estimated by the ForceSMIP methods; a follow-up paper will use method weighting to generate a definitive ForceSMIP forced response estimate with uncertainties. It is illustrative to first examine the forced responses for individual skillful methods. In Figs. 8, 9, and 10, we show the forced and internal components of observed 1980-2022 trends in SST, PR, and SLP, respectively, as estimated by selected ForceSMIP methods, alongside the raw observed trends over this period. Methods are selected to illustrate the range of different forced trend estimates, based on an EOF analysis in Appendix A.

The strong pattern observed in the 1980-2022 SST trend, with cooling in the East Pacific and Southern Ocean and intensified warming in the West Pacific and North Atlantic, unlike the more uniform East-Pacific intensified warming the climate models show for this period, has generated substantial interest from the climate science community (Wills et al. 2022; Seager et al. 2022; Watanabe et al. 2024; Simpson et al. 2025). This lack of agreement with models is apparent in the comparison in Fig. 8 with the TrainingEM method (26), which is equal (up to an amplitude rescaling) to the ensemble mean of the 5 training models. The residual internal variability estimated by TrainingEM is large, and has been shown to be larger than is consistent with internal variability in most climate models (Wills et al. 2022; Seager et al. 2022).

Several of the other ForceSMIP methods shown have a smaller amplitude of estimated internal variability in 1980-2022 SST trends, indicating that they are estimating a forced response that is closer to the full observed trends than is the TrainingEM forced response. However, the degree to which individual methods' forced response estimates are more similar to the full observed trends or to the TrainingEM forced response varies substantially. LFCA-2 is one end member, estimating that almost all of the observed trend over 1980-2022 is forced. EOF-SLR is another end member, with



563 FIG. 8. Forced and internal components of observed SST trends (1980-2022) for TrainingEM and selected
564 skillful methods, chosen as representative examples from the EOF analysis in Figure A1.

580 a forced response similar to TrainingEM except for reduced El-Niño-like warming and somewhat
581 more warming in the Atlantic. GPCA and UNet3D-LOCEAN are in between these end members,
582 but each with their own unique features. The differences across these methods, all of which
583 are shown to be skillful in the method evaluation (Fig. 3a), illustrates the epistemic uncertainty
584 in estimating the forced response from observations, where epistemic uncertainty refers to the

uncertainty and potential systematic biases associated with the method used for forced response estimation. While EOF-SLR and UNet3D-LOCEAN are modestly more skillful than the other methods in the method evaluation, we cannot say with certainty which of these six forced response estimates is closer to the truth.

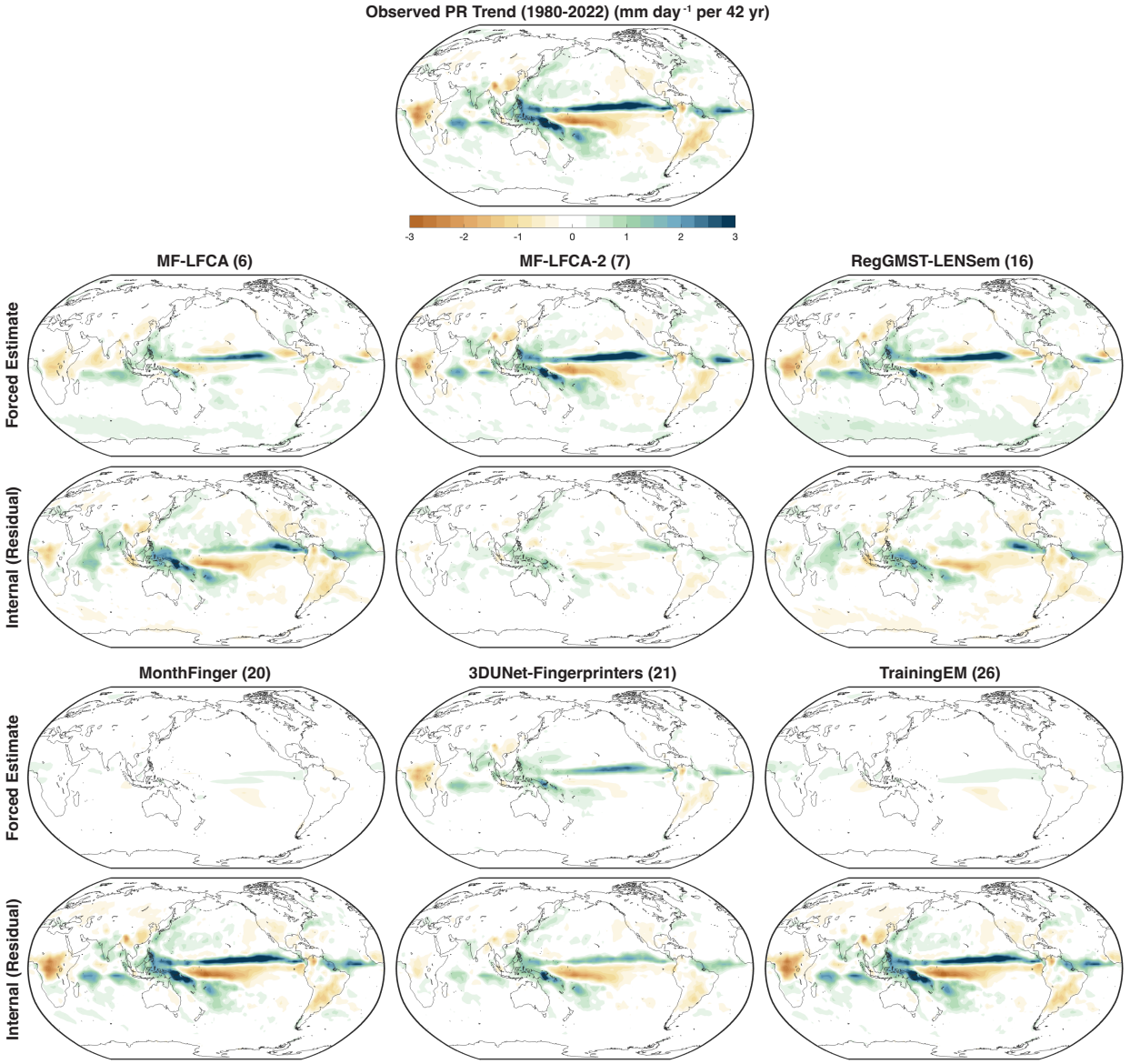


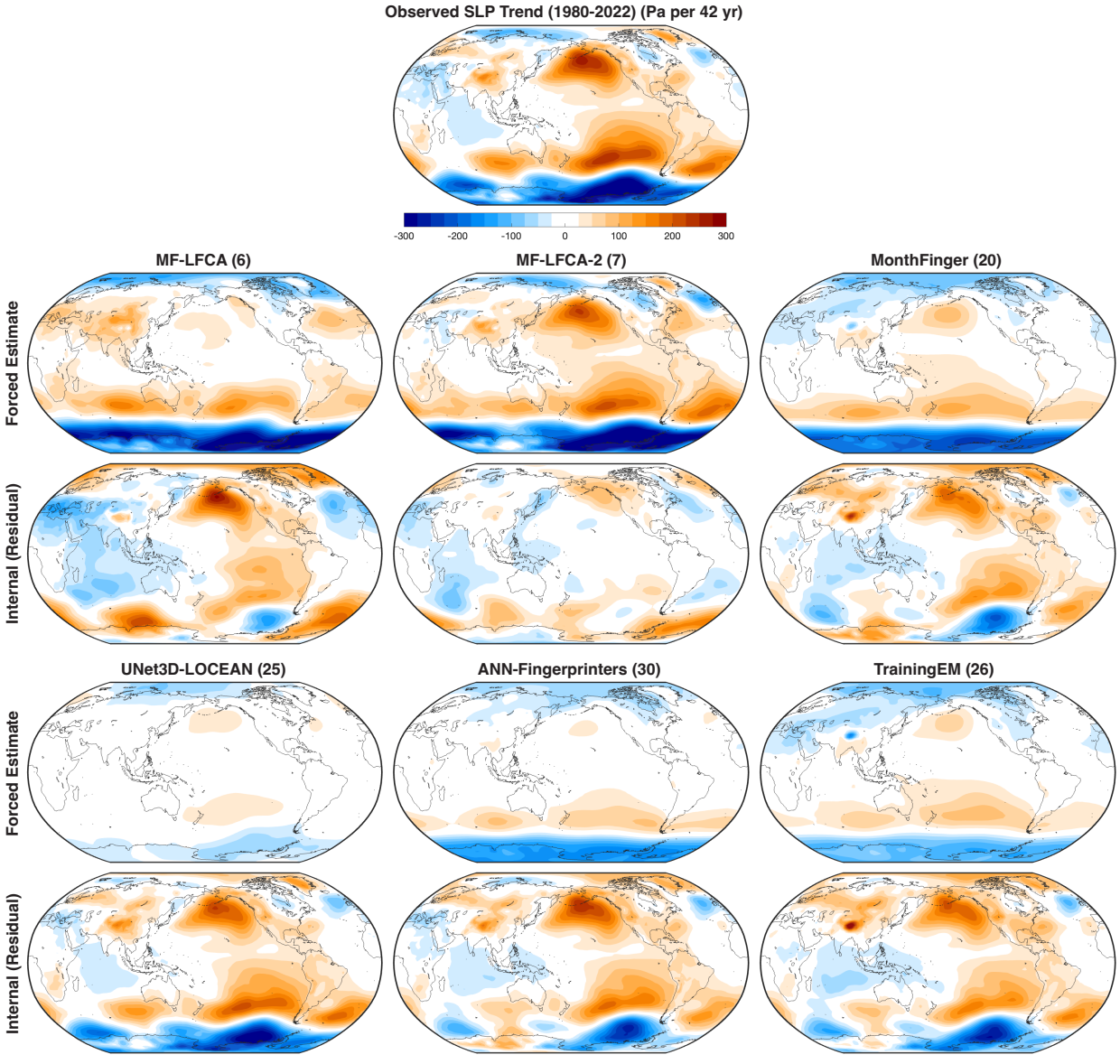
FIG. 9. Forced and internal components of observed PR trends (1980-2022) for TrainingEM and selected skillful methods, chosen as representative examples from the EOF analysis in Figure A2.

There is even wider spread of forced response estimates for precipitation (Fig. 9; see also Fig. A2), ranging from MF-LFCA-2 estimating that most of the observed 1980-2022 trend is forced to MonthFinger and TrainingEM estimating that almost none of it is. MF-LFCA and SNMP-OF are somewhere in between, with forced and internal contributions of similar amplitudes. It is worth noting that by focusing on forced responses that are robust across models, the estimated forced responses by TrainingEM and MonthFinger are smaller in amplitude than the forced precipitation response in individual models (cf. Fig. 2b), due to structural differences in models' forced responses.

The estimated 1980-2022 forced trends in SLP are all quite different from one another (Fig. 10). They agree on the poleward shift of the Southern Hemisphere westerly winds indicated by the positive and negative bands of SLP trends north and south of $\sim 50^{\circ}$S, but they have more than a factor of four spread in the magnitude of this circulation change. Some methods show that the Aleutian low weakening is mostly forced (MF-LFCA-2, consistent with the SST estimate from LFCA-2 in Fig. 8) while others show it is almost entirely internal variability (MF-LFCA, UNet3D-LOCEAN, ANN-Fingerprinters). There is a similar lack of agreement on whether North Atlantic SLP trends are forced or unforced. The large uncertainty in the forced response of SLP is consistent with the literature (Knutson and Ploshay 2021).

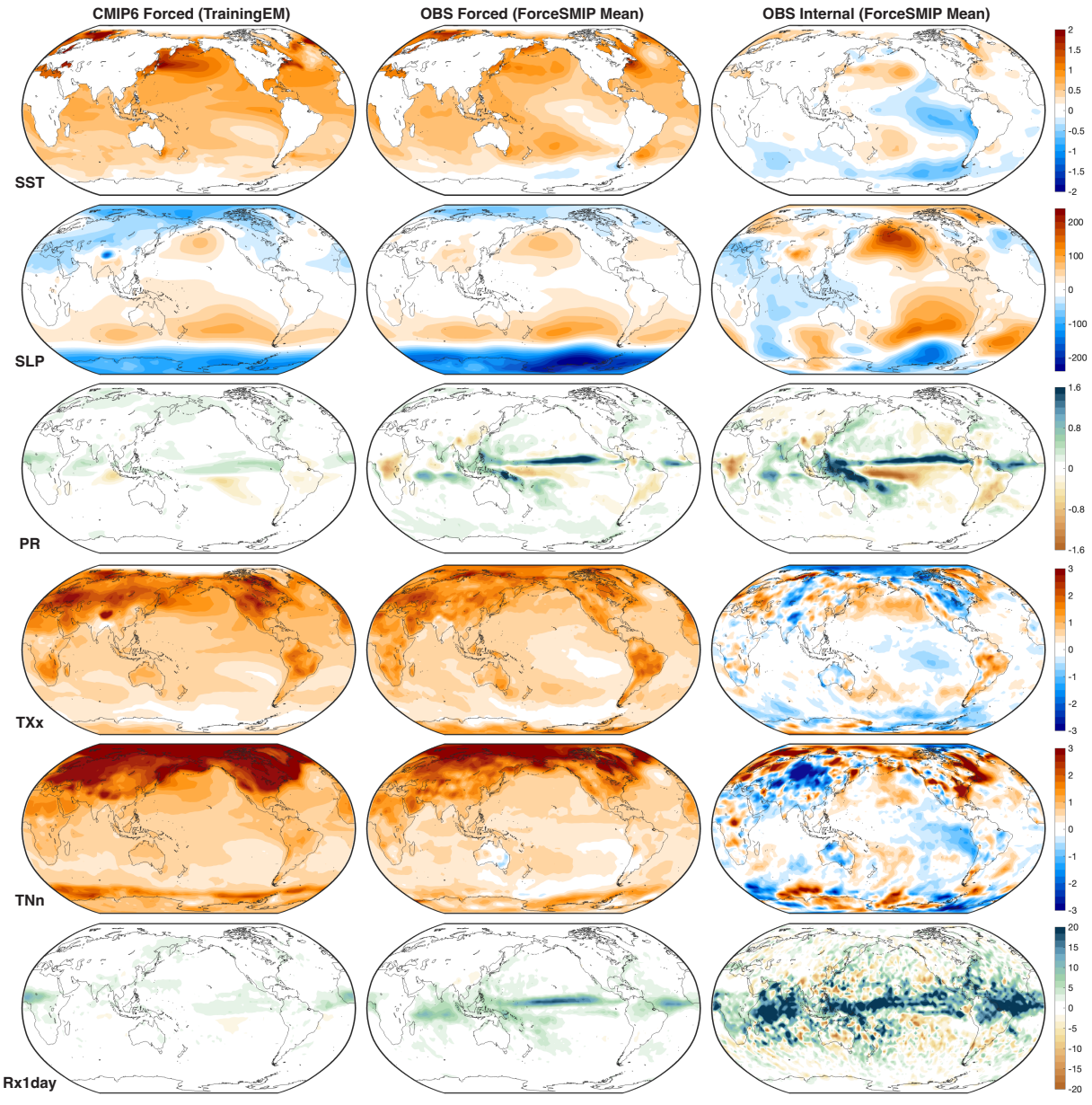
To get a sense for the average separation of 1980-2022 trends into forced and internal components by the ForceSMIP methods, we average the forced response estimates over all ForceSMIP methods determined to be skillful for each variable. Methods are included if the improvement in RMSE exceeds the deterioration of pattern correlation ($\delta\text{RMSE}_i/\text{RMSE}_{\text{RAW}} < \delta r_i / r_{\text{RAW}}$; below the white lines in Fig. 3 and 4). Figs. 11 and and S3 show the forced trend averaged over these skillful methods, the residual internal variability component of the trends, and the forced trend estimated by TrainingEM, which gives a sense of what climate models say the forced response should be over this time period.

The ForceSMIP-skillful-method-mean (hereafter ForceSMIP-mean) forced SST trend over 1980-2022 shows near-zero warming in the East Pacific and South Pacific, where the full observed SST trend shows cooling. The ForceSMIP-mean therefore attributes some but not all of the difference in 1980-2022 SST trend pattern between models and observations to internal variability. Similarly, the observed cooling of the Southern Ocean, which is not reproduced by models, is attributed



599 FIG. 10. Forced and internal components of observed SLP trends (1980-2022) for TrainingEM and selected
600 skillful methods, chosen as representative examples from the EOF analysis in Figure A3.

628 to a combination of forced response and internal variability. The ForceSMIP-mean also shows
629 stronger weakening of the Aleutian Low and stronger strengthening of the Amundsen Sea Low
630 than TrainingEM, which are both similar to La Niña teleconnections. ForceSMIP also suggests a
631 more La–Niña-like forced trend in precipitation, with a much larger amplitude than the estimate by
632 TrainingEM. However, as noted previously, the TrainingEM estimate for precipitation is smaller



618 FIG. 11. (center column) Mean estimates of forced trends (1980-2022) over all skillful ForceSMIP methods
619 (defined as $\delta\text{RMSE}_i/\text{RMSE}_{\text{RAW}} < \delta r_i / r_{\text{RAW}}$, i.e., below the white line in Figs. 3 and 4) for SST, SLP, precipitation,
620 TXx, TNn, and Rx1day. Units are °C per 42 yr, Pa per 42 yr, or mm day^{-1} per 42 yr accordingly. (right column)
621 The residual trends attributed to internal variability. (left column) The TrainingEM reference method, obtained
622 from the multi-model-mean of the five training models, is shown for comparison.

than the forced response in individual models because it focuses on the common response across all 5 training models.

The ForceSMIP-mean 1980-2022 forced trends in T2m, TXx, and TNn are broadly similar over ocean regions (Figs. 11 and S3), where they show a more La-Niña-like forced response than TrainingEM and less warming in the Kuroshio-Oyashio extension, consistent with what was found for SST. The forced trend in TXx shows more warming than the forced trend in T2m in tropical land regions and less in high-latitude land regions, whereas the opposite is true for the forced trend in TNn. This is consistent with the reduction (increase) in temperature variability in high-latitude (tropical) land regions (Kotz et al. 2021), and is also seen in TrainingEM. TXx and TNn both have larger estimated contributions of internal variability to 1980-2022 trends than does T2m, illustrating the added value of the ForceSMIP methods for noisy extreme-event statistics. Rx1day has by far the largest estimated contribution of internal variability to 1980-2022 trends, though the estimated forced response is still larger than that estimated from TrainingEM.

To visualize the ForceSMIP-estimated forced responses in the six climate indices, Figure 12 shows the likely (66%) range (i.e., the 17th and 83rd percentiles) of the ForceSMIP methods determined to be skillful compared to the raw observed data , as well as TrainingEM and five example methods. Methods are considered skillful and thus included in the likely range if they show a fractional reduction in nRMSE that exceeds any fractional reduction in their correlation (below the white lines in Fig. 7). Example methods are chosen that have varying complexity, high skill across most variables, and produce different forced response estimates from one another.

Compared to the raw data, all skillful methods smooth out some of the interannual variability in GMST (Fig. 12a). On a quantitative level, there is a 66% uncertainty range in the estimated forced 1950-2022 GMST trend of 0.89-1.07°C per 72 yr. The smoothing out of interannual variability is even more important for metrics such as Continental Europe TXx, where the forced responses estimates are all much smoother than the raw data (Fig. 12f). Methods consistently attribute the multi-year negative excursion between 1975 and 1980 to internal variability. The ratio of estimated forced trends in Continental Europe TXx and GMST has a 66% range of 1.89-2.79.

While the forced responses in GMST and Continental Europe TXx could be guessed to some degree of accuracy by simply smoothing the raw data, estimating the forced components of the other four indices is much more challenging. The ForceSMIP estimated observed forced response

in 10-yr running-mean Niño3.4 (minus GMSST) ranges from increasing (El-Niño-like warming) in Anchor-OPLS and TrainingEM to monotonically decreasing (La-Niña-like warming) in MF-LFCA and SNMP-OF (Fig. 12b), with MF-LFCA-2 even showing a strong increase through 1980 followed by a strong decrease. Nevertheless, all methods agree that the large negative excursion in the early 1970s and the large positive excursion in the early 1990s resulted from internal variability. The 66% range in the estimated 1950-2022 forced trend in Niño3.4 minus GMSST is -0.27-0.10°C per 72 yr, indicating that even the sign of the long-term forced trend remains uncertain.

The estimates of how much the AMV is forced range from almost all of it to none of it, as well as everything in between (Fig. 12c). ForceSMIP thus helps to reconcile research that indicates

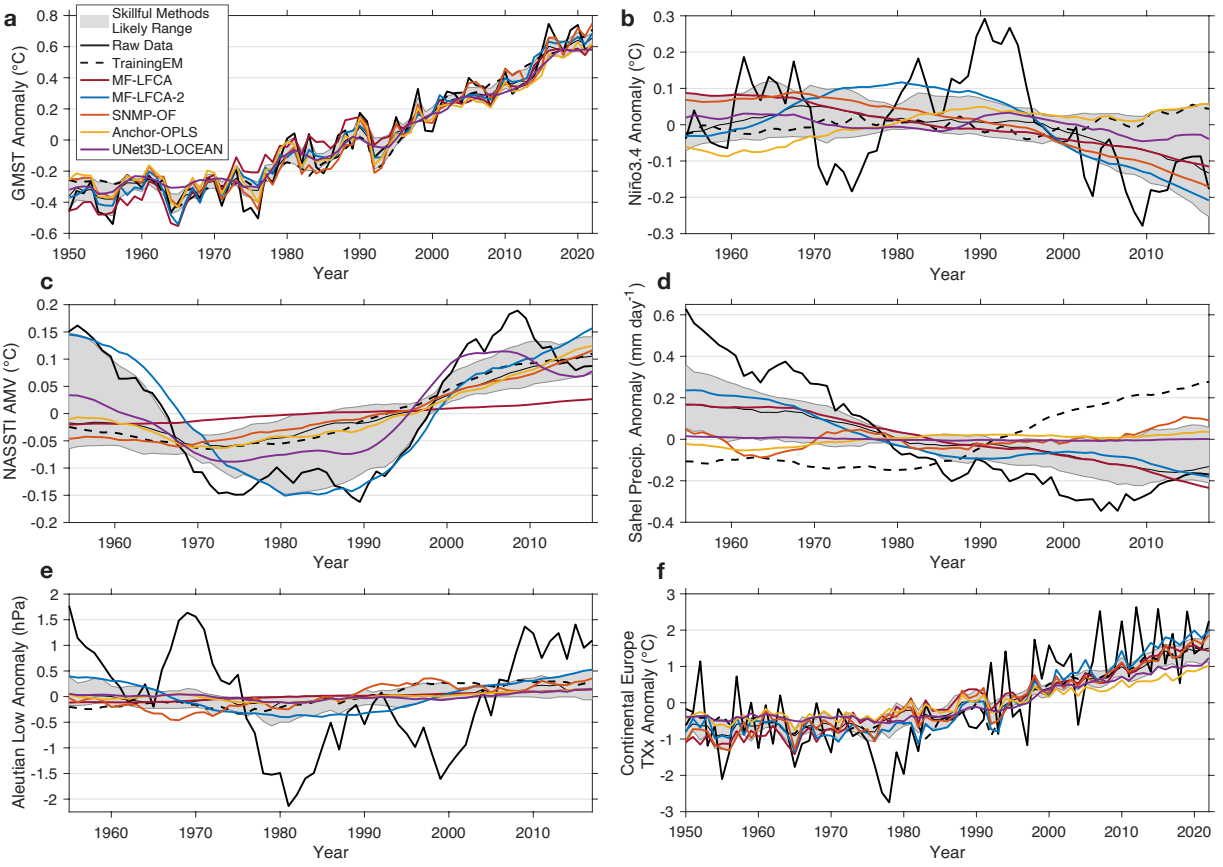


FIG. 12. Raw data timeseries, scaled training models ensemble mean (TrainingEM), skillful methods likely (66%) range, and selected methods for each index in Figure 7. Skillful methods are defined as those with a fractional reduction in nRMSE that exceeds any fractional reduction in their correlation (below the white lines in Fig. 7).

that the AMV is mostly forced (Booth et al. 2012; Wills et al. 2020; He et al. 2023) with research suggesting that it is mostly internal variability (Ting et al. 2009; Zhang et al. 2013; Qin et al. 2020; Latif et al. 2022) by demonstrating that either could be true. Interestingly, the two end members with most and least forced AMV are MF-LFCA and MF-LFCA-2, which differ only in the number of low-frequency patterns included, this illustrates how the hyperparameter sensitivity of the LFCA method may actually help to quantify the epistemic uncertainty in the forced response estimate. Given the association between the AMV and Sahel precipitation (Zhang and Delworth 2006), it is not surprising that there is also a large spread in the forced response estimates for Sahel precipitation (Fig. 12d). What is interesting however is that all of the ForceSMIP estimates either show a drying or a much weaker wettening trend than TrainingEM. This suggests that CMIP6 models, at least those used for training, have systematic biases in Sahel precipitation trends. Finally, the ForceSMIP methods consistently show a small forced response in the Aleutian Low, attributing its large decadal excursions to internal variability (Fig. 12e).

Overall, ForceSMIP provides an ensemble of estimates of the observed forced response, and we highlight cases where there are consistent differences from the forced response in climate models (e.g., the La-Niña-like forced response in observations) as well as cases where epistemic uncertainty limits the ability to draw conclusions (e.g., on the amplitude of forced AMV).

6. Conclusions, Discussion, and Outlook

We have demonstrated that many different types of StatML methods exhibit skill in estimating the forced response from individual ensemble members of a climate model large ensemble, where skill means that they give a better forced response estimate than the raw data. Skillful methods include simple regression approaches, LFCA, LIM-based methods, as well as fingerprinting and ML methods custom built for the ForceSMIP project. Methods are most skillful in absolute terms for thermodynamic responses, such as in SST and surface air temperature, but the added value of these methods compared to the raw data is largest for responses in fields with large amplitude internal variability such as SLP, precipitation, and extreme-event indices. The ForceSMIP methods are skillful for long-term regional-scale trends (e.g., over 1980-2022), grid-scale spatiotemporal variability, and large-scale climate indices. No single method outperforms the others across all variables, but rather the most skillful methods vary depending on the metric of evaluation.

Armed with an array of skillful methods for forced response estimation, we investigated the forced response in observations in Section 5. We found that the ForceSMIP methods systematically estimate that the observed forced response is more La-Niña-like than indicated by models, with a local minimum in warming in the Southeast Pacific, but also that the discrepancy in 1980-2022 SST trends between observations and models is partly due to internal variability. The observed forced response obtained from the average of skillful ForceSMIP methods also exhibits La-Niña-like teleconnections in other variables, including SLP and precipitation. Despite these commonalities, there is a large spread in the estimated forced SST trend pattern across methods that display similar skill in the large ensemble evaluation data, and an even wider spread of forced responses for SLP and precipitation. The spread across estimates of the forced response is sufficiently large that most statements about the relative contributions of external forcing and internal variability (for example to the AMV) cannot be made with great certainty. Overall, ForceSMIP suggests that there are systematic differences in the forced response between climate models and observations while also illustrating the intrinsic epistemic uncertainty in estimating the forced response from observations. The intrinsic uncertainty in the extent to which multi-decadal SST fluctuations and regional details of trend patterns are forced or unforced is important to consider in the context of climate change attribution, model evaluation, and climate impact assessments.

a. Which method should I use?

At this point, you may be wondering, which method should I use for forced response estimation in my own work? While the method evaluation in Figs. 3-7 may give some guidance, it's quite likely that this paper did not consider your metric of interest. Furthermore, there are almost always many good choices for any given metric. Nevertheless, we can give a few recommendations:

1. Use more than one type of method to get a better sense of how the forced response estimate varies across methods. It's worth keeping in mind that simple methods tend to stay closer to the observed trends, whereas most fingerprinting and ML methods will give observational forced response estimates more similar to the forced response in the climate models used for training, and will thus be more subject to any systematic biases in the training dataset.

2. Either use methods that generalize well across metrics or train/test the methods you use for your metric of interest within a large ensemble dataset. The diversity of variables and

metrics considered by ForceSMIP makes it likely that methods consistently showing skill in ForceSMIP will generalize well to other applications.

3. The ForceSMIP evaluation dataset (Wills et al. 2025) is a useful resource for evaluating new methods and/or for evaluating which methods work best for a specific application of interest.

Finally, another relevant consideration is that the ML methods would all need to be re-trained for other applications, whereas most of the other methods work out of the box and do not need further customization. However, the need to train ML methods can also be an advantage, because it means they will be tailored for the application of interest.

b. Lessons for further method development

Several lessons can be learned from the successes and failures of individual ForceSMIP methods. One of the clearest lessons is that - perhaps to no great surprise - LIMs only perform well for variables that have sufficiently large autocorrelation on the timescale of interest (monthly anomalies in our case). This is exemplified by the much higher skill of LIMnMCA and Colored-LIMnMCA compared to other LIM-based methods for variables such as precipitation, SLP, and RX1day. What's different about these two methods is that they applied a LIM to SST and then used maximum covariance analysis to identifying the covarying forced patterns in other variables. Another approach could be to merge each field variable with SST and apply a joint analysis to both fields at once. This approach was used for MF-LFCA, where it led to modest improvement in skill for precipitation and SLP over the one-field-at-time LFCA. We highlight these cases due to the clean comparisons they offer, but several other methods used multiple fields at once (Table 1). Many of the methods that analyzed one field variable at a time could likely be improved by applying them to two or more field variables at a time, especially if the additional variable is a field with a clear forced response, such as SST.

Another lesson is that methods focused on reducing RMSE or related metrics may end up guessing a near-zero forced response in cases where internal variability is larger than the forced response. To control against this, methods could expand the skill metrics they consider, for example by incorporating correlation or amplitude-error metrics and computing skill metrics on different timescales. This could draw on the experiences of the machine-learning weather prediction community (e.g., Nathaniel et al. 2024), which is grappling with similar issues. Some methods

may also give better forced response estimates if they were reformulated to explicitly estimate both forced and unforced climate variations, as was already done in UNet3D-LOCEAN (see also Po-Chedley et al. 2022).

An additional important consideration for further method development is that the ML methods are by design more trainable to optimize for a specific task. We intentionally did not specify exact evaluation targets in advance for this phase of ForceSMIP, to avoid all methods overfitting to particular metrics. Further development of these methods can now focus on correcting for some of the problems displayed in this round of evaluation. Future work should focus on cataloging a comprehensive set of forced response metrics of interest, so that methods can be trained to optimize across many relevant metrics at once.

c. An observational forced response estimate and its applications

A primary goal of ForceSMIP is to generate a forced response in observations, including a quantification of the associated epistemic uncertainty, i.e., uncertainty from different methods of estimation getting different answers. In this study, we have provided one such estimate: a 30-method ensemble of different forced response estimates (openly available on Zenodo; Wills et al. 2025). We additionally quantified the expected error based on evaluation within large ensembles and gave demonstrations of the types of information that can be obtained from such a multi-method ensemble, showing both differences in the estimated forced response across methods as well as the multi-method-mean forced response estimate for skillful methods. The method weighting is intentionally kept simple in this paper, with methods given full weight for skill above a threshold and zero weight otherwise. A follow-up paper will apply a systematic method weighting scheme, following Merrifield et al. (2023), to provide a skill weighted forced response estimate and uncertainty range. We also encourage others to generate their own forced response estimates from this dataset that are customized to specific applications.

We foresee many possible applications of an observational forced response estimate. One set of applications is for model evaluation. An observational forced response from ForceSMIP could be combined with an estimate of the residual variance due to estimation uncertainty and internal variability, e.g., based on the nRMSE evaluated in Section 4, and this would then provide a comparison point for evaluating forced trends in models against observations (cf. Simpson et al.

2025). The flip-side of evaluating forced trends in models is evaluating their amplitude of internal decadal variability, which has been suggested based on instrumental and paleoclimate data to be too weak in some regions (Laepple and Huybers 2014; Dee et al. 2017; Laepple et al. 2023). ForceSMIP can help to evaluate whether there are discrepancies in forced or internal multi-decadal variance compared to large ensembles. However, our results already suggest that, for metrics with large multi-decadal variability such as the AMV, the separation between forced and internal components remains extremely challenging, with some methods estimating a forced response more like the raw observations and some methods estimating a forced response more like the ensemble mean of the training models. In these cases, it will remain difficult to distinguish between model discrepancies in the forced response and model discrepancies internal variability.

Another set of applications of forced response estimates from ForceSMIP is for monitoring internal climate variability and generating observational large ensembles. Indices of internal variability, where the forced response is often removed either by removing the linear trend or by subtracting GMSST, can increasingly be influenced by climate change. For example, Deser and Phillips (2023b) show how not fully removing the forced response from indices of the AMV can lead to spurious implied connections with the tropical Pacific. We therefore suggest that the ForceSMIP forced response, if continuously updated, could serve as a standard estimate of the forced response to remove from indices of internal variability such as ENSO, AMV, PDO, and NAO. Removal of the forced response also allows for generation of an observational large ensemble, e.g., using the phase randomization approach of McKinnon and Deser (2018, 2021). Such an observational large ensemble can help to explore long-term trends and extreme events that could have happened in the real world under different phasing of internal variability (e.g., as in Deser and Phillips 2023a).

Underlying all of these applications of ForceSMIP observational forced response estimates is the intrinsic interest in the observational forced response itself, which can help to understand and communicate how anthropogenic activities have affected historical climate and give a glimpse into the changes expected in the near future.

Acknowledgments. This research benefited greatly from synchronous in-person hackathons in Boulder, CO and Zurich, Swizerland in August 2023, which were funded by the U.S. National Science Foundation, the Swiss National Science Foundation (Award IZSEZ0-220740), the International CLIVAR Project Office, and the Packard Foundation. R. C. J. Wills was supported by

the Swiss National Science Foundation (Award PCEFP2-203376). C. Deser and A. Phillips were supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the NSF under the Cooperative Agreement 1852977. K. A. McKinnon was supported by the Packard Foundation. S. Po-Chedley, C. Bonfils, S. Duan, and M. A. Fernandez were funded by the Regional and Global Model Analysis program area of the U.S. Department of Energy's (DOE) Office of Biological and Environmental Research (BER) as part of PCMDI, an Earth System Model Evaluation Project. Work by S. Po-Chedley, C. Bonfils, and S. Duan was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. S. Sippel acknowledges the climXtreme project funded by the German Federal Ministry of Education and Research (Phase 2, project PATTETA, Grant No. 01LP2323C) and the EU Horizon project AI4PEX (Grant agreement No. 101137682). C. Bône and G. Gastineau acknowledge the support of the EUR IPSL Climate Graduate School project managed by the ANR under the "Investissements d'avenir" programme with the reference ANR-11-IDEX-0004-17-EURE-0006. G. Camps-Valls, H. Durand, and G. Varando acknowledge funding from the European Research Council (ERC) under the ERC Synergy Grant USMILE (grant agreement 855187) and funding from the Horizon project AI4PEX (grant agreement 101137682). N. Mankovich acknowledges support from the project "Artificial Intelligence for complex systems: Brain, Earth, Climate, Society" funded by the Department of Innovation, Universities, Science, and Digital Society, code: CIPROM/2021/56. J.-R. Shi was supported by U.S. National Science Foundation under Grant OCE-2048336. The EOF-SLR and LDM-SLR methods were developed under the support of the state assignment of the Institute of Applied Physics of the Russian Academy of Sciences (Project FFUF-2022-0008). The results from UNet3D-LOCEAN were performed using HPC resources from GENCI-IDRIS AD011013295R2 and AD011013295R3. We would like to acknowledge computing support from the Casper system (https://ncar.pub/casper) provided by the NSF National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation. The authors thank all participants in the ForceSMIP hackathons for valuable discussions.

Data availability statement. The CMIP6 source data are available via ESGF, and the processed large ensemble data used in ForceSMIP has recently been made available by Maher et al. (2025). ERA5 data is available from https://cds.climate.copernicus.eu/datasets. ERSSTv5 data is available

from https://psl.noaa.gov/data/gridded/data.noaa.ersst.v5.html. The ForceSMIP Tier 1 data, i.e., the raw data, ensemble means, and estimated forced responses for each variables and each evaluation member is available on Zenodo (Wills et al. 2025). The code for all StatML methods is made available via Github (`https://github.com/ForceSMIP/tier1-methods`). Scripts for evaluating methods using the ForceSMIP Tier 1 data are made available in a separate Github repository (`https://github.com/ForceSMIP/tier1-evaluation`), including an example script for evaluating forced trends in Python and all MATLAB scripts used for analysis in this paper. Permanent identifiers for these Github repositories will be included in the final published version.

APPENDIX

Analysis of Inter-Method Variance

In order to illustrate the inter-method differences (i.e., epistemic uncertainty) in estimated forced trends, we perform an EOF analysis on the forced trends estimated by skillful methods. Methods are included if $\delta\text{RMSE}_i/\text{RMSE}_{\text{RAW}} < \delta r_i / r_{\text{RAW}}$ (below the white lines in Fig. 3). The results are shown for the EOF analysis of estimated 1980-2022 forced trends in SST, PR, and SLP in Figs. A1, A2, and A3, respectively. Panels (a) and (b) show the EOF patterns and the percentage of the variance they explain. Panels (c) show the corresponding principal components, i.e., the contribution of each EOF to the forced trend estimated by each method. The distribution of principal components are used to inform the selection of methods shown in Figs. 8-10, which are highlighted with red symbols in panels (c) of Figs. A1-A3.

Estimated 1980-2022 forced trends in SST differ from one another in a pattern (EOF1) similar to what has been called the Interdecadal Pacific Oscillation (IPO; Power et al. 1999), indicating that some methods estimate the IPO to be mostly forced, while others do not. Methods also differ in their estimates of the amount of forced warming in the Northern Hemisphere ocean basins (EOF2). The net result is that there is uncertainty in the forced SST trend in the East Pacific, Southern Ocean, Kuroshio-Oyashio Extension, and subpolar North Atlantic (Fig. A1d).

The EOF analysis for estimated 1980-2022 forced trends in PR (Fig. A2) shows a large fraction of variance explained by EOF1, which resembles the full observed trend (Fig. 9). The amplitude of PC1 shows clusters near -1 and 1.5 (Fig. A2c), which are methods estimating that very little or most of the observed trend is forced, respectively.

The leading EOF of estimated 1980-2022 forced trends in SLP (Fig. A3a) includes positive anomalies in the Aleutian low region and South Pacific and negative anomalies around Antarctic, resembling the SLP pattern associated with the IPO. Combined with EOF2 (Fig. A3b), the net result is uncertainty in the midlatitudes in all ocean basins as well as around Antarctica (Fig. A3d).

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. *Journal of Climate*, **21 (2)**, 385–402.

Bellucci, A., A. Mariotti, and S. Gualdi, 2017: The role of forcings in the twentieth-century North Atlantic multidecadal variability: The 1940–75 North Atlantic cooling case study. *Journal of Climate*, **30 (18)**, 7317–7337.

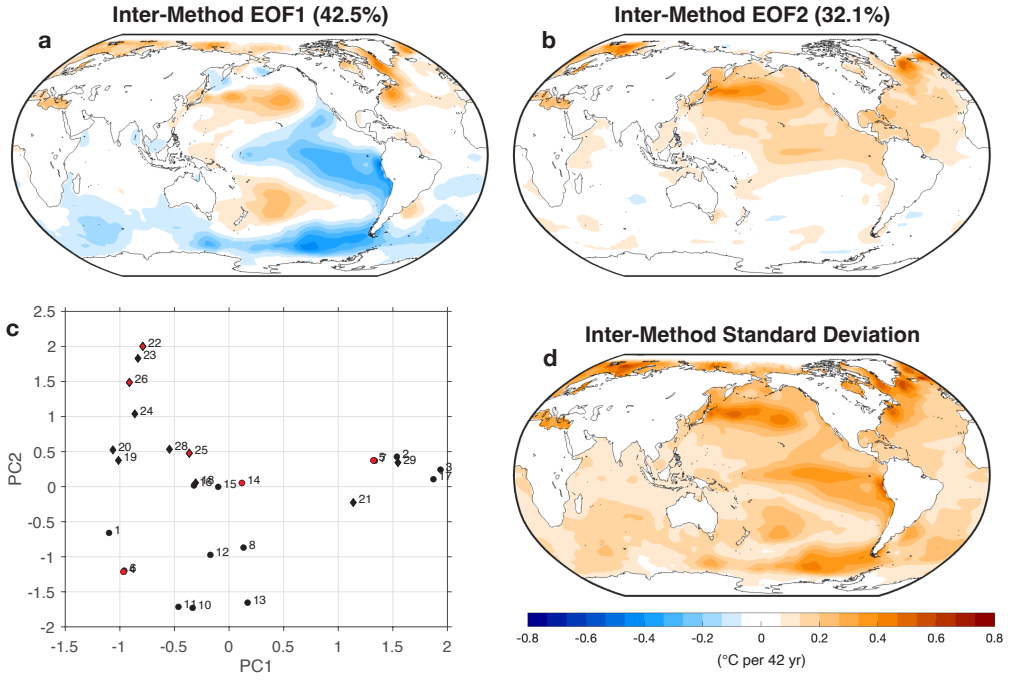


FIG. A1. Inter-method EOF analysis of estimated forced SST trends over 1980-2022, including only skillful methods (defined as $\delta\text{RMSE}/\text{RMSE}_{\text{RAW}} < \delta r / r_{\text{RAW}}$, i.e., below the white line in Figs. 3 and 4). (a) Inter-method EOF1, (b) inter-method EOF2, and (c) the principal component amplitudes for each method. The percentage of total variance explained by each EOF is shown in the title of (a) and (b). (d) Total inter-method variance, expressed as a standard deviation. Red symbols in (c) indicate methods shown in Fig. 8.

Bethke, I., and Coauthors, 2021: NorCPM1 and its contribution to CMIP6 DCPP. *Geoscientific Model Development*, **14 (11)**, 7073–7116.

Blackport, R., and J. C. Fyfe, 2022: Climate models fail to capture strengthening wintertime North Atlantic jet and impacts on Europe. *Science Advances*, **8 (45)**, eabn3112.

Bône, C., G. Gastineau, S. Thiria, P. Gallinari, and C. Mejia, 2024: Separation of internal and forced variability of climate using a U-Net. *Journal of Advances in Modeling Earth Systems*, **16 (6)**, e2023MS003 964.

Booth, B. B., N. J. Dunstone, P. R. Halloran, T. Andrews, and N. Bellouin, 2012: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature*, **484 (7393)**, 228–232.

Boucher, O., and Coauthors, 2020: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model. *Journal of Advances in Modeling Earth Systems*, **12 (7)**, e2019MS002 010, https://doi.org/doi.org/10.1029/2019MS002010.

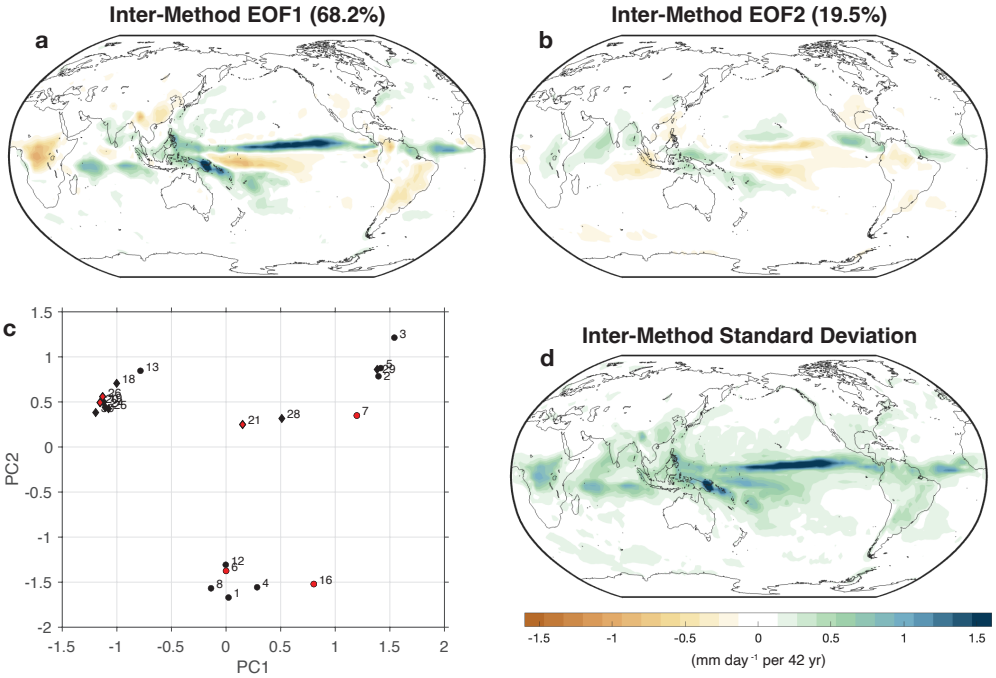
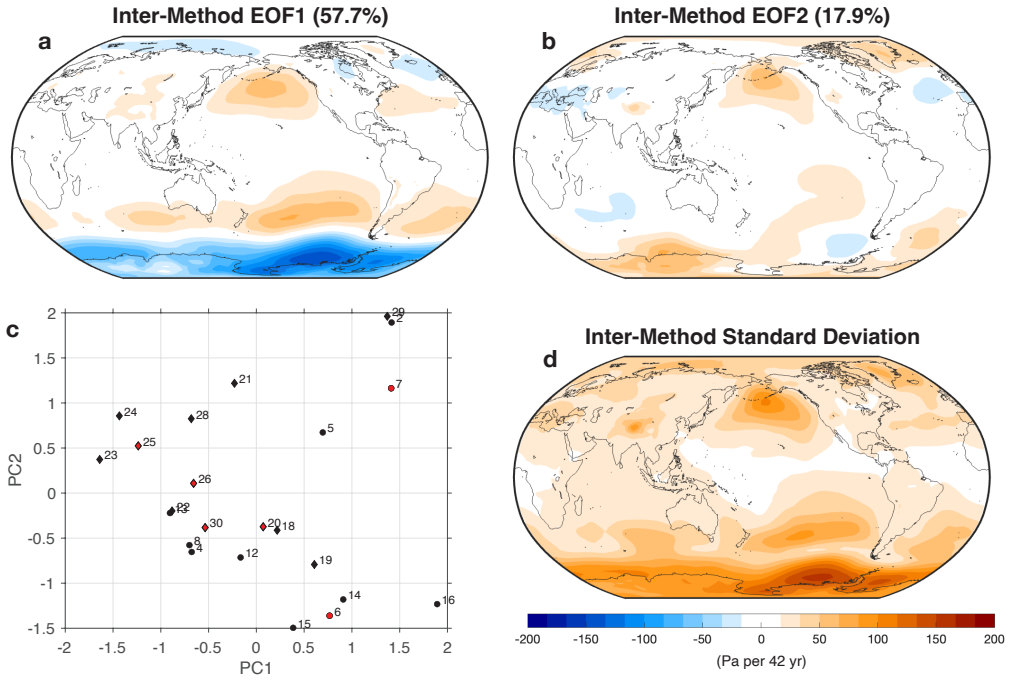


FIG. A2. Same as A1, but for estimated forced PR trends over 1980-2022. Red symbols in (c) indicate methods shown in Fig. 9.



878 FIG. A3. Same as A1, but for estimated forced SLP trends over 1980-2022. Red symbols in (c) indicate
879 methods shown in Fig. 10.

913 Brunner, L., M. Hauser, R. Lorenz, and U. Beyerle, 2020: The ETH Zurich CMIP6 next generation
914 archive: Technical documentation. Zenodo, https://doi.org/10.5281/zenodo.3734127.

915 Compo, G. P., and P. D. Sardeshmukh, 2010: Removing ENSO-related variations from the climate
916 record. *Journal of Climate*, **23 (8)**, 1957–1978.

917 Dai, A., J. C. Fyfe, S.-P. Xie, and X. Dai, 2015: Decadal modulation of global surface temperature
918 by internal climate variability. *Nature Climate Change*, **5 (6)**, 555–559.

919 Dee, S., L. Parsons, G. Loope, J. Overpeck, T. Ault, and J. Emile-Geay, 2017: Improved spectral
920 comparisons of paleoclimate models and observations via proxy system modeling: Implications
921 for multi-decadal variability. *Earth and Planetary Science Letters*, **476**, 34–46.

922 DelSole, T., M. K. Tippett, and J. Shukla, 2011: A significant component of unforced multidecadal
923 variability in the recent acceleration of global warming. *Journal of Climate*, **24 (3)**, 909–926.

Delworth, T. L., and Coauthors, 2020: SPEAR: The next generation GFDL modeling system for seasonal to multidecadal prediction and projection. *Journal of Advances in Modeling Earth Systems*, **12 (3)**, e2019MS001 895, https://doi.org/doi.org/10.1029/2019MS001895.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: the role of internal variability. *Climate dynamics*, **38**, 527–546.

Deser, C., and A. S. Phillips, 2021: Defining the internal component of Atlantic multidecadal variability in a changing climate. *Geophysical Research Letters*, **48 (22)**, e2021GL095 023.

Deser, C., and A. S. Phillips, 2023a: A range of outcomes: the combined effects of internal variability and anthropogenic forcing on regional climate trends over Europe. *Nonlinear Processes in Geophysics*, **30 (1)**, 63–84.

Deser, C., and A. S. Phillips, 2023b: Spurious Indo-Pacific connections to internal Atlantic Multidecadal variability introduced by the global temperature residual method. *Geophysical Research Letters*, **50 (3)**, e2022GL100 574.

Deser, C., A. S. Phillips, M. A. Alexander, and B. V. Smoliak, 2014: Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *Journal of Climate*, **27 (6)**, 2271–2296.

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, **10 (4)**, 277–286.

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

Frankcombe, L. M., M. H. England, M. E. Mann, and B. A. Steinman, 2015: Separating internal variability from the externally forced climate response. *Journal of Climate*, **28 (20)**, 8184–8202.

Frankignoul, C., G. Gastineau, and Y.-O. Kwon, 2017: Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic multidecadal oscillation and the Pacific decadal oscillation. *J. Climate*, **30 (24)**, 9871–9895.

Gavrilov, A., S. Kravtsov, M. Buyanova, D. Mukhin, E. Loskutov, and A. Feigin, 2024: Forced response and internal variability in ensembles of climate simulations: Identification and analysis using linear dynamical mode decomposition. *Climate Dynamics*, **62 (3)**, 1783–1810.

Gavrilov, A., S. Kravtsov, and D. Mukhin, 2020: Analysis of 20th century surface air temperature using linear dynamical modes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **30 (12)**.

Hajima, T., and Coauthors, 2020: Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks. *Geoscientific Model Development*, **13 (5)**, 2197–2244, https://doi.org/10.5194/gmd-13-2197-2020.

Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, **90 (8)**, 1095–1108.

He, C., A. C. Clement, S. M. Kramer, M. A. Cane, J. M. Klavans, T. M. Fenske, and L. N. Murphy, 2023: Tropical Atlantic multidecadal variability is dominated by external forcing. *Nature*, **622 (7983)**, 521–527.

Hegerl, G. C., H. von Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, 1996: Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *Journal of Climate*, **9 (10)**, 2281–2306.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.*, **146 (730)**, 1999–2049.

Huang, B., and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *Journal of Climate*, **30 (20)**, 8179–8205.

Hyvärinen, A., and E. Oja, 2000: Independent component analysis: Algorithms and applications. *Neural networks*, **13 (4-5)**, 411–430.

Knutson, T. R., and J. Ploshay, 2021: Sea level pressure trends: Model-based assessment of detection, attribution, and consistency with CMIP5 historical simulations. *Journal of Climate*, **34 (1)**, 327–346.

Kotz, M., L. Wenz, and A. Levermann, 2021: Footprint of greenhouse forcing in daily temperature variability. *Proceedings of the National Academy of Sciences*, **118 (32)**, e2103294 118.

Kravtsov, S., C. Grimm, and S. Gu, 2018: Global-scale multidecadal variability missing in state-of-the-art climate models. *npj Climate and Atmospheric Science*, **1 (1)**, 34.

Laepple, T., and P. Huybers, 2014: Ocean surface temperature variability: Large model–data differences at decadal and longer periods. *Proceedings of the National Academy of Sciences*, **111 (47)**, 16 682–16 687.

Laepple, T., and Coauthors, 2023: Regional but not global temperature variability underestimated by climate models at supradecadal timescales. *Nature Geoscience*, **16 (11)**, 958–966.

Latif, M., J. Sun, M. Visbeck, and M. Hadi Bordbar, 2022: Natural variability has dominated Atlantic meridional overturning circulation since 1900. *Nature Climate Change*, **12 (5)**, 455–460.

Lehner, F., C. Deser, N. Maher, J. Marotzke, E. M. Fischer, L. Brunner, R. Knutti, and E. Hawkins, 2020: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth System Dynamics*, **11 (2)**, 491–508.

Lien, J., Y.-N. Kuo, H. Ando, and S. Kido, 2025: Colored linear inverse model: A data-driven method for studying dynamical systems with temporally correlated stochasticity. *Physical Review Research*, **7 (2)**, 023 042.

Maher, N., and Coauthors, 2025: The updated Multi-Model Large Ensemble Archive and the Climate Variability Diagnostics Package: New tools for the study of climate variability and change. *EGUsphere*, **2024**, 1–28.

McKinnon, K. A., and C. Deser, 2018: Internal variability and regional climate trends in an observational large ensemble. *Journal of Climate*, **31 (17)**, 6783–6802.

McKinnon, K. A., and C. Deser, 2021: The inherent uncertainty of precipitation variability, trends, and extremes due to internal variability, with implications for Western US water resources. *Journal of Climate*, **34 (24)**, 9605–9622.

Merrifield, A. L., L. Brunner, R. Lorenz, V. Humphrey, and R. Knutti, 2023: Climate model selection by independence, performance, and spread (ClimSIPS v1. 0.1) for regional applications. *Geoscientific Model Development*, **16 (16)**, 4715–4747.

Milinski, S., N. Maher, and D. Olonscheck, 2020: How large does a large ensemble need to be? *Earth System Dynamics*, **11 (4)**, 885–901.

Nathaniel, J., Y. Qu, T. Nguyen, S. Yu, J. Busecke, A. Grover, and P. Gentine, 2024: Chaosbench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. *arXiv preprint arXiv:2402.00712*.

Olonscheck, D., M. Rugenstein, and J. Marotzke, 2020: Broad consistency between observed and simulated trends in sea surface temperature patterns. *Geophysical Research Letters*, **47 (10)**, e2019GL086 773.

Olonscheck, D., and Coauthors, 2023: The new Max Planck Institute Grand Ensemble with CMIP6 forcing and high-frequency model output. *Journal of Advances in Modeling Earth Systems*, **15 (10)**, e2023MS003 790, https://doi.org/10.1029/2023MS003790.

Oudar, T., P. J. Kushner, J. C. Fyfe, and M. Sigmond, 2018: No impact of anthropogenic aerosols on early 21st century global temperature trends in a large initial-condition ensemble. *Geophysical Research Letters*, **45 (17)**, 9245–9252.

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *Journal of climate*, **8 (8)**, 1999–2024.

Po-Chedley, S., J. T. Fasullo, N. Siler, Z. M. Labe, E. A. Barnes, C. J. Bonfils, and B. D. Santer, 2022: Internal variability and forcing influence model–satellite differences in the rate of tropical tropospheric warming. *Proceedings of the National Academy of Sciences*, **119 (47)**, e2209431 119.

Power, S., T. Casey, C. Folland, A. Colman, and V. Mehta, 1999: Inter-decadal modulation of the impact of ENSO on Australia. *Climate dynamics*, **15**, 319–324.

Proctor, J. L., S. L. Brunton, and J. N. Kutz, 2016: Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, **15 (1)**, 142–161.

Qin, M., A. Dai, and W. Hua, 2020: Quantifying contributions of internal variability and external forcing to Atlantic multidecadal variability since 1870. *Geophysical Research Letters*, **47 (22)**, e2020GL089 504.

Rader, J. K., C. Connolly, M. A. Fernandez, and E. M. Gordon, 2025: Attribution of the record-high 2023 SST using a deep-learning framework. *Environmental Research Communications*.

Rodgers, K. B., and Coauthors, 2021: Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, **12 (4)**, 1393–1411, https://doi.org/10.5194/esd-12-1393-2021.

Rugenstein, M., S. Dhame, D. Olonscheck, R. J. Wills, M. Watanabe, and R. Seager, 2023: Connecting the SST pattern problem and the hot model problem. *Geophysical Research Letters*, **50 (22)**, e2023GL105 488.

Santer, B. D., and Coauthors, 2023: Exceptional stratospheric contribution to human fingerprints on atmospheric temperature. *Proceedings of the National Academy of Sciences*, **120 (20)**, e2300758 120.

Schneider, T., and I. M. Held, 2001: Discriminants of twentieth-century changes in Earth surface temperatures. *Journal of Climate*, **14 (3)**, 249–254.

Schulzweida, U., 2023: CDO User Guide (2.3.0). Zenodo, https://doi.org/10.5281/zenodo.10020800.

Seager, R., N. Henderson, and M. Cane, 2022: Persistent discrepancies between observed and modeled trends in the tropical Pacific Ocean. *Journal of Climate*, **35 (14)**, 4571–4584.

Simpson, I. R., and Coauthors, 2025: Confronting Earth System Model trends with observations. *Science Advances*, **11 (11)**, eadt8035.

Sippel, S., N. Meinshausen, A. Merrifield, F. Lehner, A. G. Pendergrass, E. Fischer, and R. Knutti, 2019: Uncovering the forced climate response from a single ensemble member using statistical learning. *Journal of Climate*, **32 (17)**, 5677–5699.

Sippel, S., N. Meinshausen, E. Székely, E. Fischer, A. G. Pendergrass, F. Lehner, and R. Knutti, 2021: Robust detection of forced warming in the presence of potentially large climate variability. *Science Advances*, **7 (43)**, eabh4429.

Smith, D. M., and Coauthors, 2016: Role of volcanic and anthropogenic aerosols in the recent global surface warming slowdown. *Nature Climate Change*, **6 (10)**, 936–940.

Solomon, A., and M. Newman, 2012: Reconciling disparate twentieth-century Indo-Pacific ocean temperature trends in the instrumental record. *Nature Climate Change*, **2 (9)**, 691–699.

Steinman, B. A., M. E. Mann, and S. K. Miller, 2015: Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures. *Science*, **347 (6225)**, 988–991.

Stolpe, M. B., I. Medhaug, and R. Knutti, 2017: Contribution of Atlantic and Pacific multidecadal variability to twentieth-century temperature changes. *Journal of Climate*, **30 (16)**, 6279–6295.

Swart, N. C., and Coauthors, 2019: The Canadian Earth System Model version 5 (CanESM5.0.3). *Geoscientific Model Development*, **12 (11)**, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019.

Tatebe, H., and Coauthors, 2019: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, **12 (7)**, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of geophysical research: atmospheres*, **106 (D7)**, 7183–7192.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2009: Forced and internal twentieth-century SST trends in the North Atlantic. *Journal of Climate*, **22 (6)**, 1469–1481.

Trenberth, K. E., and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophysical research letters*, **33 (12)**.

Varando, G., M.-Á. Fernández-Torres, J. Muñoz-Marí, and G. Camps-Valls, 2022: Learning causal representations with Granger PCA. *UAI 2022 Workshop on Causal Representation Learning*.

Wadoux, A. M.-C., D. J. Walvoort, and D. J. Brus, 2022: An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. *Geoderma*, **405**, 115 332.

Watanabe, M., S. M. Kang, M. Collins, Y.-T. Hwang, S. McGregor, and M. F. Stuecker, 2024: Possible shift in controls of the tropical Pacific surface warming pattern. *Nature*, **630 (8016)**, 315–324.

Wills, R. C., T. Schneider, J. M. Wallace, D. S. Battisti, and D. L. Hartmann, 2018: Disentangling global warming, multidecadal variability, and El Niño in Pacific temperatures. *Geophysical Research Letters*, **45 (5)**, 2487–2496.

Wills, R. C. J., K. C. Armour, D. S. Battisti, and D. L. Hartmann, 2019: Ocean–atmosphere dynamical coupling fundamental to the Atlantic multidecadal oscillation. *Journal of Climate*, **32 (1)**, 251–272.

Wills, R. C. J., D. S. Battisti, K. C. Armour, T. Schneider, and C. Deser, 2020: Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations. *J. Climate*, **33 (20)**, 8693–8719.

Wills, R. C. J., Y. Dong, C. Proistosecu, K. C. Armour, and D. S. Battisti, 2022: Systematic climate model biases in the large-scale patterns of recent sea-surface temperature and sea-level pressure change. *Geophysical Research Letters*, **49 (17)**, e2022GL100 011.

Wills, R. C. J., A. L. Merrifield, A. Phillips, C. Deser, K. McKinnon, S. Po-Chedley, S. Sippel, and ForceSMIP Tier1 contributors, 2025: ForceSMIP Tier 1 data repository [dataset]. Zenodo, https://doi.org/10.5281/zenodo.15577520.

Wyser, K., T. Koenigk, U. Fladrich, R. Fuentes-Franco, M. P. Karami, and T. Kruschke, 2021: The SMHI Large Ensemble (SMHI-LENS) with EC-Earth3.3.1. *Geoscientific Model Development*, **14 (7)**, 4781–4796, https://doi.org/10.5194/gmd-14-4781-2021.

Xu, T., M. Newman, A. Capotondi, S. Stevenson, E. Di Lorenzo, and M. A. Alexander, 2022: An increase in marine heatwaves without significant changes in surface ocean temperature variability. *Nature Communications*, **13 (1)**, 7396.

Zhang, R., and T. L. Delworth, 2006: Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. *Geophysical research letters*, **33 (17)**.

Zhang, R., and Coauthors, 2013: Have aerosols caused the observed Atlantic multidecadal variability? *Journal of the Atmospheric Sciences*, **70 (4)**, 1135–1144.

Zhang, X., L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers, 2011: Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, **2 (6)**, 851–870.

Ziehn, T., and Coauthors, 2020: The Australian Earth System Model: ACCESS-ESM1.5. *JSHESS*, **70 (1)**, 193–214.