Supplemental Material for "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 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]Department of Statistics and Operational Research, University of Valencia, Valencia, Spain

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

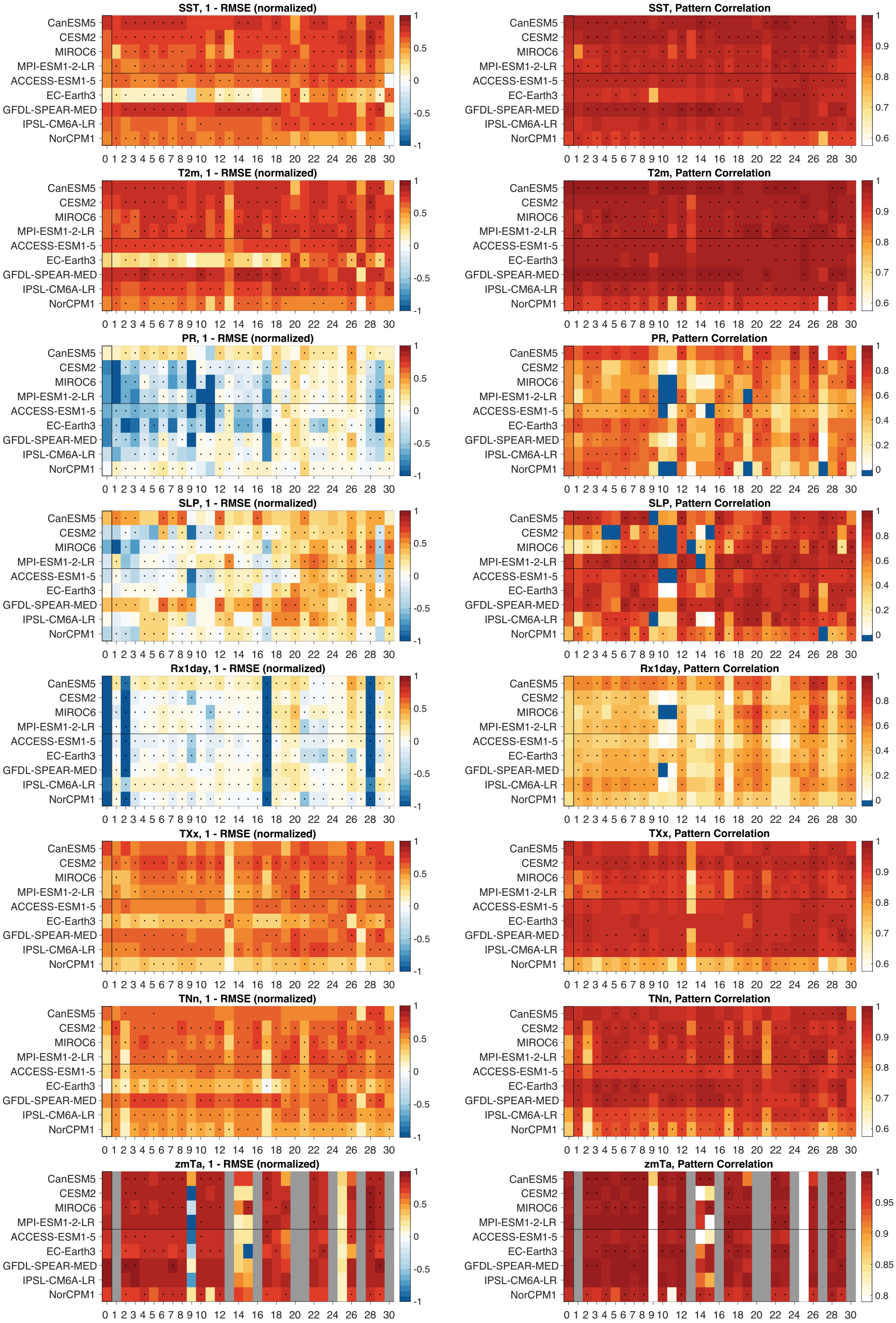

Figure S1. Skill summary scorecard for all variables and methods as in Fig. 5, but separately for each evaluation model. The horizontal line separates the training models and unseen models. A separate scorecard is shown for each variable.

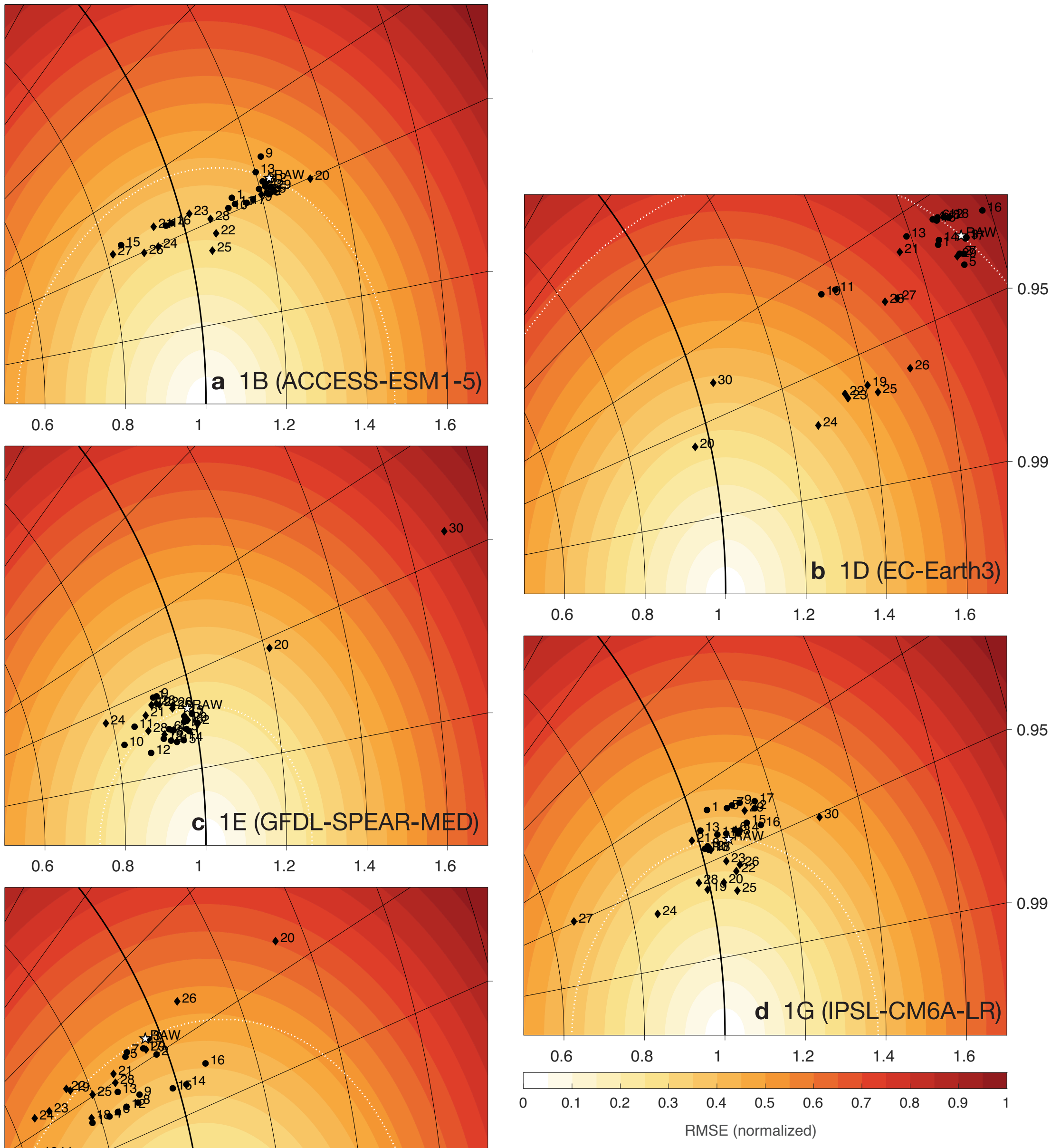

Figure S2. Taylor diagram of method skill for SST trends over 1980-2022 separately in each unseen model. Colors, lines, and symbols as described in Fig. 1. Outlier methods excluded from plots are: (a) 30, (b) 9, (c) 27, (d) none, (e) 27, 30.

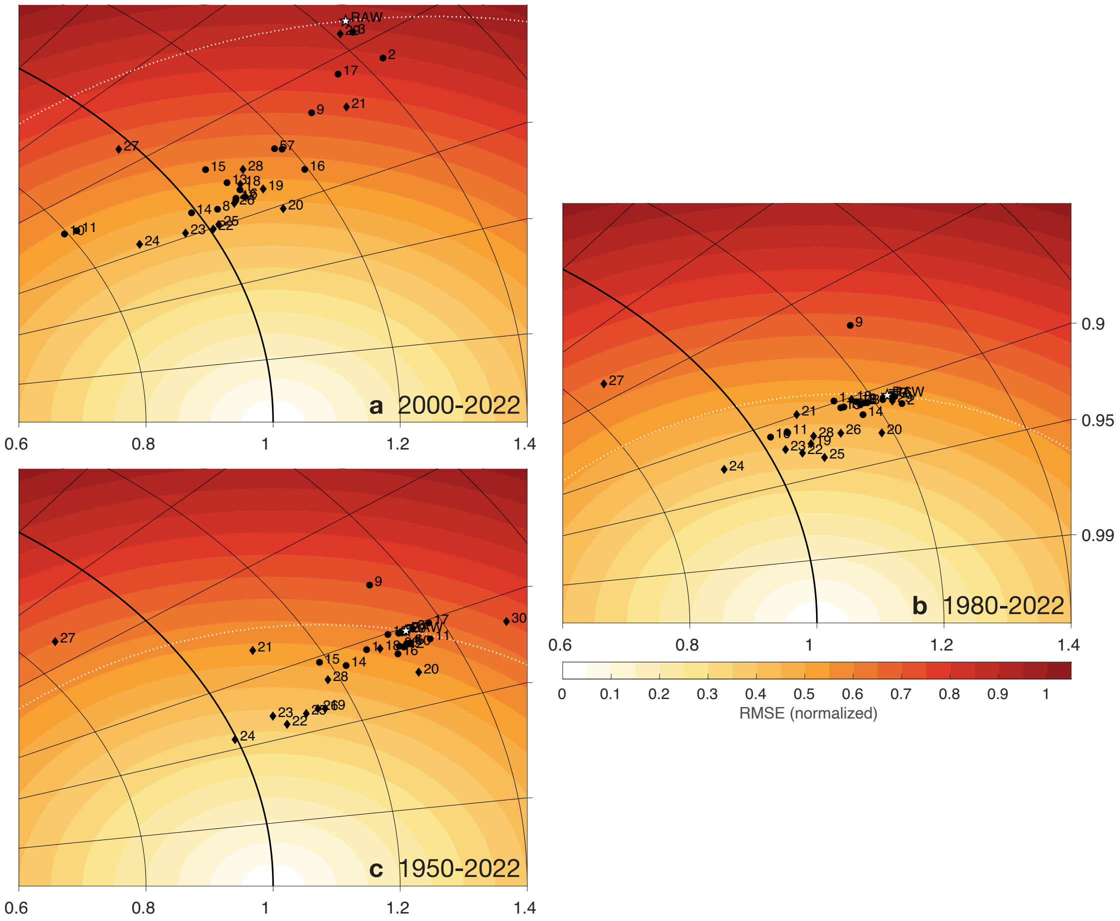

Figure S3. Taylor diagram of method skill for SST trends over (a) 2000-2022, (b) 1980-2022, and (c) 1950-2022. Colors, lines, and symbols as described in Fig. 1. Method 30 excluded from (a).

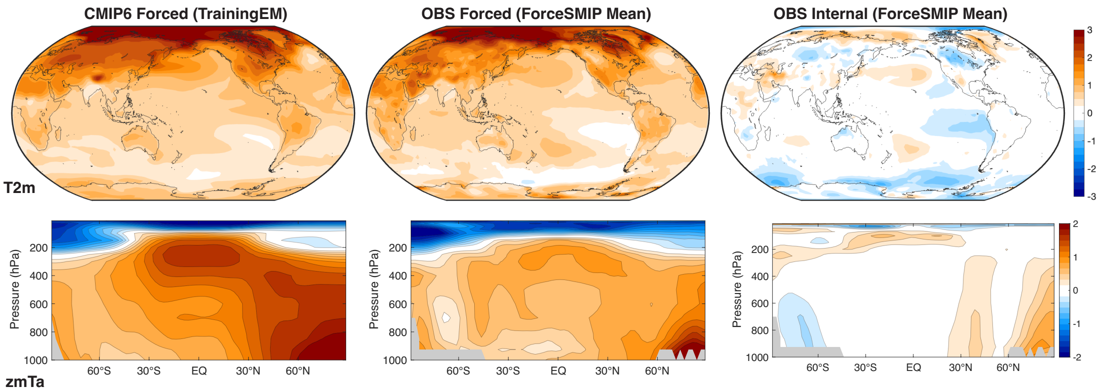
CMIP6 Forced (TrainingEM)
OBS Forced (ForceSMIP Mean)
OBS Internal (ForceSMIP Mean)
T2m
3
2
1
0
-1
-2
-3
zmTa
Pressure (hPa)
200
400
600
800
1000
60°S
30°S
EQ
30°N
60°N
2
1
0
-1
-2

Figure S4. Same as Fig. 11, but for surface air temperature (T2m) and zonal-mean atmospheric temperature (zmTa).

1 Detailed Method Descriptions

These method descriptions are meant to give an idea of the overall approach and key details of each method. Additional details and code for each method can be found at `https://github.com/ForceSMIP/tier1-methods`. For questions about individual methods, please contact either the contact author(s) listed with each method or the corresponding author Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch).

1.1 RegGMST

Contacts: Clara Deser (cdeser@ucar.edu), Adam Phillips

This method uses global-mean surface temperature (GMST) time series as a proxy for the temporal evolution of the global-mean forced response. Regressing a gridded anomaly field, for example precipitation, onto the GMST time series yields the pattern of precipitation anomalies that is linearly associated with the GMST record. However, internal variability in GMST is not removed, and this may affect the efficacy of this method (see Deser and Phillips (2023) and the related discussion in Section 1.16). This method has no tunable parameters and is thus not influenced by the training data.

1.2 4th-Order-Polynomial

Contacts: Stephen Po-Chedley (pochedley1@llnl.gov), Robert Jnglin Wills

The forced response is estimated by a 4th-order polynomial fit to the timeseries at each grid point. This method has no tunable parameters and is thus not influenced by the training data.

1.3 10yr-Lowpass

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch)

A pointwise 10-yr Lanczos lowpass filter with reflected boundary conditions (for anomalies about the linear trend) is applied to the data. The entire 1950-2022 linear trend is allowed to pass through the filter. This method has no tunable parameters and is thus not influenced by the training data.

1.4 LFCA

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch)

LFCA estimates the forced component based on the slowest evolving patterns, which are empirically determined within the evaluation dataset. This assumes that the forced response is on a longer timescale than most types of internal variability. This method errors on the side of including some internal variability, in particular decadal variability, in the forced component. This method is described in detail in Wills et al. (2020). As in that paper, we estimate the forced response based on the first N_{LFP} low-frequency patterns, combined with their timeseries.

For ForceSMIP, the method is applied to the training data for various choices of N_{LFP} and the fraction of variance included (which allows the EOF truncation to vary for different fields). Because training is only used to find the optimal value for these two hyperparameters (i.e., the fraction of variance included and N_{LFP}), this method should not overfit to the training data and should transfer well to an unseen model.

Parameters of $N_{\mathrm{LFP}} = 1$ and 70% of variance included are chosen for Tier 1, based on a subjective optimization of the global-mean spatiotemporal RMSE and squared correlation across the 8 different target fields. Based on the training, we do not expect the method to be skillful in terms of squared correlation for PR, SLP, and Rx1day (it is skillful in RMSE for those variables).

1.5 LFCA-2

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch)

Same as LFCA, but with $N_{\mathrm{LFP}} = 2$.

1.6 MF-LFCA

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch), Maren Höver

Same as LFCA, except applied to two fields at once instead of just one. The idea is that variables with low signal-to-noise ratio (PR, SLP, Rx1day) may benefit from the inclusion of a field with higher signal-to-noise ratio. For this second field, SST is used, except in the case where SST itself is the target variable, in which case T2m is used as the second field. Each field is normalized by the trace of its covariance matrix, such that each field is treated equally by the LFCA. The fields are then concatenated into a single data matrix and input into the LFCA. The methodology then follows that described in Section 1.4.

1.7 MF-LFCA-2

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch), Maren Höver

Same as MF-LFCA, but with $N_{\mathrm{LFP}} = 2$.

1.8 LIMnMCA

Contacts: Yan-Ning Kuo (yk545@cornell.edu), Justin Lien

LIMnMCA mainly targets finding the forced component. We were motivated by results during the ForceSMIP hackathon in Boulder, which showed the optimal linear inverse model (optLIM) has a great performance on capturing forced response in SST but not PR. This is the same case for LIM - we found LIM did a great job obtaining forced response of SST but not PR. We decided to further develop a method based on the trustworthy SST forced response from LIM with Maximum Covariance Analysis (MCA) to find other variables' forced responses.

Mainly following Penland and Sardeshmukh (1995) and codes provided by Xu et al. (2022), we used a linear inverse model (LIM) with lag-1 autocovariance to find the least damped mode (LDM) in SST. Then, we applied a Maximum Covariance Analysis (MCA) between the SST-LDM and the other target variables (with dimension reduced to their PCs space that explains 90% of total variance) to find the dominant covarying mode of other target variables to the SST-LDM as an estimate of their forced responses.

Before EOF analysis, the seasonality and long-term mean of all variables are removed and the data is weighted by the cosine of latitude ($\cos(lat)$). We found that the results are not sensitive to whether weighting with $\cos(lat)$ or $\sqrt{\cos(lat)}$ is used with the 2.5° lat/lon data, and suggestions and details on choices of weighting can be found in Baldwin et al. (2009). In addition, the EOFs/PCs are standardized by the square root of their eigenvalues.

1.9 ICA-lowpass

Contact: Moritz Günther (moritz.guenther@mpimet.mpg.de)

This method uses Independent Component Analysis (ICA), which is well established in other fields (Hyvärinen & Oja, 2000), but has rarely been applied to climate science. It is a method of blind source separation that aims to separate a mixture of signals into the individual source signals. In this case, the signals are the modes of internal variability and the climate change signal. To achieve this, assumptions must be made. ICA approaches the problem by assuming that each signal is independent from each other. Based on some independence criterion (e.g. Gaussianity, mutual information) one can "maximize independence" between the signals. This is achieved in an iterative process. From all the resulting signals (i.e. modes) the signal that represents the forced response must be picked.

Like the more commonly used Principal Component Analysis (PCA), ICA is a method that separates a spatiotemporal field into components. ICA finds pairs of independent components (ICs, the analog to PCs in PCA) and independent patterns (IPs, the analog to EOFs in PCA) that maximize independence. Independence is a much stronger constraint than uncorrelatedness. Independence is maximized when the sources are as far away from a Gaussian distribution as possible. This idea is rooted in the central limit theorem, which states that the mixture of two sources is always "more Gaussian" then the two sources themselves. ICA is not subject to some of the constraints of PCA. In particular the modes are not required to be orthogonal, there are no constraints on locality or non-locality, and the resulting modes are independent if ICA is applied properly.

The ICA method is preceded by PCA for dimension reduction, and will yield as many independent components as input dimensions. Therefore, the number of dimensions to retain is a critical parameter. Choosing this number too low means including unforced variability in the estimated forced response, choosing it too high means excluding part of the true forced response from the estimated forced response. We found it optimal to retain of 3 - 4 PCs for the model analysis, but this might be different for observations if they feature modes of variability that models cannot reproduce.

Specifically for the method ICA-lowpass, ICA is performed on the 10-yr (Lanczos) low-pass filtered field and the forced signal as the pattern belonging to the independent pattern with the highest global mean. The idea is that, for example, the SST pattern of global warming is expected to have a high mean value, while the SST pattern of modes of natural variability would rather tend to redistribute SST. This method is expected to work only for variables with global mean changes like SST, but was adapted for SLP by instead choosing the independent pattern with the highest mean over latitudes equatorward of ±45°. The forced response estimate is rescaled at every time step so that its global mean equals the low pass-filtered global mean of the input signal, which improves the results. Other than determining the optimal hyperparameters during tuning, the method does not rely on any information derived from climate models.

1.10 LIMopt

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch)

This method uses the machinery of Linear Inverse Models (LIMs; Penland and Sardeshmukh (1995)) to find a pattern that grows the most over a period of 8 years and uses that to define the forced response. It is therefore assuming that the forced component is the biggest contributor to long-term changes. This method has a tendency to include some internal variability with the forced response for fields with low signal-to-noise ratio. For fields with particularly low signal-to-noise ratio, it can sometimes completely miss the forced response and leave it within the internal component.

This method finds a linear combination of the included EOFs for which an initial perturbation of this form would grow into the maximum size in a period of 8 years, as determined by the LIM optimal perturbation (LIMopt) pattern. This is as described in Frankignoul et al. (2017) and Wills et al. (2020) (supplementary material), but here we use an 8-year optimal perturbation time (instead of 2.5 years). The lag time used in the LIM is 3 months, but all months are still included in the analysis. The time evolution of the LIMopt pattern is found by regression of the left singular vector associated with the LIMopt pattern (leading right singular vector) onto the monthly data. The forced component estimate is the combination of the two. Unlike LIMopt-filter, this method does not apply the LIMopt filter that was described in Frankignoul et al. (2017) and Wills et al. (2020).

1.11 LIMopt-filter

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch)

Same as LIMopt, but additionally applying an iterative filter that is meant to capture the subsequent evolution of the LIMopt pattern over a time window and not just the initial perturbation, providing some additional smoothing.

Starting from the identified LIMopt pattern used in LIMopt (they are run in the same script), this method applies the iterative filter described in Frankignoul et al. (2017) and Wills et al. (2020) (supplementary material). This projects the LIMopt pattern only onto departures from the expected evolution of the LIMopt pattern from previous timesteps within a specified window period. Here, we apply the pattern and filter only to 3-month averages centered at Feb, May, Aug, and Nov and linearly interpolate the neighboring months from the resulting values at these months. We use a window period of only 12 months (as opposed to 20 years in Frankignoul et al. (2017)), because we saw that there was a tendency for the algorithm to diverge for longer window periods, dramatically reducing the skill.

1.12 Colored-LIMnMCA

Contacts: Yan-Ning Kuo (yk545@cornell.edu), Justin Lien

Colored-LIMnMCA mainly targets finding the forced component. The forced response can be described as an outcome of a system in which the forcings (e.g., anthropogenic emissions of GHGs, aerosols, etc) are the main perturbation. Such a system resembles a red noise-driven system with a distinct power spectrum biased in the low frequencies. Lien et al. (2025) developed a colored noise Linear Inverse Model (Colored-LIM) and specifically aimed to solve the red noise-driven system dynamics (compared to LIM, which is a white noise-driven system).

However, key limitations of using Colored-LIM are: 1) The target variable (X) needs to have red noise characteristics (e.g., the power spectrum has stronger signal at the low frequencies). 2) The target variable (X) has fewer high-frequency/white noise-like signals so that the solution L can be stable (i.e., each eigenvalue has a negative real part). To satisfy this condition, low-pass filtering for input data is necessary for solving L with Colored-LIM.

SST is the only variable in ForceSMIP's target that can fairly satisfy the first condition (other atmospheric variables are more white-noise-like). Therefore, we used Colored-LIM to solve the least damped mode (LDM) in SST to represent the most unstable long-term red-noise-driven signal (forced response). Then, we used Maximum Covariance Analysis (MCA) to find other target variables' covarying patterns to the LDM from SST.

The workflow is described as follows: We first reduced the dimension of the seasonal cycle removed SST to its principal components (PCs) space (nPC = 9); then, we applied the Lanczos filter (Duchon, 1979) to separate the PCs' signals into high- and low- frequencies (cutoff = 1 yr). Assuming the low-frequency variability (> 1 yr) has a dependence time over 1 year (12 months), we used a Colored-LIM to solve LDM from the low-frequency, red noise-driven signals in the PC space. For the same PC space's high-frequency variability (< 1 yr), we used a traditional LIM with lag-1 month autocovariance for obtaining its LDM from these high frequency- white-noise-like behavior. We use a linear combination of two LDMs from Colored-LIM and LIM as the total LDM, and reconstruct the estimated forced signal of SST. Then, we applied a maximum covariance analysis (MCA) on the total LDM (from SST) and the other target variables (e.g., T2m, SLP, PR, etc) in their PCs space (truncated PCs with 90% of the total explained variance) to get the maxima covarying mode to LDM of SST. Finally, we reconstruct the target variable's leading covarying mode to LDM of SST as their forced responses.

Note: during the data training process, we also compared the ColoredLIM method against LIM. There is no significant distinction between the performance of the two approaches (for some models' ensemble members, Colored-LIM performed better on capturing the long-term trend (compared to ensemble mean from that model), while for the other models LIM works better. The LIM is already good enough to detect the forced response in SST.

1.13 DMDc

Contacts: Nathan Mankovich (nathan.mankovich@gmail.com), Gherardo Varando, Homer Durand, Gustau Camps-Valls

Dynamic mode decomposition with control (DMDc) estimates the forced component using the least damped component of a LIM plus the linear contribution of the forcing. Although it is certainly possible to use training data to optimize the number of PCs, spatial modes, and control signals, we choose to not use the training data to tune parameters for the Tier 1 submission.

Dynamic mode decomposition (DMD; Schmid (2010)) is essentially a linear inverse model (LIM; Penland and Sardeshmukh (1995)). The least damped mode of a LIM has been used in climate science to predict the forced response (Frankignoul et al., 2017; Xu et al., 2022). Dynamic mode decomposition with control (DMDc; Proctor et al. (2016)) is a variant of DMD (LIM) which includes a forcing component and has been used to analyze forced response patterns across climate scenarios (Mankovich et al., 2025). It assumes that *the signal at time t is a linear function of the signal and forcing at time $t-1$*. Namely $\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{z}_{t-1}$, where $\mathbf{x}_t$ is the climate signal and $\mathbf{z}_t$ is the forcing at time t. The forcing $\mathbf{z}_t$ is interpolated to monthly resolution from the annual-mean total effective radiative forcing from the IPCC AR6 Chapter 7 (Forster et al. (2021); https://raw.githubusercontent.com/IPCC-WG1/Chapter-7/main/data_output/AR6_ERF_1750-2019.csv). Due to data volume, we run DMDc on the principal components (PCs) of the climate signal rather than the climate signal itself. The 73 years of effective radiative forcing data are counted as 73 "tunable parameters", in addition to the method's 2 hyperparameters.

We perform 3 steps to use DMDc to predict the first response. (1) We take the first principal components (PCs) that explain 30% of the variance of the anomalies. (2) We run DMDc using the package pydmd (Demo et al., 2018; Ichinaga et al., 2024) with a time lag of 1 month on these PCs, controlling for the total radiative forcing. We reconstruct $\mathbf{A}$ using the least damped mode (associated with the eigenvalue with the largest real part) and call it $\hat{\mathbf{A}}$. Then we reconstruct the PCs using the DMDc model with $\hat{\mathbf{A}}$ and $\mathbf{B}$. (3) Finally, we estimate the forced response as the reconstruction of the anomaly signal from the output of (2). This pipeline should effectively predict the forced response in multiple ways. First, we know the reconstructed signal from the first principal component is close to the forced response, so we take the first few principal components (the ones that explain 30% of the variance). Next, DMDc assumes that there is a fixed linear map for the forcing contribution over all time steps ($\mathbf{B}\mathbf{z}_{t-1}$). This could predict the forced response well if the forcing contribution is linear. Since this is likely not the case, we include the signal from the least damped mode in our prediction. In summary, we add this linear forcing contribution with the standard 'LIM forced response prediction' using the least damped mode to predict the forced response.

1.14 GPCA

Contacts: Gherardo Varando (gherardo.varando@uv.es), Homer Durand, Nathan Mankovich, Gustau Camps-Valls

Granger-rotated principal component analysis (GPCA) is an unsupervised method which estimates the forced component by employing the forcing time series signal. The training data is used only for hyperparameters calibration (even if it was not done automatically and optimally now). The method is inspired by the idea that under some assumption on the dynamical-system we could recover the causal effect of an exogenous signal.

The GPCA approach estimates the "direct Granger effect" of the exogenous forcing signal which is obtained by rotating the PCA/EOF decomposition to align the first(s) mode to be Granger-effects of the forcings. In particular, the algorithm is a small extension of the method proposed in Varando et al. (2022).

The underlying assumption is that for a subspace defined by the leading EOFs (X) the system is a vector autoregressive process $X(t) = AX(t-1) + BY(t - lags) + CZ(t-1)$ where X are the first PCAs of the variable of interest (e.g. T2m), Y is the exogenous forcing signal (expanded to non-linear terms) and Z are additional control variables (e.g., some PCs of SLP used as an estimation of the internal variability, which are only used in GPCA-DA). The forcing timeseries Y is interpolated to monthly resolution from the annual-mean total effective radiative forcing from the IPCC AR6 Chapter 7 (Forster et al. (2021); https://raw.githubusercontent.com/IPCC-WG1/Chapter-7/main/data_output/AR6_ERF_1750-2019.csv). The 73 years of effective radiative forcing data are counted as 73 "tunable parameters", in addition to the method's 15 hyperparameters.

After disentangling the "direct Granger effect" of the forcing series on the variable of interest (e.g. T2m), we estimate the full forced response as a forward reconstruction of the above vector autoregressive process by zeroing the internal variability contribution (Z) and where A and B are estimated, partially, with an elastic-net penalty (linear combination of L_1 and L_2).

1.15 GPCA-DA

Contacts: Gherardo Varando (gherardo.varando@uv.es), Homer Durand, Nathan Mankovich, Gustau Camps-Valls

Same as GPCA with additional dynamical adjustment via using some EOFs of SLP as control variables.

1.16 RegGMST-LENSem

Contacts: Clara Deser (cdeser@ucar.edu), Adam Phillips

This method uses the ensemble-mean global-mean surface temperature time series from a model large ensemble as a proxy for the temporal evolution of the global-mean forced response. By regressing gridded observational fields onto this time series, we hope to estimate the forced response. By subtracting this estimated forced response from the original data, we obtain the unforced component. The advantages of the method are that is it extremely simple and easy to implement. The disadvantage is that the pattern (not the amplitude) of the forced response is assumed to be stationary in time (e.g., over the time interval used for the regression analysis). However, the pattern of the forced response may evolve over time due to slow oceanic adjustment processes as well as the contribution from regional forcing agents such as anthropogenic aerosols. As such, the method probably underestimates the forced response. The method translates directly to an unseen model simulation and observations, because it recomputes the forced response pattern directly within the new dataset based only on the timing of the forced response in the training data.

The so-called “global temperature residual” method, which employs linear regression analysis to remove variability at each location that is linearly congruent with variability in global-mean temperature ($G(t)$), has been widely used in the literature (Dai et al., 2015; Zhang et al., 2019; Yan et al., 2019; Deser & Phillips, 2021, 2023). The rationale behind this approach is that variations in $G(t)$ are mainly a reflection of external radiative forcing, making it a convenient metric for tracking the temporal evolution of forced climate change. However, as shown by Deser and Phillips (2023), the internal component of $G(t)$ is non-negligible and is associated with a distinctive pattern known as the “Interdecadal Pacific Oscillation (IPO)”. This association has been highlighted in the literature on the global surface warming hiatus of the early 21st century (e.g., Kosaka and Xie (2013)). The internal contribution to $G(t)$ can be removed by using the ensemble-mean $G(t)$ from a model Large Ensemble in place of the observed $G(t)$ or a single model simulation of $G(t)$. This modified method (the ”ensemble-mean global temperature residual method”) was introduced by Deser and Phillips (2023) with application to internal Atlantic Multi-decadal Variability, and tested within the framework of model large ensembles. It is important to note that this method does not account for changes in the pattern of the forced response over time. Further, the method assumes that external forcing projects entirely on global-mean temperature, whereas other forcing agents such as anthropogenic aerosols may impart inter-hemispheric asymmetries in the forced temperature response.

In its implementation for ForceSMIP, $G(t)$ is taken from the ensemble-mean timeseries of global-mean surface air temperature from the 50 CESM2 large ensemble members in the training data. This timeseries is 73 years long and has monthly resolution, and we thus count this as 876 “tunable parameters”.

1.17 MLR-Forcing

Contact: Stephen Po-Chedley (pochedley1@llnl.gov)

The main assumption this method makes is that the local forced response scales with global mean indices of a subset of climate forcings (regional aerosol emissions, and the global mean volcanic, solar, and greenhouse gas DAMIP time series of surface air temperature). Since the response to volcanoes can vary depending on eruption location, we treated each volcanic time series as a separate predictor. In general, this method appears to capture most of the temporal variance in regional and global spatial averages but has biased trends over sub-periods. For example, the method generally overestimates (too positive) the unforced response toward the end of the time series (e.g., the satellite era), which leads to a trend bias. Note that de-trended Niño3.4 T2m predictors were included in the regression model (to try to explicitly account for variability “noise” in the forced estimate). After fitting the regression model, these predictors were not used to produce the forced response estimate.

We first constructed a forced predictor matrix that included a greenhouse gas, solar, and volcanic term. These terms were derived from the surface air temperature timeseries from DAMIP (Gillett et al., 2016). Each term was smoothed using a Savitzky-Golay filter. The volcanic time series were broken into six individual time series for the eruptions of Krakatoa, Santa Maria, Novarupta, Mt. Agung, El Chichon, Mt. Pinatubo. In each case the values were zero, except for the four years following the volcanic eruption. We also included time series of sulphate aerosol emissions (Hoesly et al., 2018) over North America, Europe, East Asia, South Asia, and Africa (regions defined in Wilcox et al., (2023)). Last we de-trended the Nino 3.4 time series (derived from tas data) using the two lowest frequency modes from ensemble empirical mode decomposition and used this de-trended time series as a predictor at 3- and 6-month lead time (https://pyemd.readthedocs.io/en/latest/eemd.html). If predictors did not extend to 2022, they were linearly extrapolated through December 2022 using the last ten years of predictor data. In all, this results in 15 predictor time series (6 volcanic, 5 aerosol, 1 GHG, 1 solar, and 2 Niño3.4 time series).

At each grid point we fit the anomaly time series of the variable of interest with these predictors on a monthly basis: $Y(t) = F(t,x) \cdot B(x) + N34(t-3) \cdot C + N34(t-6) \cdot D + E$. The forced response was then estimated by dropping the Niño3.4 terms: $Y(t) = F(t,x) \cdot B(x) + E$.

Two of the 15 time series are Niño3.4 time series at different lags, which are derived from the raw data. There is additionally 1 tunable parameter to apply EEMD smoothing and 1 parameter to select the lag. The number of tunable parameters is therefore $13 \times n_{\text{time}} + 4$, where $n_{\text{time}} = 876$ months.

1.18 SNMP-OF

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch), Maren Höver

This method estimates the forced component. It learns from climate models which grid points have the largest signal-to-noise ratio and then looks at the signal at those points in the new dataset, but it calculates the forced response pattern from the new dataset, so it will generalize well as long as the regions with largest signal-to-noise ratio are similar between the datasets.

This method combines signal-to-noise maximizing pattern (SNMP) analysis with optimal fingerprinting (OF). SNMP is used as described in Wills et al. (2020), except it is applied to all 5 training models simultaneously instead of one model at a time. The fingerprint patterns obtained from this analysis (termed $\mathbf{u}_k$ in Wills et al. (2020)), which contain information about regions that have the highest signal-to-noise ratio, are then projected onto (matrix multiplied with) the data matrix $\mathbf{X}_{\text{eval}}$ from the evaluation member of interest, generating timeseries $\mathbf{t}_{\text{k,eval}}$ of each SNMP in the new dataset. These $\mathbf{t}_{\text{k,eval}}$ then need to be renormalized before the final step, which is to matrix multiply the $\mathbf{t}_{\text{k,eval}}$ one more time $\mathbf{X}_{\text{eval}}$ to get the forced response pattern $\mathbf{v}_{\text{k,eval}}$ in the evaluation dataset. M patterns $\mathbf{v}_{\text{k,eval}}$ are then combined with their timeseries $\mathbf{t}_{\text{k,eval}}$ to make a forced response, similar to the use of M SNMPs in Wills et al. (2020).

Training was based on applying this procedure with 4 training models at a time, treating the left out model as the evaluation data, and then comparing the resulting forced component estimate with the ensemble mean. Various choices of the EOF truncation (retained variance) and M were chosen, and 70% variance retained and $M = 1$ were chosen for the Tier 1 submission. Choosing the EOF truncation based on retained variance allows it to adapt to the different EOF properties of different fields, choosing less EOFs for zmTa or SST and more for PR or Rx1day. For all fields, only one SNMP is used, such that the number of tunable parameters is $nlon \cdot nlat + 2 = 1.037 \times 10^4$, where 2 is for the 2 hyperparameters.

1.19 AllFinger

Contacts: Céline Bonfils (bonfils2@llnl.gov), Jia-Rui Shi, Shiheng Duan, Gavin D. Madakumbura, Stephen Po-Chedley

This approach builds on Klaus Hasselmann's pioneering efforts to develop a statistical framework designed to distinguish signal from noise and identify a forced warming signal within a noisy climate background (Hasselmann, 1979). A key principle of Hasselmann's approach is that each external driver—whether anthropogenic or natural—as well as internal climate variability, manifests in distinct spatial and temporal patterns, leaving unique geographical imprints (or "fingerprints") on the Earth's climate. The extraction of the forced component out of noisy background serves the foundation of many modern fingerprint detection and attribution techniques. In the pattern-based fingerprint method (e.g, Santer et al. (2023)), the forced pattern is obtained by averaging across models and extracting the leading EOF, simultaneously amplifying the forced signal and reducing the noise. The target field (observations or unseen realizations) is projected onto the fingerprint to generate a pseudo-PC time-series, which measures the similarity between the fingerprint and the target's time-varying patterns. A novel aspect of our methodology is the use of a machine learning concept, which trains on large dataset while validating on a held-out portion of the training data. Similarly, our validation involves: (1) estimating the fingerprint from all-but-one ensembles, (2) predicting the forced component using a single realization from the held-out ensemble, and (3) comparing the predicted trend map to the “ground truth”, defined as the averaged trend map across the held-out ensemble realizations. Errors in our methodology can arise from uncertainties in forcings and responses, the presence of residual unforced variability within the forced fingerprint, intrinsic model errors, and observational uncertainties.

We propose two versions of this method: AllFinger, which relies on a standard fingerprinting analysis using monthly anomaly data (described here) and MonthFinger, an improved version that utilizes standardized data for each individual month (Section 1.19). As a first step, we estimate the model-based fingerprint for a target variable T by using an EOF analysis applied to the large ensembles of historical simulations. The EOF is calculated using the multi-model average of anomalies $T_{hist,av}(x,t)$, where x represents the number of grid cells, and t denotes time in years multiplied by 12 months. The anomalies in $T_{hist}(i,j,x,t)$ for the j-th model and the i-th realization are defined relative to local climatological monthly means over the period 1950 to 2022. The multi-model average $T_{hist,av}(x,t)$ is obtained by first averaging across realizations and then across models. The fingerprint is then extracted as the leading EOF of $T_{hist,av}(x,t)$. In the second step, we derived the forced component of each target field (observations or unseen individual realization) by projecting its data in anomaly form onto the fingerprint. The resulting pseudo-PC time-series measures the pattern similarity between the fingerprint and the target field at each time step. The 3D forced response field is reconstructed using the forced pattern and pseudo-PCs. Its trend map is then calculated and compared to the “ground truth”.

Here we focus exclusively on the leading EOF mode as the forced response. The trend map of the forced response is independent of the target dataset and is solely determined by the training ensembles. However, the temporal evolution and magnitude of changes are strongly influenced by the target field itself. The fingerprint derived from anomaly data from all months at once demonstrates good performance for the temperature and/or annual-mean fields. This approach has potential for further refinement by standardizing the anomalies (Section 1.20) or by adjusting the number of EOF modes utilized.

1.20 MonthFinger

Contacts: Shiheng Duan (duan5@llnl.gov), Jia-Rui Shi, Céline Bonfils, Gavin D. Madakumbura, Stephen Po-Chedley

This method is similar to AllFinger (Section 1.19), with the key difference being that the fingerprints (EOFs) are calculated separately for each individual month, using the multi-model average of standardized anomalies $T_{hist,av}(x,t)$, where t now represents time in years for that specific month. The monthly anomalies in $T_{hist}(i,j,x,t)$ for the j-th model and the i-th realization are defined relative to local monthly climatology over the period 1950 to 2022 and standardized using the local temporal standard deviation $\sigma(x)$ calculated for each month based on the concatenated unforced component across ensembles, determined by subtracting the ensemble mean from individual model realizations. Compared to AllFinger (Section 1.19) the use of 12 individual fingerprints (one for each month) in the MonthFinger demonstrates clear advantages when applied to the monthly precipitation fields.

1.21 3DUNet-Fingerprinters

Contacts: Gavin D. Madakumbura (gavindayanga@g.ucla.edu), Céline Bonfils, Jia-Rui Shi, Stephen Po-Chedley

The application of machine learning (ML)-based pattern identification for forced signal estimation was introduced by Barnes et al. (2019). In their foundational work, they modeled the year of the data, considered a proxy for anthropogenic forcing, using a shallow fully connected neural network and spatial maps of the variable of interest as inputs (e.g., surface temperature, precipitation rate). While this ML-based approach has been particularly useful in identifying a global estimate of the forced response (e.g., Barnes et al., (2020); Madakumbura et al., (2021)), for the objective of identifying the globally spatiotemporally complete forced response, ML-based pattern recognition can be utilized using more complex ML architectures (Goodfellow et al., 2016). Specifically, the task can be categorized as an image-to-image (or video-to-image, if temporal lags of variables are included) translation, and the U-Net architecture (Ronneberger et al., 2015) has been popular for such tasks in climate science (Lin et al., 2023; Quesada-Chacón et al., 2022; Serifi et al., 2021).

In 3DUNet-Fingerprinters, we utilized a U-Net. We use spatial maps of the original field (e.g., temperature) to model the spatial maps of the forced response, estimated as an ensemble mean of a large ensemble simulation. We tuned the hyperparameters by leaving one training model out and, therefore, we believe the generalizability of the trained model to unseen models or observations is adequate. Regarding the forced trends, the holdout model ensembles gave quite high pattern correlations and low errors in the magnitudes, compared to using a fully connected and 3D convolutional neural network. The U-Net consists of convolutional layers, downsampling, a bottleneck, and upsampling components, with skip connections. In the final submitted version, we used the current year and 9 years of lags (total 10 years) to model the forced component of the current year. While using 9-year temporal lags comes with the drawback of losing the forced component of the first 9 years, requiring a secondary data infilling step, it provided more accurate results compared to shorter or zero temporal lags. To model the monthly forced component, we employed one U-Net model for each month and combined results from 12 models. Training was done for the period 1950-2100. The model architecture and hyperparameters were tuned by cross-validation, leaving one large ensemble out during training and evaluating the model with the holdout model data.

1.22 EOF-SLR

Contacts: Andrei Gavrilov (gavrilov@uv.es), Sergey Kravtsov, Alexander Feigin

The philosophy of the EOF-SLR method is based on two main points: (i) the forced response of all models and of real climate system can be efficiently approximated as a mapping from a low-dimensional space of variables (components), and (ii) there exists an operator which approximates the low-dimensional forced components as a function of observable time series. This operator is assumed to be independent of the data source, so it should be able to decode the physics affecting the resulted response directly from the observed time series. Both steps are implemented using linear mappings between the spaces of observable and reduced variables, making it a fingerprinting method.

In particular, at the first step we estimate ensemble EOFs across a 125-member ensemble (25 members for each of 5 models), estimate the forced response of each model as a model-wise ensemble mean and project it onto 20 leading EOFs to get 20 forced response components for each of 5 training models. At the second step we perform a smoothed linear regression (SLR) procedure: We look for a fingerprinting regression pattern which allows reproduction of the forced response components from temporally-smoothed original data. Here we pre-compress the data by projecting it again onto K leading ensemble EOFs, and the smoothing timescale is different for each of 20 forced response components we try to approximate. Both K and the smoothing timescale are optimized using cross-validation. The SLR is performed across 125 available pairs of ensemble members and their forced response components. Since the forced components are equal for different realizations of internal variability within one model, the fingerprinting patterns are thus trained to be robust with respect to internal variability and model-related specifics, and can be applied to estimate the forced components of unseen data. These forced response components are then composed with the patterns found at the first step.

The potential errors may arise from poor statistics (e.g., statistical uncertainty of the ensemble mean, especially slow internal variability filtering) or principal absence of SLR-mapping with appropriate robustness level. Also, for the precipitation variables the method is applied to a logarithmically scaled data, which may effect performance evaluation based on unscaled precipitation, as is done in the main text.

The number of parameters in EOF-SLR is $K \cdot nlon \cdot nlat + (K+1) \cdot 20$, where K ranges from 20 to 200 depending on the target variable.

1.23 LDM-SLR

Contacts: Andrei Gavrilov (gavrilov@uv.es), Sergey Kravtsov, Alexander Feigin

LDM-SLR method is similar to EOF-SLR, but its first step is based on the linear dynamic mode (LDM) decomposition (Gavrilov et al., 2019, 2020, 2024). It finds low-dimensional components (linear dynamical modes) of forced and unforced variability in the ensemble climate data, involving Bayesian optimization of their timescales. Here we generalized the method for the case of a multi-model ensemble. The strength of this method lies in the possibility of efficient separation of forced and internal variability modes with large timescales in small ensembles (Gavrilov et al., 2024; Buyanova et al., 2025). In this application, however, we only use the forced response part of the decomposition, and the ensemble size is quite large.

Potential errors may still arise from poor statistics (e.g. for slow internal and forced modes with substantially overlapping spatial structures), principal absence of an SLR-mapping with appropriate robustness level, or suboptimal LDM dimension (usually due to computational cost reasons).

1.24 Anchor-OPLS

Contacts: Homer Durand (durand.homer@gmail.com), Gherardo Varando, Nathan Mankovich

The method is an extension of anchor regression for fingerprint extraction using orthonormalised partial least squares (OPLS) instead of ordinary least squares as learning algorithm. The methods aims at estimating the forced component using the variable of interest, with robustness to a potential shift in the magnitude of internal variability at regional scale. If the unseen model has larger amplitude internal variability at regional scales, the method is expected to perform better than a classical unregularised OPLS estimation. As the method assumes a linear model from the variable of interest to the forced response, it is not expected to perform well in models (or regions) where the relation is highly nonlinear. Thus the performance is not expected to be as high in predicting PR and SLP variables as for temperature variables.

Our method extends the anchor regression framework for fingerprint extraction, as originally introduced by Sippel et al. (2021). Given the multivariate nature of the target variable, OPLS regression is employed as the learning algorithm due to its advantageous properties in such scenarios. An essential aspect of this extension lies in the selection of anchor variables, which represent proxies of internal variability at the regional scale, ensuring robust estimation against potential increases in the magnitude of regional internal variability. While similar to the approach outlined in Sippel et al. (2021), this method utilizes OPLS instead of Ordinary Least Squares Regression to leverage potential correlation structures within the target variable and employs regional proxies of internal variability as anchor variables.

Regarding the training procedure, due to the large size of the training dataset, a sampling strategy is adopted. Specifically, 100 random ensemble members from the training model are sampled for training, and a train/validation split is performed to select ridge hyperparameters and determine the number of components for the OPLS algorithm. The ridge hyperparameter is chosen from the set [10, 100, 1000], while the number of components ranges linearly from 1 to 200 with 100 values. With a causal regularization parameter set to 5, reflecting an expectation of a moderate increase in regional internal variability magnitude, equal weighting is given to each model by randomly sampling 1500 samples for hyperparameter selection. To enhance computational efficiency, predictors are coarsened by a factor of 6, resulting in 288 predictors and $nlon \cdot nlat = 10,368$ target variables.

1.25 UNet3D-LOCEAN

Contacts: Guillaume Gastineau (guillaume.gastineau@locean.ipsl.fr), Constantin Bône

UNet3D-LOCEAN uses a refined version of the method by Bône et al. (2024). It is a U-Net with three-dimensional kernels that learns the spatiotemporal features of the forced and internal variability. Several modifications have been implemented compared to Bône et al. (2024). First, the U-Net architecture was modified for the use of monthly data in the 1950-2022 period, with the horizontal resolution of ForceSMIP data. The padding was improved. Several dropout layers were added to avoid overfitting. The training was changed to a more classic supervised learning, as the noise-to-noise procedure used in Bône et al. (2024) was found to result in slightly larger validation error for the variable T2m.

The U-Net is trained on monthly maps of anomalies from a single member, with 4-5 input channels—SST, T2m, zmTa, SLP, and the target variable (when it is not one of the previous four variables). The output consists of two channels: the monthly ensemble mean and the deviations from the ensemble mean, corresponding to the estimated forced response and internal variability. All data were normalized before the training and denormalized after the inference. The loss function optimized in the learning process is the area-weighted time-averaged squared error between the predicted and actual output. Although multiple fields are used as input, a separate training is performed for each variable, resulting in eight different U-Nets. Each U-Net is specialized to identify the forced and internal variability of its respective variable.

A cross-validation was conducted using the training data for the choice of hyperparameters (number of epochs, dropout rate, kernel size, and number of layers), but this was only applied to the U-Net specialized in identifying the internal and forced variability of T2m, and the same hyperparameters were used in the other U-Nets. We also investigate the influence of the period selected to train the models. The lowest errors obtained in the 1950-2022 periods are obtained for T2m when using the period 1940-2012, 1950-2022 and 1960-2032 in the training. For each target variable, several U-Net are trained leaving out the data from each model successively, and by using a random translation of the input data. The estimated response provided in ForceSMIP is the median of the 100-members obtained from each estimate. Further details regarding the method and the tests performed will be provided in a forthcoming publication.

1.26 TrainingEM

Contacts: Stephen Po-Chedley (pochedley1@llnl.gov), Robert Jnglin Wills

The multi-ensemble-mean of the five training models (at monthly resolution) was scaled such that the 1950-2022 trend in global-mean surface air temperature matches the data set of interest (e.g., observations). The same scaling constant is applied across all target variables. The number of tunable parameters is $nlon \cdot nlat \cdot ntime = 9.1 \times 10^6$.

1.27 RandomForest

Contact: Stephen Cropper (croppers@g.ucla.edu)

This method estimates forced climate variability by training a random forest (RF) regression model (Breiman, 2001; Geurts et al., 2006), on the ForceSMIP training ensemble dataset. An RF regression model is used to estimate a continuous target value based on a vector of features. Mathematically, it averages together the predictions of an ensemble of decision trees to produce a cumulative estimate that is designed to be more accurate and resistant to overfitting when compared to other methods. The RF regression model is optimized against the training data using the mean squared error as the criteria.

Specifically for ForceSMIP, we are training the RF model to estimate the forced component of climate variability (the target value) using model output from an arbitrary climate model in the ForceSMIP ensemble. The model output can be from any realization and is subselected to only include the variable of interest (only the PR climate variable is used to estimate the PR forced component, for example).

The training procedure for each GCM model is as follows. First, for a given GCM model, one realization is selected to compute the four largest modes of variability using Principal Component Analysis (PCA) for the climate variable of interest. This procedure reduces the size of the training data from an original dimension of samples (N) × months of lead time (60) × latitude (72) × longitude (144) into samples (N) × PCA components (4) whereby the magnitudes of the four largest PCA modes are kept for each sample. In this case, N is simply the number of months that the user is interested in estimating the forced component (e.g., for a single realization, this could be up to 86 years × 12 months - 60 months lag time = 972).

Second, the forced data is estimated using the mean across all realizations. The target data is of size samples × 1 where the target variable from each sample is selected from the forced data at some arbitrary latitude and longitude. Third, a distinct Random Forest Regression model is trained for each latitude and longitude spanning the 72 × 144 horizontal dimensions of the data and for each GCM model in the training data ($N = 72 \times 144 = 10368$).

For a final prediction of the target value, our method produces one forced component estimate for each GCM model in the training data and averages them together. This method can be then be applied for evaluation purposes on any out-of-sample models, including observation data, provided the data is gridded to the same horizontal resolution and is of the same monthly temporal resolution.

Some advantages of this method include: (1) the ability to generate infinite estimates of the forced component of variability by randomizing the initial state of the RF model; (2) the interpretability of the RF model; (3) the resistance of the RF model to overfitting.

Some disadvantages include: (1) the intense computational resources required to construct the decision tree ensembles; (2) high degrees of freedom; (3) the requirement that the feature data match the shape of the original data.

1.28 EncoderDecoder

Contacts: Jamin Rader (jaminrader.science@gmail.com), Charlotte Connolly, M.A. Fernandez, Emily Gordon

We use a feed-forward, high-dimensional encoder-decoder neural network architecture to predict the unforced component (internal variability). The neural network takes single-month maps as input (e.g., January for SLP). There is a different network trained for each month. In this method, the single-month map is input into the encoder which feeds through hidden layers until it reaches the latent space or “code.” From the code layer, the information is passed forward through the decoder layers which output predictions of internal variability at each grid point. We have two encoding and decoding layers, with 1000 nodes each, and use 100 nodes in the code layer. The number of nodes in the latent space was chosen to be small enough to prevent overfitting without loss of information in the decoding step. We use tanh as our activation function, a learning rate of 0.0001, a batch size of 64, and ran the network for 15 epochs.

The input maps are standardized as: 1) For each individual input map, subtract its own mean and divide by its own standard deviation; 2) using the maps produced in Step 1, we calculate the mean and deviation across each ensemble member. Subtract this mean, and divide by this standard deviation. The output maps (internal variability) are standardized as: 1) Take the mean and standard deviation of the internal variability from the entire training set (i.e. including all GCMs), subtract the mean, divide by the standard deviation. This method does not use time-evolving information (i.e., past states of the climate). All predictions for Tier 1 were done with a 1:1 prediction approach: We use a single map, for a single month and a single variable, to predict the forced response in that month and that variable.

In predicting the unforced component, we make the assumption that our training set, in this case five GCMs with 25 members each, accurately captures observed global teleconnections. Our methodology therefore assumes that climate model realizations of internal variability are faithful to observed internal variability regardless of climate model representations of the forced response. We expect this method will not work as well as predicting the forced response directly when applied to evaluation GCMs that are similar in their representations of the forced response to the training set, but we believe it may generalize to different climates (e.g., observations) better. Our method is likely more robust to realizations of the climate system that are outside of the training distribution than a comparable method that directly predicts the forced response. However, internally driven precipitation signals may be weak in GCMs, and GCMs disagree with observations and each other, and thus predicting the forced trend in precipitation with our methodology may not be appropriate.

We further developed this approach and applied the updated method in an attribution framework to the record-high 2023 SST in a recent work (Rader et al., 2025).

1.29 EnsFMP

Contact: Robert Jnglin Wills (r.jnglinwills@usys.ethz.ch), Maren Höver

The ensemble fingerprint maximizing pattern (EnsFMP) method estimates the forced response in an unknown dataset based on the agreement of its pattern with the forced response in one of the training ensembles, based on maximizing the signal-to-noise ratio of a fingerprint of the model forced response in the unknown dataset. This is done multiple times with different combinations of the training data used as the target forced response, and a weighted ensemble average of these estimates is used to generate the final forced response estimate, with the weights determined based on the pattern correlation of the linear trends in each forced response estimate with those of the raw data, to avoid cases where the fingerprint was not found in the unknown dataset.

We expect it to have a more complete estimate of the forced response (i.e., including subtle details like changes in seasonality and responses to volcanoes) than other methods, but it may also include some variability in the forced response estimate. Compared to the methodologically similar SNMP-OF method, this method relies more heavily on the training data (as evident in its larger number of tunable parameters).

EnsFMP method is based on signal-to-noise maximizing pattern (SNMP) analysis described in Wills et al. (2020) and references therein. In this case, the SNMP analysis framework is applied to maximize the projection onto a signal defined by the forced response in the training models (two at a time), from total variance that includes one member each from the same models as well as the evaluation member (e.g., observations). This allows the SNMP to be influenced by observed variance while looking for patterns that look like the forced response in the utilized model. The forced response estimate is then a combination of the leading pattern(s) with their timeseries, including a number of patterns N_{FMPs} that is specified or determined within the training data. The forced response estimates based on 5 different combinations of 2 training models are then averaged together into a single "ensemble" forced response estimate.

This method was found to have issues for cases where the training models and evaluation models have different climate sensitivities, which led to underestimation or overestimation of the global warming rate found in the raw data. To address this, we removed the 4th-order polynomial trend in the annual means at each grid point before the analysis, then added if back in afterwards. This has the disadvantage that the forced response estimate is closely tied to the 4th-order polynomial trend, but it allows the method to focus on precise details like the the spatiotemporal evolution (seasonality, volcanic eruptions, etc.).

1.30 ANN-Fingerprinters

Contacts: Shiheng Duan (duan5@llnl.gov), Céline Bonfils

The ANN model takes the anomaly map as the input and forced component as the target. It adopts an encoder-decoder architecture. At the bottleneck of the architecture, the year index is added, providing temporal context to the model. Subsequently, the decoder generates the forced component as its output. While the ANN model does not explicitly include temporal features, the addition of the year index is expected to effectively convey temporal information and enable it to capture the time evolving forced signal within the dataset.

References

Baldwin, M. P., Stephenson, D. B., & Jolliffe, I. T. (2009). Spatial weighting and iterative projection methods for EOFs. *J. Climate*, *22*(2), 234–243.

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019). Viewing forced climate patterns through an AI lens. *Geophys. Res. Lett.*, *46*(22), 13389–13398.

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002195.

Bône, C., Gastineau, G., Thiria, S., Gallinari, P., & Mejia, C. (2024). Separation of internal and forced variability of climate using a U-Net. *Journal of Advances in Modeling Earth Systems*, *16*(6), e2023MS003964.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Buyanova, M., Gavrilov, A., & Mukhin, D. (2025, 4). Analysis of forced response and internal climate variability in the inmcm earth system model. *Russian Journal of Numerical Analysis and Mathematical Modelling*, *40*, 91-106. doi: 10.1515/rnam-2025-0008

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

Demo, N., Tezzele, M., & Rozza, G. (2018). PyDMD: Python dynamic mode decomposition. *Journal of Open Source Software*, *3*(22), 530.

Deser, C., & Phillips, A. S. (2021). Defining the internal component of Atlantic multidecadal variability in a changing climate. *Geophysical Research Letters*, *48*(22), e2021GL095023.

Deser, C., & Phillips, A. S. (2023). Spurious Indo-Pacific connections to internal Atlantic Multidecadal variability introduced by the global temperature residual method. *Geophysical Research Letters*, *50*(3), e2022GL100574.

Duchon, C. E. (1979). Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology (1962-1982)*, 1016–1022.

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., ... Zhang, H. (2021). The earth's energy budget, climate feedbacks, and climate sensitivity [Book Section]. In V. Masson-Delmotte et al. (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (p. 923-1054). Cambridge, UK and New York, NY, USA: Cambridge University Press. Retrieved from `https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC`$_{A}R6_{W}GI_{C}hapter07.pdf$ doi: 10.1017/9781009157896.009

Frankignoul, C., Gastineau, G., & Kwon, Y.-O. (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., Kravtsov, S., Buyanova, M., Mukhin, D., Loskutov, E., & Feigin, A. (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., Kravtsov, S., & Mukhin, D. (2020). Analysis of 20th century surface air temperature using linear dynamical modes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(12).

Gavrilov, A., Seleznev, A., Mukhin, D., Loskutov, E., Feigin, A., & Kurths, J. (2019). Linear dynamical modes as new variables for data-driven ENSO forecast. *Climate Dynamics*, *52*, 2199–2216.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42.

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., ... Tebaldi, C. (2016). The detection and attribution model intercomparison project (DAMIP v1. 0) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3685–3697.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1) (No. 2). MIT press Cambridge.

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

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... others (2018). Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS). *Geoscientific Model Development*, *11*(1), 369–408.

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

Ichinaga, S. M., Andreuzzi, F., Demo, N., Tezzele, M., Lapo, K., Rozza, G., ... Kutz, J. N. (2024). PyDMD: A Python package for robust dynamic mode decomposition. *Journal of Machine Learning Research*, *25*(417), 1–9.

Kosaka, Y., & Xie, S.-P. (2013). Recent global-warming hiatus tied to equatorial Pacific surface cooling. *nature*, *501*(7467), 403–407.

Lien, J., Kuo, Y.-N., Ando, H., & Kido, S. (2025). Colored linear inverse model: A data-driven method for studying dynamical systems with temporally correlated stochasticity. *Physical Review Research*, *7*(2), 023042.

Lin, H., Tang, J., Wang, S., Wang, S., & Dong, G. (2023). Deep learning downscaled high-resolution daily near surface meteorological datasets over East Asia. *Scientific Data*, *10*(1), 890.

Madakumbura, G. D., Thackeray, C. W., Norris, J., Goldenson, N., & Hall, A. (2021). Anthropogenic influence on extreme precipitation over global land areas seen in multiple observational datasets. *Nature Communications*, *12*(1), 3944.

Mankovich, N., Bouabid, S., Nowack, P., Bassotto, D., & Camps-Valls, G. (2025). Analyzing climate scenarios using dynamic mode decomposition with control. *Environmental Data Science*, *4*, e16.

Penland, C., & Sardeshmukh, P. D. (1995). The optimal growth of tropical sea surface temperature anomalies. *J. Climate*, *8*(8), 1999–2024.

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

Quesada-Chacón, D., Barfus, K., & Bernhofer, C. (2022). Repeatable high-resolution statistical downscaling through deep learning. *Geoscientific Model Development Discussions*, *2022*, 1–28.

Rader, J. K., Connolly, C., Fernandez, M. A., & Gordon, E. M. (2025). Attribution of the record-high 2023 SST using a deep-learning framework. *Environmental Research Communications*. Retrieved from `http://iopscience.iop.org/article/10.1088/2515-7620/add322`

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–miccai 2015: 18th international conference, munich, germany, october 5-9, 2015, proceedings, part iii 18* (pp. 234–241).

Santer, B. D., Po-Chedley, S., Zhao, L., Zou, C.-Z., Fu, Q., Solomon, S., ... Taylor, K. E. (2023). Exceptional stratospheric contribution to human fingerprints on atmospheric temperature. *Proceedings of the National Academy of Sciences*, *120*(20), e2300758120.

Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, *656*, 5–28.

Serifi, A., Günther, T., & Ban, N. (2021). Spatio-temporal downscaling of climate data using convolutional and error-predicting neural networks. *Frontiers in Climate*, *3*, 656479.

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

Varando, G., Fernández-Torres, M.-Á., Muñoz-Marí, J., & Camps-Valls, G. (2022). Learning causal representations with Granger PCA. In *UAI 2022 workshop on causal representation learning.*

Wilcox, L. J., Allen, R. J., Samset, B. H., Bollasina, M. A., Griffiths, P. T., Keeble, J., ... others (2023). The regional aerosol model intercomparison project (RAMIP). *Geoscientific Model Development*, *16*(15), 4451–4479.

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

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

Yan, X., Zhang, R., & Knutson, T. R. (2019). A multivariate AMV index and associated discrepancies between observed and cmip5 externally forced amv. *Geophysical Research Letters*, *46*(8), 4421–4431.

Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y.-O., Marsh, R., Yeager, S. G., ... Little, C. M. (2019). A review of the role of the Atlantic Meridional Overturning Circulation in Atlantic multidecadal variability and associated climate impacts. *Reviews of Geophysics*.