Optimal Fingerprinting with Estimating Equations

SAI MA,[a] TIANYING WANG,[b] JUN YAN,[a] AND XUEBIN ZHANG[c]

[a] *Department of Statistics, University of Connecticut, Storrs, Connecticut*
[b] *Department of Statistics, Colorado State University, Fort Collins, Colorado*
[c] *Climate Research Division, Environment and Climate Change Canada, Toronto, Ontario, Canada*

(Manuscript received 7 September 2022, in final form 12 June 2023, accepted 11 July 2023)

ABSTRACT: Climate change detection and attribution have played a central role in establishing the influence of human activities on climate. Optimal fingerprinting, a linear regression with errors in variables (EIVs), has been widely used in detection and attribution analyses of climate change. The method regresses observed climate variables on the expected climate responses to the external forcings, which are measured with EIVs. The reliability of the method depends critically on proper point and interval estimations of the regression coefficients. The confidence intervals constructed from the prevailing method, total least squares (TLS), have been reported to be too narrow to match their nominal confidence levels. We propose a novel framework to estimate the regression coefficients based on an efficient, bias-corrected estimating equations approach. The confidence intervals are constructed with a pseudo residual bootstrap variance estimator that takes advantage of the available control runs. Our regression coefficient estimator is unbiased, with a smaller variance than the TLS estimator. Our estimation of the sampling variability of the estimator has a low bias compared to that from TLS, which is substantially negatively biased. The resulting confidence intervals for the regression coefficients have coverage rates close to the nominal level, which ensures valid inferences in detection and attribution analyses. In applications to the annual mean near-surface air temperature at the global, continental, and subcontinental scales during 1951–2020, the proposed method led to shorter confidence intervals than those based on TLS in most of the analyses.

SIGNIFICANCE STATEMENT: Optimal fingerprinting is an important statistical tool for estimating human influences on the climate and for quantifying the associated uncertainty. Nonetheless, the estimators from the prevailing practice are not as optimal as believed, and their uncertainties are underestimated, both owing to the unreliable estimation of the optimal weight matrix that is critical to the method. Here we propose an estimation method based on the theory of estimating equations; to assess the uncertainty of the resulting estimator, we propose a pseudo bootstrap procedure. Through extensive numerical studies commonly used in statistical investigations, we demonstrate that the new estimator has a smaller mean-square error, and its uncertainty is estimated much closer to the true uncertainty than the prevailing total least squares method.

KEYWORDS: Error analysis; Numerical analysis/modeling; Statistics; Climate models; Climate change

1. Introduction

The successive assessments of the Intergovernmental Panel on Climate Change (IPCC) have established that human influence has resulted in global warming, mainly through the emission of greenhouse gases (Hegerl et al. 2007; Bindoff et al. 2013; Eyring et al. 2021). Climate change detection and attribution provided the critical evidence leading to the IPCC conclusions. Optimal fingerprinting (OF), a multiple linear regression model, is the most widely used method for the detection and attribution of climate change. It regresses the observed climate variable of interest on the fingerprints, the expected responses of the climate system to external forcings (Hegerl et al. 1996; Allen and Tett 1999; Allen and Stott 2003). The regression coefficients are called scaling factors. Their point estimates scale the fingerprints to best match the observed climate change. If the confidence interval (interval estimate in statistics) of a scaling factor is significantly above 0, then the effect of the corresponding external forcing is said to be "detected" in the observed data. If, in addition, the confidence interval covers 1, then this is necessary (not sufficient) evidence that the observed changes can be "attributed" to that external forcing.

Both point and interval estimations of the scaling factors are the center of the statistical inference in detection and attribution analyses. The point estimate of a scaling factor reflects how well the model-simulated response has properly estimated the magnitude of the observed changes. A good point estimator should be unbiased, with a variance as small as possible. The interval estimate is important because it provides a quantification of the estimation. Another aspect that has not been widely focused on in the climate literature is the so-called coverage rate of a confidence interval. Often, the

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-22-0681.s1.

Corresponding author: Tianying Wang, tianying.wang@colostate.edu

DOI: 10.1175/JCLI-D-22-0681.1

statistical meaning of a confidence interval is misinterpreted. In (frequentist) statistics, a 90% confidence interval for a target does not mean the target will be within the confidence interval with a 90% probability. Rather, it means that if the estimation is repeated many times, there is a 90% chance that the target is covered by the confidence interval. A proper confidence interval should have a coverage rate, the percentage of times that it covers the target in repeated estimations, the same as its nominal level.

The linear regression in the OF setting has two distinguishing challenges compared to the standard setting. First, the predictors or the fingerprints of the external forcings are not observed but estimated from climate model simulations, typically as multimodel ensemble averages. Because individual simulations contain natural variations of the climate, averaging will not remove uncertainty completely. That is, the average contains noises, or errors in the predictors, leading to the so-called errors-in-variables (EIV) issue, also known as measurement errors in statistics (Carroll et al. 2006). Different climate models may produce different climate responses, and the model structural differences can be treated (Huntingford et al. 2006), but this is not considered here. If ignored, EIVs may yield a severely biased estimator of the scaling factors. Under the assumption that the natural climate variability in individual model simulations is the same as that in the observations, the errors in the estimated fingerprints have the same covariance structure as the internal climate variability $\mathbf{\Sigma}$, but the magnitudes are different depending on the number of simulations being used in the averaging. Allen and Stott (2003) for the first time addressed the EIV issue for OF with total least squares (TLS), which remains commonly used.

The second challenge is that the response variable of the OF (e.g., the observed climate variable) is spatially and temporally dependent. As a result, the covariance matrix $\mathbf{\Sigma}$ of the regression error vector is needed to prewhiten the data. Nonetheless, $\mathbf{\Sigma}$ is not known and cannot be estimated from the data, as there is only one observation per site and time point. In practice, $\mathbf{\Sigma}$ is estimated using climate model simulations under the assumption, again, that model-simulated natural variability properly represents real-world natural climate variability. As $\mathbf{\Sigma}$ is of high dimension, available model simulations may not be sufficient to provide a reliable estimation. Methods have been proposed to improve the estimation, including the use of a regularized estimator of $\mathbf{\Sigma}$, which ensures its positive definiteness (Ribes et al. 2013). Confidence intervals of the scaling factors can be constructed based on the normal approximation of the estimator (Ribes et al. 2013; DelSole et al. 2019; Li et al. 2021) or bootstrap (Pešta 2013; DelSole et al. 2019).

The impact of using an estimated $\mathbf{\Sigma}$, especially when it is based on relatively small samples, has not been studied until recently. The resulting scaling factor estimator is no longer optimal in root-mean-square error (RMSE) (Li et al. 2023). The inverse of $\mathbf{\Sigma}$ acts as a weight, and it is possible for other weights to yield a better estimator of the scaling factors in terms of RMSE when $\mathbf{\Sigma}$ is estimated with a high level of uncertainty. Further, the resulting confidence intervals do not account for the uncertainty in the estimated $\mathbf{\Sigma}$ to provide enough coverage for the scaling factors. To reduce the effect of using an estimated $\mathbf{\Sigma}$ on the uncertainty of the interval estimate, a common practice is to produce two separate estimates for $\mathbf{\Sigma}$, one for prewhitening and the other for inferences (Hegerl et al. 1996; Allen and Stott 2003). This approach has been reported to not give confidence intervals with sufficient coverage (Li et al. 2021). Hannart (2016) proposed an integrated OF method, where the unobserved measurement errors in predictors and the unknown $\mathbf{\Sigma}$, under an inverse Wishart prior, are both integrated out of the likelihood with a closed form. However, the prior used in the study was too informative to be practical, and it is not clear how to properly specify the parameters of this prior distribution. Li et al. (2021) proposed to fix the undercoverage rate issue by using a parametric bootstrap calibration method that enlarges the confidence intervals such that their coverage rates match their nominal levels. The method may not work well, however, when the sample size of data for estimating $\mathbf{\Sigma}$ is more limited.

To tackle both the point estimation and the interval estimation tasks in the OF regression, we propose a novel framework based on estimating equations (EEs). The EE method is a widely used estimation technique in statistics (e.g., Godambe 1991; Heyde 2008). The method of least squares and the method of maximum likelihood frequently used in the climate literature are special cases of the EE method. For interested readers, we provide a brief tutorial of EEs in appendix A. For linear regression with EIVs, the bias-corrected EE method is a standard approach for regression coefficient estimation (e.g., Carroll et al. 2006). The extra challenge in OF is to appropriately account for the spatiotemporal dependence in both point and interval estimations. As will be shown in numerical studies, our point estimators are unbiased with smaller RMSEs, and our confidence intervals provide coverage rates close to the nominal level.

The rest of the paper is organized as follows. In section 2, we propose an EE method for estimating the scaling factors in OF and a pseudo bootstrap method to estimate the variance of the estimator, which leads to confidence intervals with desired coverage rates. In section 3, we report a simulation study that shows the competitiveness of our method in comparison with existing approaches in both point and interval estimations. The methods are applied in section 4 to the detection and attribution analyses of the annual mean temperature of 1951–2020 at the continental and subcontinental scales. A discussion concludes in section 5. To improve readability, we relegate technical details, including the basics of EEs, bias correction, constructing $\mathbf{\Sigma}$ with a block Toeplitz structure, and data description, to the appendixes.

2. Methods

In the following, we will start with the basic concepts, including the main ingredients and assumptions involved in OF. Then, we present our estimation for the scaling factors and construct their confidence intervals. We will also describe the diagnostics of our statistical model.

a. Statistical model

Three ingredients are used in typical detection and attribution analyses.

1) Observational data. These are a dataset of climate variability that potentially contains climate change signals to be detected. This can be, for example, a spatial map of long-term temperature trends. It can also be a time series of the annual mean temperature over the globe or over a region. But most often, it consists of the spatial and temporal evolutions of a climate variable that may enable detecting the effects of different external forcings separately.
2) Signal(s) or expected climate response(s) to one or more external forcings. Since they are not known, they are often estimated based on the responses simulated by climate models under different external forcings, such as anthropogenic (ANT) forcing, external natural (NAT) forcing, or combined ANT and NAT (ALL) forcing.
3) Information about natural internal climate variability or noise. This is typically based on climate model simulations under the "control" condition without external forcing. The residual of individual ensemble members after the removal of model-simulated responses can also be used to supplement the control simulation (in this paper, we will not make the distinction and refer to them as control simulations).

The three parts will be further explained as we introduce the notations next.

We consider a general setting with T time periods and S spatial regions (referred to sites below). For ease of notation, we assume that there are no missing data, but the method still works if there are missing values, as we will show in section 2b. Let $\mathbf{Y}_{ts}$ be the observation of the climate variable, covering time $t = 1, \ldots, T$ and space $s = 1, \ldots, S$. Indexing the time and space separately facilitates different treatments of the temporal and spatial dependences in section 2b. Suppose that one is interested in separately detecting signals from J external forcings. Let $\mathbf{X}_{tsj}$ be the true (unobserved) fingerprint of the jth external forcing, $j = 1, \ldots, J$. Let m_j be the number of simulations in the ensemble whose average $\tilde{\mathbf{X}}_{tsj}$ is used to estimate $\mathbf{X}_{tsj}, j = 1, \ldots, J$. With each time period treated as a map or a cluster, define for cluster t: $\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{tS})^{\mathrm{T}}$, $\mathbf{X}_{tj} = (X_{t1j}, \ldots, X_{tSj})^{\mathrm{T}}$, and $\mathbf{X}_t = (X_{t1}, \ldots, X_{tJ})$.

The OF framework assumes that the responses of the climate system to different forcings are additive. It links the observational climate variables to the signals by linear regression model (e.g., Allen and Stott 2003)

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{1}$$

and

$$\tilde{\mathbf{X}}_{tj} = \mathbf{X}_{tj} + \boldsymbol{\nu}_{tj}, \quad j = 1, \ldots, J, \qquad t = 1, \ldots, T, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^{\mathrm{T}}$ is a J-dimensional unknown scaling factor, $\boldsymbol{\epsilon}_t$ is an S-dimensional regression error, and $\boldsymbol{\nu}_{tj}$ are S-dimensional noises (measurement errors in statistics) in estimating true signal $\mathbf{X}_{tj}$ with $\tilde{\mathbf{X}}_{tj}$. The regression error $\boldsymbol{\epsilon}_t$ and measurement errors $\boldsymbol{\nu}_{tj}$ have mean zero and covariance matrices $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Omega}_{tj}$, respectively. Let $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\epsilon}_T^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\nu}_j = (\boldsymbol{\nu}_{1j}^{\mathrm{T}}, \ldots, \boldsymbol{\nu}_{Tj}^{\mathrm{T}})^{\mathrm{T}}, j = 1, \ldots, J$. Let $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}_j$ be the $TS \times TS$ covariance matrices of $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}_j$, respectively. With each block representing one time point, the diagonal blocks of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}_j$ are, respectively, $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Omega}_{tj}$.

1) Assumption 1

The noises in the estimated signals $\{\boldsymbol{\nu}_j: j = 1, \ldots, J\}$ are mutually independent and are independent of $\boldsymbol{\epsilon}$.

Assumption 1 implies that the measurement errors $\boldsymbol{\nu}_j$ are independent across ensemble members and between observations and model simulations. Note that this assumption does not assume independence across different time points within each ensemble member. This assumption is valid because simulations for individual signals are conducted separately, and they are also independent of the evolution of the natural climate.

The internal natural variability of the climate system is represented by $\boldsymbol{\Sigma}$, which is unknown but critical in making inferences about β. A common strategy is to estimate $\boldsymbol{\Sigma}$ from control runs of climate model simulations, which are assumed to reflect the pattern of internal climate variability. Let $\{\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(L)}\}$, each with dimension ST, be independent control simulations with sample size L. Let $\boldsymbol{\Psi}$ be the variance matrix of $\boldsymbol{\epsilon}^{(\ell)}, \ell = 1, \ldots, L$.

2) Assumption 2

The internal natural variability simulated by the climate models has the same temporal and spatial structures as those of the observations, though the magnitudes may differ. That is, for a scale parameter $a > 0$, $\boldsymbol{\Omega}_j = a\boldsymbol{\Sigma}/m_j, j = 1, \ldots, J$, and $\boldsymbol{\Psi} = a\boldsymbol{\Sigma}$.

Assumption 2 implies that internal climate variability is not affected by external forcing. Although there are examples, such as Arctic sea ice extent, where internal variability may change regionally under certain external forcing (Bonan et al. 2021; Swart et al. 2015), this assumption is reasonable for many OF settings over the period for which we have historical data. Typical OF applications are usually implemented assuming $a = 1$, which is checked through the residual consistency test (Allen and Tett 1999). Assumption 2 makes our method less restricted and adaptable to more general cases. It is possible to relax this assumption even more by assuming different models simulate different magnitudes of variability. As this is not the main focus of the paper, we will keep our statistical model simple by assuming a unified a and provide a test for statistical model diagnostics, as detailed later.

Estimating $\boldsymbol{\Sigma}$ with control simulations is challenging, as L could be much smaller than ST, which motivated the regularized OF (Ribes et al. 2013). Imposing some structures on $\boldsymbol{\Sigma}$ helps to improve the estimation if the structures are reasonable.

3) Assumption 3

The natural internal climate variability does not change over time, that is, it is temporally stationary.

While a typical OF analysis does not explicitly make this assumption, the estimation of $\boldsymbol{\Sigma}$ from control simulations in practice does not consider temporal changes. Additionally,

there is no evidence indicating covariance change on the time and space scales in a typical OF analysis. This explicit assumption greatly reduces the number of parameters in $\boldsymbol{\Sigma}$. It implies that $\boldsymbol{\Sigma}$ has a block Toeplitz structure with the diagonal blocks $\boldsymbol{\Sigma}_t$ to be the same for all t and that the covariance of $\sum_{t,t'}$ of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\epsilon}_{t'}$ depends only on the time lag $|t - t'|$. This assumption does not, however, state that the off-diagonal blocks are zero. In general, we expect that $\sum_{t,t'}$ decreases as $|t - t'|$ increases. As will be clear next, we discard the off-diagonal blocks of $\boldsymbol{\Sigma}$ in exchange for a much more reliable weight construction in our point estimation; the uncertainty caused by the unspecified off-diagonal blocks in the estimation is accounted for by a pseudo bootstrap procedure.

b. *Estimating the scaling factor: Point estimator*

If $\mathbf{X}_t$ were known, $\boldsymbol{\beta}$ can be easily estimated using the EE method; see appendix A for a tutorial. The basic principle of the EE method is to construct a set of equations based on the sample data and the unknown parameters, which are called EEs. The estimator is the solution EEs. Consider a simpler situation where $\boldsymbol{\Sigma}_t$ are known, and let us discard the temporal dependence for now. An EE can be obtained by weighted least squares.

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}) = 0. \tag{3}$$

Solving this equation for $\boldsymbol{\beta}$ gives a closed-form estimator of $\boldsymbol{\beta}$. As the expectation of the left side of Eq. (3) is zero, it is an unbiased EE, and the resulting estimator is a consistent estimator according to the EE theories (e.g., Godambe 1991; Heyde 2008). That is, as sample size T increases, the estimator converges in probability to the true parameter value of $\boldsymbol{\beta}$.

Now, only $\tilde{\mathbf{X}}_t$ instead of $\mathbf{X}_t$ are known. The left side of Eq. (3), with $\mathbf{X}_t$ substituted with $\tilde{\mathbf{X}}_t$, does not have an expectation of zero anymore, resulting in a biased estimating equation. With assumptions 1–3, it can be shown that (appendix B)

$$E\{\tilde{\mathbf{X}}_t^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}(\mathbf{Y}_t - \tilde{\mathbf{X}}_t\boldsymbol{\beta})\} = -aS\,\mathrm{diag}\left(\frac{1}{m_1}, \ldots, \frac{1}{m_J}\right)\boldsymbol{\beta}. \tag{4}$$

Thus, a bias-corrected estimating equation for $\boldsymbol{\beta}$ is

$$\frac{1}{T}\sum_{t=1}^{T}G_t(\boldsymbol{\beta}; \boldsymbol{\Sigma}_t) = 0, \tag{5}$$

where

$$G_t(\boldsymbol{\beta}; \boldsymbol{\Sigma}_t) = \tilde{\mathbf{X}}_t^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}(\mathbf{Y}_t - \tilde{\mathbf{X}}_t\boldsymbol{\beta}) + aS\,\mathrm{diag}\left(\frac{1}{m_1}, \ldots, \frac{1}{m_J}\right)\boldsymbol{\beta}.$$

Equation (5) is not implementable, because $\boldsymbol{\Sigma}_t$ is unknown. Under assumption 2, $\boldsymbol{\Psi}_t = a\boldsymbol{\Sigma}_t$, where $\boldsymbol{\Psi}_t$ are the diagonal blocks of $\boldsymbol{\Psi}$. So we can estimate the pattern of $\boldsymbol{\Sigma}_t$ using the control simulations $\{\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(L)}\}$. Under assumption 3, $\boldsymbol{\Sigma}_t$ are identical for all $t \in \{1, \ldots, T\}$. Therefore, the L replicates at all T time periods can be pooled to form a sample of size LT to estimate this $S \times S$ covariance matrix. This is in contrast to typical OF analyses with TLS, where a sample of size L is used to estimate $\boldsymbol{\Sigma}$ of dimension $TS \times TS$. Let $\hat{\boldsymbol{\Sigma}}_+$ and $\hat{\boldsymbol{\Psi}}_+$ be the pooled estimates of $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Psi}_t$, respectively. Substituting $\hat{\boldsymbol{\Sigma}}_+ = a^{-1}\hat{\boldsymbol{\Psi}}_+$ into $\boldsymbol{\Sigma}_t$ in Eq. (5), an implementable EE for $\boldsymbol{\beta}$ is

$$\frac{1}{T}\sum_{t=1}^{T}G_t(\boldsymbol{\beta}; \hat{\boldsymbol{\Sigma}}_+) = \frac{1}{T}\sum_{t=1}^{T}G_t(\boldsymbol{\beta}; a^{-1}\hat{\boldsymbol{\Psi}}_+) = 0. \tag{6}$$

Solving the EE [Eq. (6)] gives a closed-form estimator

$$\hat{\boldsymbol{\beta}}_T = \frac{1}{T}A_T\sum_{t=1}^{T}\tilde{\mathbf{X}}_t\hat{\boldsymbol{\Psi}}_+^{-1}\mathbf{Y}_t, \tag{7}$$

where

$$A_T = \left\{\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{X}}_t^{\mathrm{T}}\hat{\boldsymbol{\Psi}}_+^{-1}\tilde{\mathbf{X}}_t - S\,\mathrm{diag}\left(\frac{1}{m_1}, \ldots, \frac{1}{m_J}\right)\right\}^{-1}. \tag{8}$$

Interestingly, the unknown scale a cancels out, and it is not needed for the estimator [Eq. (7)].

Recall the EE [Eq. (6)] discards the temporal dependence, as it only uses the diagonal blocks of $\boldsymbol{\Sigma}$, which means we are losing efficiency. Estimating all T blocks of the Toeplitz structure of $\boldsymbol{\Sigma}$ is possible, but blocks further away from the main diagonals are estimated less reliably, because the number of available pairs drops. The fully estimated $\boldsymbol{\Sigma}$ may still not be of full rank, a challenge similarly encountered in typical OF analyses. Discarding the off-diagonal blocks in estimation has the potential to lose efficiency, but the diagonal blocks are estimated more reliably, which leads to a more reliable weight. The gain from a more reliable weight may well exceed the loss from discarding the temporal dependence, especially when the temporal dependence is not strong, as shown in our numerical studies. Indeed, in typical OF applications where interannual variability, such as the effect of ENSO, is averaged out, the temporal dependence is weak, and the efficiency loss can be minimal.

The estimator $\hat{\boldsymbol{\beta}}_T$ in Eq. (7) has nice properties. For a large T, it is approximately unbiased (consistent) as long as the expectation in Eq. (4) holds; no distributional assumptions have been made beyond the expectation. The efficiency loss from discarding the temporal dependence in point estimation is offset by the potentially big gain from a more reliable pooled estimator of the $S \times S$ weight matrix $\hat{\boldsymbol{\Sigma}}_+^{-1}$ or $\hat{\boldsymbol{\Omega}}_+^{-1}$. Incorporating the temporal dependence like the prevailing TLS method does through a much bigger $TS \times TS$ weight matrix has the potential to achieve higher efficiency in point estimation, but this potential cannot be realized because of the large uncertainty in estimating a much higher-dimensional weight matrix. Further, in practice, missing data do not affect the proposed estimator. We simply use the available observations to construct the contribution of each cluster (time point) to the EEs, with the rows and columns corresponding to the missing observations removed from $\hat{\boldsymbol{\Sigma}}_+$.

Although a is not needed in the point estimation of $\boldsymbol{\beta}$, it is needed in constructing the confidence intervals of $\boldsymbol{\beta}$, and an estimator of a provides a diagnosis of the restriction $a = 1$ in typical OF analyses. We give details of how to obtain an estimator $\hat{a}_T$ of a with an EE in appendix C.

c. Confidence intervals

Confidence intervals for $\boldsymbol{\beta}$ can be constructed based on the normal approximation of the estimator $\hat{\boldsymbol{\beta}}_T$. By the theory of EEs, as $T \to \infty$,

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}) \to N(0, \mathbf{ABA}^{\mathrm{T}}),$$

where $\mathbf{A} = \lim_{T\to\infty}\mathbf{A}_T$, $\mathbf{B} = \lim_{T\to\infty}\mathbf{B}_T$, and $\mathbf{B}_T = \mathrm{cov}\{T^{-1/2}\sum_{t=1}^{T} G_t(\boldsymbol{\beta}; \hat{\boldsymbol{\Sigma}}_+)\}$. Detailed derivation can be found in appendix D. The two components $\mathbf{A}$ and $\mathbf{B}$ in the variance can be consistently estimated by their sample counterparts, with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}_T$. In particular, $\mathbf{A}$ can be estimated by $\mathbf{A}_T$ easily, but estimating $\mathbf{B}$ is more challenging, because there may be unspecified temporal dependence, which, if ignored, would lead to confidence intervals with undercoverage issues.

We propose a pseudo residual bootstrap approach to fully utilize the control runs in estimating $\mathbf{B}$. Define the block residual $\mathbf{r}_t(\boldsymbol{\beta}) = \mathbf{Y}_t - \tilde{\mathbf{X}}_t\boldsymbol{\beta}$, $t = 1, \ldots, T$. Let $\mathbf{r}(\boldsymbol{\beta}) = [\mathbf{r}_1^{\mathrm{T}}(\boldsymbol{\beta}), \ldots, \mathbf{r}_T^{\mathrm{T}}(\boldsymbol{\beta})]^{\mathrm{T}}$. Note that $\mathbf{r}(\boldsymbol{\beta})$ has an expectation of zero and covariance matrix $\delta^2(\boldsymbol{\beta})\boldsymbol{\Sigma}$ with $\delta(\boldsymbol{\beta}) = (1 + a\sum_{j=1}^{J}\boldsymbol{\beta}_j^2/m_j)^{1/2}$. In a standard residual block bootstrap procedure, one would resample the residuals $\mathbf{r}(\hat{\boldsymbol{\beta}}_T)$ in blocks to preserve the spatial and temporal dependences and add the bootstrap copies of the residuals to $\tilde{\mathbf{X}}_t\hat{\boldsymbol{\beta}}_T$ to form bootstrap copies of $\mathbf{Y}_t$, $t = 1, \ldots, T$. The performance of the standard procedure here, however, is questionable, because it requires a large sample size T, whereas in a real-world application where multiyear averages are considered, T can be quite small. A valid bootstrap procedure has to preserve the spatial and temporal dependences in the data. We resort to control runs to meet this requirement, because they are assumed to have a similar covariance structure as the climate system (assumption 2).

The control runs need to be appropriately scaled to form bootstrap residuals with spatial and temporal dependences preserved. Recall that $\hat{\boldsymbol{\Psi}} = \hat{a}_T\hat{\boldsymbol{\Sigma}}$. With control runs $\{\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(L)}\}$, we use

$$\left\{\frac{\delta(\hat{\boldsymbol{\beta}}_T)\boldsymbol{\epsilon}^{(1)}}{\sqrt{\hat{a}_T}}, \ldots, \frac{\delta(\hat{\boldsymbol{\beta}}_T)\boldsymbol{\epsilon}^{(L)}}{\sqrt{\hat{a}_T}}\right\}$$

in place of bootstrapped residuals, which by construction preserve the spatial and temporal dependence perfectly. This leads to L pseudo bootstrap samples. Each pseudo bootstrap sample gives a copy of $T^{-1/2}\sum_{t=1}^{T} G_t(\boldsymbol{\beta}; \hat{\boldsymbol{\Sigma}}_+)$. We estimate $\mathbf{B}$ by the sample covariance matrix $\hat{\mathbf{B}}_T$ of these L copies. Note that the scale of $\hat{a}_T$ plays an important role in determining the residual magnitudes and, hence, the width of the confidence intervals. This procedure is computationally efficient, as it only requires evaluating the EE [Eq. (6)] with the L copies from control runs, neither bootstrapping nor resolving it. The procedure is itself general and could be applied to estimate the variance of other estimators, including the TLS estimator.

Note that one can also incorporate the runs for the J external forcings, after centering, through the aforementioned scaling procedure to increase the number of bootstrapped copies. In our simulations and data analysis, we used $L + m_1 + m_2 - 2$ runs in total, where m_1 and m_2 are the numbers of runs for the ANT and NAT forcings, respectively. The subtraction of 2 is because both the ANT and NAT forcings were centered in the analysis.

With $\mathbf{A}_T\hat{\mathbf{B}}_T\mathbf{A}_T^{\mathrm{T}}$ as an estimator for $\mathbf{ABA}^{\mathrm{T}}$, we are ready to construct confidence intervals for $\boldsymbol{\beta}$ based on the normal approximation of $\hat{\boldsymbol{\beta}}$. For each β_j, a $100(1-\alpha)\%$ confidence interval is

$$\left(\hat{\beta}_{T,j} - \frac{1}{\sqrt{T}}z_{\alpha/2}\hat{\sigma}_{\hat{\beta}_{Tj}}, \hat{\beta}_{T,j} + \frac{1}{\sqrt{T}}z_{\alpha/2}\hat{\sigma}_{\hat{\beta}_{Tj}}\right), \quad j = 1, \ldots, J,$$

where $\hat{\beta}_{T,j}$ is the jth component of $\hat{\boldsymbol{\beta}}_T$, $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution, and $\hat{\sigma}_{\hat{\beta}_{T,j}}$ is the jth diagonal element of $\mathbf{A}_T\hat{\mathbf{B}}_T\mathbf{A}_T^{\mathrm{T}}$. As shown in the simulation study in the next section, the coverage rates of the confidence intervals constructed this way are close to the nominal level.

To assess the quality of a confidence interval, we use the interval score of Gneiting and Raftery (2007) as a measure that combines both the interval length and its deviation from the target. Consider a symmetric $100(1-\alpha)\%$ confidence interval (L, U) for a target τ. The interval score with level α is

$$\mathrm{IS}_\alpha = (U - L) + \frac{2}{\alpha}(L - \tau)1(L > \tau) + \frac{2}{\alpha}(\tau - U)1(U < \tau),$$

where $1(\cdot)$ is an indicator function. The optimal score is achieved when the target τ is covered by (L, U), with the interval length being minimal. This score is used to compare different confidence intervals in the simulation study.

d. Diagnostics of the statistical model

The first diagnosis is to test the null hypothesis H_{01}: $a = 1$, which is assumed in typical OF analyses. The asymptotic normal distribution of $\hat{a}_T$ can be used to construct a Z statistic $Z = (\hat{a}_T - 1)/\hat{\mathrm{var}}^{1/2}(\hat{a}_T)$, which follows a standard normal distribution under the null asymptotically. The variance of $\hat{a}_T$ can be estimated using the same pseudo parametric bootstrap procedure for $\hat{\boldsymbol{\beta}}_T$.

Alternatively, H_{01} can be tested without an estimation of a. Consider prewhitened residuals $\mathbf{r}_t^* = a^{-1/2}\delta^{-1}(\hat{\boldsymbol{\beta}}_T)\hat{\boldsymbol{\Psi}}_+^{-1/2}\mathbf{r}_t$, $t = 1, \ldots, T$, fixing $a = 1$. Define $\mathbf{r}^* = (r_1^*, \ldots, r_T^*)^{\mathrm{T}}$. Under H_{01}, $\mathbf{r}^*$ should have a variance of 1, although there may be some temporal dependence. So we consider H_{02} that the prewhitened residual $\mathbf{r}^*$ has a variance of 1 and use the sample variance $S^2(\mathbf{r}^*)$ of $\mathbf{r}^*$ as the testing statistic for H_{02}. The null distribution of $S^2(\mathbf{r}^*)$ depends on the unspecified temporal dependence. Denote prewhitened control runs $\boldsymbol{\epsilon}_t^{(l)*} = \hat{\boldsymbol{\Psi}}_+^{-1/2}\boldsymbol{\epsilon}_t^{(l)}$ and $\boldsymbol{\epsilon}^{(l)*} = [\epsilon_1^{(l)}, \ldots, \epsilon_T^{(l)}]^{\mathrm{T}}$. The variance of $\boldsymbol{\epsilon}^{(l)*}$, $l = 1, \ldots, L$ should also be 1, and its temporal dependence should be the same as that of $\mathbf{r}^*$. Therefore, the null distribution of $S^2(\mathbf{r}^*)$ can be approximated by the empirical distribution $\hat{F}(\cdot)$ of the sample variances $S^2(\boldsymbol{\epsilon}^{(l)*})$, $l = 1, \ldots, L$ of $\boldsymbol{\epsilon}^{(l)*}$. When L is small, to increase the accuracy, we could also generate some block bootstrapped versions of the prewhitened control runs and pool them with the original prewhitened control runs to approximate the null distribution of the testing statistic. The approximate p value is $2\min\{\hat{F}[\mathrm{var}(\mathbf{r}^*)], 1 - \hat{F}[\mathrm{var}(\mathbf{r}^*)]\}$. The same procedure could be applied with a evaluated at $\hat{a}_T$ as a diagnostic test to check H_{02} after allowing $a \neq 1$.

The effect of prewhitening with $\hat{\boldsymbol{\Psi}}_+$ also needs to be tested. The EE method is based on the assumption that the regression error is temporally stationary and that after prewhitening, the

errors at each time become uncorrelated. So we test the composite null hypotheses H_{03}: there is no spatial autocorrelation in the prewhitened residual $\mathbf{r}_t^*$ for all $t = 1, \ldots, T$. For each time t, we use Moran's I statistic to test no zero spatial correlation at that t (Moran 1950; Li et al. 2007). The p value p_t at each time t can be calculated by function Moran.I() from the R package ape (Paradis et al. 2004). To combine the T individual p values to form an overall diagnosis for the prewhitening effect, we use the recently developed Cauchy combination rule to define a combined statistic (Liu and Xie 2020)

$$C_T = \frac{1}{T}\sum_{t=1}^{T} \tan[(0.5 - p_t)\pi].$$

The tail of the null distribution of C_T is well approximated by the standard Cauchy distribution under arbitrary dependency structures among the individual p values. The Cauchy combination test suits our situation well to obtain a single decision about the adequacy of the prewhitening with $\hat{\boldsymbol{\Psi}}_+$. Of note, the prewhitening is only needed for model diagnostics but not for point or interval estimation.

3. Simulation study

a. Simulation settings

To evaluate the performance of the proposed estimator in realistic settings, we conducted simulation studies mimicking detection and attribution analyses of changes in the mean temperature. Both global and regional scales were considered. The global setting was based on $S = 54$ grid boxes of size 40° × 30°. The regional setting was based on eastern North America (ENA), with $S = 21$ grid boxes of size 5° × 5°. For ease of comparison with the results in Li et al. (2021), 5-yr mean temperatures were considered over the time period 1951–2010. Mimicking the application in the next section, the observed data and simulations under the external forcings were anomalies relative to their 30-yr average over 1961–90. Due to this centering, the period of 1961–65 was then removed from the analysis, resulting in $T = 11$ clusters.

For each setting, we first set the true signals $\mathbf{X}$ and the true covariance $\boldsymbol{\Sigma}$ of the TS-dimensional regression error $\boldsymbol{\epsilon}$. Two signals were considered, $\mathbf{X}_1$ for the ANT forcing and $\mathbf{X}_2$ for the NAT forcing. Their true values were set to be the average of 35 and 46 simulations under the ANT and NAT forcings, respectively, from phase 5 of the Coupled Model Intercomparison Project (CMIP5). The true $\boldsymbol{\Sigma}$ was set to be the estimate based on 223 control runs from CMIP5 with a block Toeplitz structure imposed to satisfy the temporal stationarity assumption. See appendix E for details on strategies to make it positive definite and make the temporal correlation decay as time lag increases.

With $\mathbf{X}_1$, $\mathbf{X}_2$, and $\boldsymbol{\Sigma}$ set, we can now generate the observed data and control runs. For $j \in \{1, 2\}$, the estimated signal $\hat{\mathbf{X}}_j = (\mathbf{X}_{j1}^{\mathrm{T}}, \ldots, \mathbf{X}_{jT}^{\mathrm{T}})^{\mathrm{T}}$ was generated from a multivariate normal distribution $N(\mathbf{X}_j, a\boldsymbol{\Sigma}/m_j)$, with $m_1 = m_2 = m \in \{20, 40\}$ and $a \in \{0.5, 1\}$. The regression error $\boldsymbol{\epsilon}$ and the control runs $[\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(L)}]$ were independently generated from a multivariate normal distribution $N(0, \boldsymbol{\Sigma})$, with $L \in \{50, 100, 200\}$. Here $L = 50$ is relatively common in OF studies, and $L = 200$ is possible but not easily obtained unless runs from different climate models are pooled. The observed temperature $\mathbf{Y} = (\mathbf{Y}_1^{\mathrm{T}}, \ldots, \mathbf{Y}_T^{\mathrm{T}})^{\mathrm{T}}$ was generated from the model [Eq. (1)] with $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}} = (1,1)^{\mathrm{T}}$.

For each combination of m and L, we ran 1000 replicates to evaluate the proposed method compared to two TLS-type methods. The first TLS method, denoted by TLS-TS, is the prevailing two-sample approach where the control runs were split into two samples, each giving an estimate of $\boldsymbol{\Sigma}$. The first one was used to prewhiten the data in the point estimation of $\boldsymbol{\beta}$; the second one was used to estimate the variance of the estimator and construct confidence intervals based on the normal approximation (DelSole et al. 2019; Li et al. 2021). Unlike the simulation-based approach of Allen and Stott (2003), no open intervals were expected. The second TLS method, denoted by TLS-PBC, uses all the control runs to estimate $\boldsymbol{\Sigma}$ and calibrates the confidence intervals by parametric bootstrap (Li et al. 2021) with 500 bootstrap replicates. The whole study was run on the high-performance cluster (HPC) of the University of Connecticut.

b. Results

Here we only discuss the simulation results from the setting where the true $\boldsymbol{\Sigma}$ is the block Toeplitz matrix, the true a is 1, and the regression error follows a multivariate normal distribution. More simulation results are provided in the supplement. Table 1 summarizes the bias and the RMSE of the three estimators in all the simulation configurations. All three point estimators appear unbiased in all settings. At the global level, the EE estimator has the smallest RMSE among the three estimators, while the TLS-TS estimator has the largest RMSE for most settings. The difference in RMSE between the EE and TLS-PBC increases as the sample size of control runs decreases or the accuracy in estimating $\boldsymbol{\Sigma}$ decreases. For $L = 50$, the EE has a much smaller RMSE for both scaling factors than TLS-PBC. The RMSE of the EE almost remains the same for different L, which means a small sample size of control runs is sufficient for an EE. In contrast, TLS-PBC needs more control runs to achieve a smaller RMSE. For the ANT forcing, the RMSE of the EE is always smaller than TLS-PBC for all m and L. For the NAT forcing, the EE is better than TLS when $L = 50$ or 100. In the extreme case where $L = 200$ and $m = 20$, TLS-PBC is slightly better than the EE, but this number of L is not easy to obtain. At the regional scale, where the signal-to-noise ratio is lower, the RMSE of the EE is very close to or smaller than those of TLS-TS and TLS-PBC except for the NAT forcing when $m = 20$. In summary, all three methods give unbiased estimators, but the EE has the smallest RMSE for most settings, especially for the ANT scaling factor.

Table 2 summarizes the average widths, the empirical coverage rates, and the average interval score (Gneiting and Raftery 2007) of the 90% confidence intervals constructed from the three methods across all the simulation settings. At both global and regional levels, EE intervals have coverage rates very close to the nominal level in almost all settings. For all three levels of L, even at $L = 50$, their coverage rates are close to the nominal level 90%, with a minimum of 87% for

TABLE 1. Summaries of the bias and RMSE from three methods in the simulation settings.

				$L = 50$		$L = 100$		$L = 200$	
Scale	Forcing	m	Method	Bias	RMSE	Bias	RMSE	Bias	RMSE
Global	ANT	40	TLS-TS	−0.003	0.148	−0.003	0.145	−0.000	0.137
			TLS-PBC	−0.001	0.145	−0.001	0.137	0.002	0.131
			EE	0.000	0.112	0.003	0.111	0.004	0.109
		20	TLS-TS	0.003	0.151	−0.002	0.148	0.000	0.140
			TLS-PBC	0.003	0.149	−0.003	0.139	0.003	0.134
			EE	−0.005	0.117	−0.003	0.118	−0.003	0.115
	NAT	40	TLS-TS	−0.006	0.299	−0.003	0.285	−0.003	0.262
			TLS-PBC	0.001	0.282	−0.002	0.260	0.001	0.241
			EE	−0.005	0.244	0.006	0.242	0.009	0.241
		20	TLS-TS	0.001	0.333	0.002	0.310	−0.007	0.280
			TLS-PBC	−0.003	0.309	0.001	0.280	−0.004	0.264
			EE	−0.043	0.288	−0.023	0.289	−0.010	0.291
Regional	ANT	40	TLS-TS	−0.001	0.175	−0.003	0.179	−0.007	0.178
			TLS-PBC	−0.007	0.181	−0.001	0.176	−0.006	0.174
			EE	−0.003	0.159	−0.002	0.157	−0.001	0.157
		20	TLS-TS	−0.005	0.188	−0.003	0.189	−0.001	0.192
			TLS-PBC	−0.006	0.194	−0.003	0.192	−0.003	0.184
			EE	0.003	0.172	0.005	0.170	0.005	0.171
	NAT	40	TLS-TS	0.005	0.334	0.004	0.343	−0.006	0.328
			TLS-PBC	0.010	0.341	0.007	0.336	0.000	0.327
			EE	0.010	0.327	0.012	0.325	0.011	0.324
		20	TLS-TS	0.010	0.348	0.011	0.353	0.013	0.355
			TLS-PBC	0.013	0.355	0.006	0.348	0.012	0.350
			EE	−0.001	0.365	−0.000	0.369	0.004	0.370

both ANT and NAT scaling factors, regardless of the noise level of the measurement error controlled by m. TLS-TS intervals have the lowest coverage, as reported in Li et al. (2021). Their coverage rate improves as L increases but remains unsatisfying and lower than 80% for both ANT and NAT even when $L = 200$. TLS-PBC requires $L = 100$ to reach the nominal level of coverage rate for the ANT scaling factor and $L = 200$ for the NAT scaling factor, whereas it becomes conservative for the ANT scaling factor when $L = 200$, especially at the regional level. In addition, EE intervals are often narrower than or comparable to TLS-PBC intervals when both methods provide desired coverage rates. The superiority of EE intervals regarding the widths and coverage rates is reflected by the lower interval scores. The undercoverage of TLS-TS intervals and the conservativeness of TLS-PBC intervals lead to larger interval scores. At the regional level, where the signal-to-noise ratio is lower, all intervals become wider than at the global level. EE intervals still give the proper coverage rate with a sample size of control runs as small as $L = 50$. To sum up, the proposed EE intervals provide valid confidence intervals with the desired level of coverage rate and a narrower width in most cases at a much lower computational cost than TLS-PBC intervals.

c. *Additional simulations*

To investigate the effect of differences in the magnitude of climate variability between observations and model simulations and the effect of non-Gaussian residuals, we also conducted additional simulations with $a \in \{0.5, 1\}$, $\boldsymbol{\epsilon}$ following a multivariate normal or t distribution with a degree of freedom of 15 and the true $\boldsymbol{\Sigma}$ set as a block Toeplitz matrix or the regularized linear shrinkage estimator from control runs (see section S2 in the online supplemental material). For the point estimators, the EE method is still almost unbiased and has the smallest RMSE across all the settings, while the two TLS methods showed a large bias when $a = 0.5$, especially for the NAT forcing. For the confidence interval, the EE method maintains a close-to-nominal coverage rate, while TLS-TS has a much lower coverage rate. TLS-PBC also could not reach the nominal coverage rate when the sample size of control runs is small, e.g., $L = 50$. These results showed the robustness of the EE method against non-Gaussian error distribution and the necessity of relaxing the well-accepted assumption $a = 1$. The advantage of the EE method would be even more obvious when the tails of the multivariate t distribution are heavier with smaller degrees of freedom.

We also investigated the performance of the EE method with a larger number of sites ($S = 108$), with $\boldsymbol{\epsilon}$ following a multivariate normal or t distribution (see section S2). Results suggest that the EE method maintains its unbiasedness, efficiency, and validity in almost all cases. Even though a minor bias is observed for the NAT forcing when the measurement error is large ($m = 20$) and the number of control runs is small ($L = 50$), it becomes negligible with the increase of L. For TLS-TS and TLS-PBC, the NAT forcing is also a challenging case, especially when $\boldsymbol{\epsilon}$ follows a multivariate t distribution; the bias and RMSE are significantly increased compared to their performances with $S = 54$. In general, the EE method retains its advantages with a large S.

TABLE 2. Summaries of the average length, empirical coverage percentages (CPs), and average interval score of the 90% confidence intervals constructed from three methods in the simulation settings.

				$L = 50$			$L = 100$			$L = 200$		
Scale	Forcing	m	Method	Width	CP (%)	Score	Width	CP (%)	Score	Width	CP (%)	Score
Global	ANT	40	TLS-TS	0.285	61.1	1.235	0.275	65.8	1.087	0.294	71.3	1.004
			TLS-PBC	0.411	84.4	1.042	0.446	89.9	1.029	0.458	91.9	1.008
			EE	0.362	89.3	0.818	0.355	90.0	0.804	0.353	89.4	0.795
		20	TLS-TS	0.405	65.3	1.403	0.312	68.3	1.110	0.320	73.8	1.012
			TLS-PBC	0.423	83.3	1.064	0.466	90.4	1.039	0.479	91.4	1.042
			EE	0.373	87.0	0.874	0.371	87.9	0.862	0.372	88.6	0.861
	NAT	40	TLS-TS	0.903	62.4	3.048	0.595	68.0	2.187	0.588	71.6	1.899
			TLS-PBC	0.793	83.0	1.964	0.798	86.7	1.875	0.786	89.9	1.782
			EE	0.768	88.0	1.766	0.761	87.7	1.746	0.760	87.6	1.738
		20	TLS-TS	2.861	65.8	7.023	0.819	70.4	2.665	0.695	77.3	2.057
			TLS-PBC	0.906	83.9	2.228	0.900	89.2	2.041	0.887	90.5	1.986
			EE	0.879	87.4	2.044	0.893	88.1	2.071	0.908	88.9	2.086
Regional	ANT	40	TLS-TS	0.316	61.7	1.371	0.356	67.8	1.305	0.411	73.9	1.284
			TLS-PBC	0.489	81.5	1.289	0.589	90.7	1.309	0.638	93.5	1.368
			EE	0.510	87.8	1.178	0.508	89.5	1.159	0.508	88.6	1.158
		20	TLS-TS	0.373	61.5	1.553	0.385	66.0	1.423	0.447	73.1	1.375
			TLS-PBC	0.509	80.0	1.399	0.614	87.6	1.423	0.666	92.5	1.447
			EE	0.538	89.1	1.259	0.539	88.9	1.243	0.539	88.9	1.246
	NAT	40	TLS-TS	0.643	65.7	2.607	0.712	68.9	2.555	0.832	80.0	2.317
			TLS-PBC	0.867	79.1	2.467	1.039	86.2	2.455	1.178	92.6	2.548
			EE	1.052	88.5	2.442	1.053	88.6	2.431	1.049	88.4	2.412
		20	TLS-TS	0.938	65.6	3.087	0.836	73.5	2.677	0.998	82.1	2.604
			TLS-PBC	0.911	78.8	2.540	1.122	87.6	2.609	1.282	92.2	2.761
			EE	1.214	90.6	2.710	1.224	90.5	2.731	1.233	90.4	2.747

In typical OF analyses, the ANT forcing is often obtained by subtracting the NAT forcing from the ALL forcing, which causes correlated measurement errors between ANT and NAT. Thus, to apply the EE method, we suggest estimating $\boldsymbol{\beta}_{\mathrm{ALL}}$ and $\boldsymbol{\beta}_{\mathrm{NAT}}$ first and then constructing the estimators for $\boldsymbol{\beta}_{\mathrm{ANT}}$ and $\boldsymbol{\beta}_{\mathrm{NAT}}$ (see details in section 2a). Although the estimating procedure is slightly different than the main simulation results presented in section 2b, which assumes that the two forcings are independent, we conducted additional simulations in section S2 mimicking the data generation and processing procedure in section 2b. The results suggest that the EE method is valid as expected and maintains its advantages over TLS methods.

As our EE method has assumed temporal stationarity (assumption 3), we further investigate if the same assumption would also improve the prevailing TLS estimator. The results in section S2 shows that the temporal stationarity assumption does have an effect in reducing the RMSE of the point estimator compared to TLS-TS and TLS-PBC at the regional scale with Gaussian errors; however, the RMSE is still larger than that of the EE estimator in most settings. Also, imposing the temporal stationarity assumption does not seem to reduce the RMSE at the global scale, possibly due to a higher spatial dimension. When the regression errors are heavy tailed (i.e., t distributed), the adapted TLS estimator is also less efficient than the EE. For confidence intervals, we assess the performance of TLS adapted with the proposed pseudo bootstrap procedure, as the prevailing TLS could not provide confidence intervals with the desired coverage rates. The results in section S2 suggest that the modified TLS method has close-to-nominal coverage rate with Gaussian errors, but it could be very conservative when the error distribution is non-Gaussian (e.g., t distribution), leading to a surprisingly large interval width. Therefore, the stationarity assumption does improve TLS-type methods overall, but our proposed EE method has clear advantages.

4. Application

To demonstrate the performance of the proposed approach in real-world applications, we conducted optimal fingerprinting analyses on the annual mean near-surface air temperature at the global (GL), continental, and subcontinental scales during 1951–2020 (Zhang et al. 2006). At the continental (and larger) scale, we consider the Northern Hemisphere (NH), NH midlatitude between 30° and 70° (NHM), Eurasia (EA), and North America (NA). At the subcontinental scale, we consider western North America (WNA), central North America (CNA), ENA, southern Canada (SCA), and southern Europe (SEU) (Giorgi and Francisco 2000).

a. Data preparation

Our observational data Y came from the HadCRUT4 dataset (Morice et al. 2012), which contains monthly anomalies of near-surface air temperature on 5° × 5° grid boxes relative to the 30-yr average over 1961–90. The annual mean temperature anomalies were computed from the monthly values. The

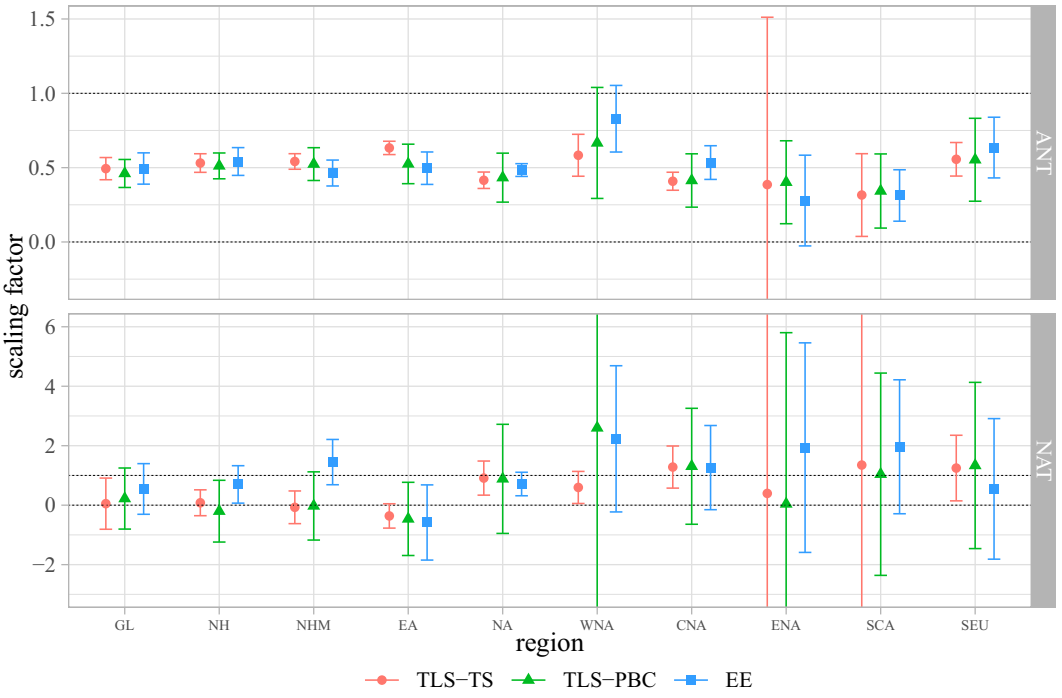

FIG. 1. Estimated scaling factors with 90% confidence intervals of the ANT and NAT forcings at the global, continental, and subcontinental scales during 1951–2020. The confidence intervals for TLS-TS and TLS-PBC in WNA, ENA, and SCA regions are truncated for better visualization.

annual value was considered missing if there were more than 3 monthly values missing in the year. Our analyses were conducted on nonoverlapping 5-yr mean temperature anomalies. If there were at least three annual values available within a 5-yr period, we computed the 5-yr average temperatures. After removal of the 1961–65 period due to centering, we have $T = 13$ values of 5-yr averages at each grid box. The spatial dimension is also reduced for the global and continental scale analyses by averaging available 5-yr 5° × 5° grid boxes within a larger box. The final box sizes used in the global and continental scale analyses are as follows: GL and NH, 40° × 30°; NHM, 40° × 10°; EA, 10° × 20°; and NA, 10° × 5°. For the subcontinental scale analyses, we only aggregated the boxes of SCA to 10° × 10° boxes. Details about the boundaries of the regions and data availability are summarized in appendix F.

We obtained estimated signals and control runs from large ensemble simulations conducted with CanESM5 climate model simulations (Swart et al. 2019). Specifically, let $\tilde{\mathbf{X}}_{\mathrm{NAT}}$ be the estimated NAT signal from CanESM5 simulations with 50 runs. Since the ALL forcing in CanESM5 ended in 2014, we appended the simulations using those under the ssp245 forcing (Danabasoglu 2019). The estimated ALL signal $\mathbf{X}_{\mathrm{ALL}}$ was obtained from 50 such appended runs. During the processing, the same missing pattern of $\mathbf{Y}$ was imposed, and the same average and aggregation procedures of $\mathbf{Y}$ were applied to obtain $\tilde{\mathbf{X}}_{\mathrm{ALL}}$ and $\tilde{\mathbf{X}}_{\mathrm{NAT}}$. Since they were centered by the 30-yr average over 1961–90, the first 5-yr period 1961–90 was excluded as well. The control runs were obtained from 50 runs under solar and volcanic forcings only, 30 runs under anthropogenic aerosols only, and 50 runs under greenhouse gas only from CanESM5 simulations; after centering under each forcing, the intraensemble variation gave 127 runs. At each grid box, a long-term linear trend was removed from the control simulations from each climate model separately.

The scaling factor of the ANT forcing is of the most interest. In typical OF analyses, one would obtain $\tilde{\mathbf{X}}_{\mathrm{ANT}}$ from subtracting $\tilde{\mathbf{X}}_{\mathrm{ALL}}$ by $\tilde{\mathbf{X}}_{\mathrm{NAT}}$ and estimate the scaling factors of ANT and NAT directly. Such transformation leads to correlated measurement errors ν_1 and ν_2, which violates assumption 1. Thus, we first fit the model and obtain the estimated scaling factors $\hat{\boldsymbol{\beta}}_{\mathrm{ALL}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{NAT}}$; then the estimated scaling factor for the ANT forcing is $\hat{\boldsymbol{\beta}}_{\mathrm{ANT}} = \hat{\boldsymbol{\beta}}_{\mathrm{ALL}}$, and the estimated scaling factor for the NAT forcing is $\hat{\boldsymbol{\beta}}_{\mathrm{ALL}} + \hat{\boldsymbol{\beta}}_{\mathrm{NAT}}$. The variances and confidence intervals can be obtained according to the respective linear transformation of $\hat{\boldsymbol{\beta}}_{\mathrm{ALL}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{NAT}}$.

b. *Estimation results*

Figure 1 shows the estimated ANT and NAT scaling factors with 90% confidence intervals based on TLS-TS, TLS-PBC, and the proposed EE method. For both forcings, the point estimators of the three methods are different, especially on the subcontinental scale. Given the unbiasedness and robustness of the EE method demonstrated in simulation studies, its results are more trustable. It is also worth noting that the point estimators for the ANT forcing are around 0.5 across almost all regions, indicating that the signal of the simulated ANT forcing in the CanESM5 model is twice as big in magnitude as expected from observations. This discovery is consistent with

TABLE 3. Estimated a and p values of model diagnostic tests for the EE method.

Region	$\hat{a}_T$	H_{01}	H_{02}		H_{03}
			$a = 1$	$a = \hat{a}_T$	
GL	0.325	0.000	0.000	0.406	0.404
NH	0.690	0.000	0.002	0.605	0.054
NHM	0.891	0.331	0.236	0.580	0.418
EA	0.651	0.000	0.000	0.331	0.847
NA	0.326	0.000	0.000	0.231	0.218
WNA	0.203	0.000	0.000	0.321	0.257
CNA	1.021	0.829	0.854	0.783	0.230
ENA	0.115	0.000	0.000	0.413	0.165
SCA	0.470	0.000	0.033	0.606	0.749
SEU	0.281	0.000	0.000	0.509	0.199

existing literature indicating that CanESM4 warms foo fast (Gillett et al. 2021, Fig. 2). Regarding the confidence intervals, the EE method provides narrower intervals than TLS-PBC on most scales, which is supported by the potential conservativeness of the TLS-PBC intervals observed in simulation studies and discussed in Li et al. (2021). The often-narrower intervals of TLS-TS are questionable owing to the undercoverage issue discussed in the simulation studies.

Across the detection and attribution analyses from the three methods, results are similar for the ANT forcing and substantially different for the NAT forcing, owing to its weaker signal. For the ANT forcing, TLS-PBC and the EE lead to both detection and attribution statements at WNA, while TLS-TS only supports the detection statement. At the ENA region, only TLS-PBC leads to the detection statement, while the other two methods do not, as their confidence intervals cover 0. In other regions, all three methods support the detection but not attribution statements for the ANT forcing. For the NAT forcing, the EE method claims its detection and attribution at NH, NHM, and NA regions; TLS-TS supports the detection and attribution statements at NA, CNA, WNA, and SEU regions; TLS-PBC does not support either detection or attribution statements in any region. Although we cannot claim which method is more accurate for this single analysis, we believe that a thorough comparison of the performances of these three methods has been reflected in the simulation studies.

c. *Model diagnostics*

Table 3 summarizes the p values of the three hypotheses described in section 2d. In particular, we consider two scenarios where the prewhitened residuals are calculated based on 1) estimated $\hat{a}$ or 2) prefixed $a = 1$. The significance level is set as $\alpha = 0.05$. The estimated $\hat{a}$ is not close to 1 in most regions, except NHM and CNA. Accordingly, for those regions except NHM and CNA, we observe that both H_0^1 and H_0^2 (assuming $a = 1$) are rejected, and H_0^2 based on estimated $\hat{a}$ is not rejected as expected. Thus, the assumption that $a = 1$ is violated for most regions in this dataset. Table 3 also supports the conclusion that there is no spatial autocorrelation in $\mathbf{r}^*(\hat{\boldsymbol{\beta}})$ in most regions. Hence, applying the proposed EE method to this dataset is reasonable, as all assumptions are valid, while the results of the two TLS methods are questionable owing to the violation of $a = 1$ in most regions.

5. Discussion

Our methodological contributions are threefold. First, we propose an efficient, bias-corrected EE approach to estimate the scaling factors in OF. Under the temporal stationarity assumption about the natural internal variability, $\boldsymbol{\Sigma}$ has a block Toeplitz structure, with each time point as a block, which greatly reduces the number of parameters in $\boldsymbol{\Sigma}$. The diagonal blocks of $\boldsymbol{\Sigma}$, which capture the spatial dependence, can be estimated more reliably than the other blocks of $\boldsymbol{\Sigma}$. Although we discarded temporal dependence in the EE method, which may lead to efficiency loss, the gain due to much-reduced uncertainty in estimating spatial dependence has resulted in a much-improved estimator. The same assumption applied to the TLS method also leads to improvement, but the EE method still has advantages when the spatial dimension is higher or the error distribution is heavy tailed. Unlike approaches that rely on distributional specifications (Hannart 2016; Katzfuss et al. 2017), no distributional assumption beyond the first two moments is needed. Our second contribution is the valid confidence intervals for the scaling factors with close-to-nominal coverage rates. The confidence intervals are constructed with a novel pseudo residual bootstrap method that takes advantage of the available control runs that preserve both spatial and temporal dependences. This pseudo residual bootstrap method provides close-to-nominal coverage rates at a much lower computation cost and with no distributional assumption on the regression errors compared to the TLS-PBC approach of Li et al. (2021). The nonparametric feature of the procedure makes it applicable to other methods too, such as TLS. Finally, our method incorporates a scale parameter a in the variances to account for the proportion of the variances among different climate simulation models, which provides additional flexibility. We further provide different ways to test the commonly made assumption $a = 1$ as part of model diagnostics. When the assumption $a = 1$ is violated, our simulation results in the supplemental material suggest the TLS methods could be heavily biased.

The proposed OF framework is promising as a solid, easy-to-implement alternative to the TLS approach in practice. The undercoverage of the confidence intervals from the TLS approach has not attracted attention until recently (DelSole et al. 2019; Li et al. 2021), but with no completely satisfying solutions. Our methods give not only point estimates with a smaller RMSE, but also confidence intervals with the desired coverage rates. Further, we have weaker variance scale assumptions on climate model simulations under external forcings and much lower computing costs. The methods could be a reasonable solution to the long-overlooked coverage issue of the confidence intervals. When the sample size of control runs L is insufficient, the runs under each external forcing of interest could be centered and appropriately scaled to supplement the control runs in estimating $\mathbf{B}$ and, hence, the variance of $\hat{\boldsymbol{\beta}}_T$. When $a \neq 1$, both TLS-TS and TLS-PBC have biased point estimators, while the EE method can estimate a and

remains valid in confidence interval coverage rates. In practice, as the application shown in section 2b, the qualitative conclusions of detection and attribution from TLS may be the same as those from EEs. Nevertheless, a reexamination of the main results supporting the attribution assessment of Eyring et al. (2021) using the EE method would not be a natural yet much feasible task with our software implementation.

The EE approach can be extended in several directions. The temporal aggregation, such as the 5-yr average, may be relaxed. One could use annual data with a more sophisticated model for the signals under each forcing, such as B splines, as used in Wang et al. (2021). The use of control runs to estimate internal variability assumes that internal climate variability is not affected by external forcing, which may need careful consideration for regional data. Several works suggest that internal climate variability may change regionally (Bonan et al. 2021; Swart et al. 2015), and estimating it from forced run residuals rather than from control runs could potentially address this issue (Ribes et al. 2013). The climate model differences were discarded in our study. That is, the m_j runs under the jth forcing were treated as if they were the same, while in reality, they can be from different climate models. A more realistic model could incorporate a random effect to capture the heterogeneity among different climate models in estimating the signals under each external forcing, which will also allow pooling simulations from different models to reach a larger sample size. Adding random effects into the EE approach merits further investigation. Similarly, the scale parameter a could also be model specific if multiple climate models are considered.

Acknowledgments. We are grateful for the feedback from participants in the presentation of an earlier version of this work at the International Detection and Attribution Group (IDAG) virtual seminar series. We thank Dr. Timothy DelSole, Dr. Francis Zwiers, and Dr. Dáithí Stone for their constructive suggestions.

Data availability statement. The data presented in numerical studies are available online as follows: 1) CMIP5, https://pcmdi.llnl.gov/mips/cmip5/; 2) HadCRUT4, https://www.metoffice.gov.uk/hadobs/hadcrut4/; and 3) CanESM5, https://crd-data-donnees-rdc.ec.gc.ca/.

APPENDIX A

Basics of Estimating Equations

Before introducing the general setting, consider the standard linear regression setting. Let $\{(\mathbf{Y}_i, \mathbf{X}_i): i = 1, \ldots, n\}$ be a random sample of response variable $\mathbf{Y}$ and a $p \times 1$ covariate vector $\mathbf{X}$. The regression model is

$$E(\mathbf{Y}_i|\mathbf{X}_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, \tag{A1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector to be estimated. The least squares estimator $\hat{\beta}_n$ minimizes the sum of squares as follows:

$$\hat{\beta}_n = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_i^n (\mathbf{Y}_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2.$$

The solution is equivalently obtained by equating the first derivative of the objective function with respect to $\boldsymbol{\beta}$ to zero,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i(\mathbf{Y}_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) = 0, \tag{A2}$$

and solving for $\boldsymbol{\beta}$. Equation (A2), also known as the normal equation of the least squares problem, is an EE. Its left-hand side is called an estimating function. The estimating function has an expectation of zero, as specified by the linear regression [Eq. (A1)], which only specifies a moment condition (mean). If the regression error $\boldsymbol{\epsilon}_i = \mathbf{Y}_i - \mathbf{X}_i^{\mathrm{T}}$ is further assumed to be normally distributed, Eq. (A2) also coincides with the score equation (first derivative of the log likelihood equated to zero). Nonetheless, the EE estimator is robust to distributional misspecification because Eq. (A2) specifies nothing beyond the expectation of the regression error.

The method of EEs is a general strategy for parameter estimation in statistical applications. The estimator is the root (zero) of a set of data-dependent functions called estimating functions. In particular, let $\{\mathcal{X}_i : i = 1, \ldots, n\}$ be the observed data of sample size n, which are not necessarily independent copies; let $\boldsymbol{\theta}$ be a $p \times 1$ parameter vector to be estimated. Consider a p-dimensional estimating function $G(\boldsymbol{\theta}; \mathcal{X}_i)$, $i = 1, \ldots, n$, which depends on both $\boldsymbol{\theta}$ and the data. If the expectation of $G(\boldsymbol{\theta}; \mathcal{X}_i)$ with respect to $\mathcal{X}_i$ is zero, then we have an EE of

$$\Phi_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n G(\boldsymbol{\theta}; \mathcal{X}_i) = 0. \tag{A3}$$

The estimator $\hat{\theta}_n$ is the root of Eq. (A3), i.e., $\boldsymbol{\Psi}_n(\hat{\theta}) = 0$.

The framework of EEs is very general. When the likelihood is available from a fully specified model, Eq. (A3) can be the score equation (the first derivative of the log likelihood). In scenarios where the estimator is obtained by optimizing an objective function (e.g., least squares, minimum risk), Eq. (A3) can be the first derivative of the objective function with respect to $\boldsymbol{\theta}$. Nonetheless, Eq. (A3) needs not correspond to an optimization problem. The only requirement is $E[G(\boldsymbol{\theta}; \mathcal{X}_i)] = 0$, $i = 1, \ldots, n$, which is much weaker than full likelihood specification. Moment conditions are often used to construct estimating functions. This is why EE estimators are more robust than likelihood estimators in general.

The asymptotic properties of $\hat{\theta}_n$ as $n \to \infty$ can be derived under fairly general regularity conditions. The estimator $\hat{\theta}_n$ converges to the true θ in probability, i.e., $\hat{\theta}_n$ is asymptotically unbiased. The distributional properties of $\hat{\theta}_n$ are inherited from the behavior of the estimating functions. A first-order Taylor approximation of $\Phi_n(\hat{\theta}_n)$ around the true $\boldsymbol{\theta}$ gives

$$0 = \Phi_n(\hat{\theta}_n) \approx \Phi_n(\boldsymbol{\theta}) + \dot{\Phi}_n(\boldsymbol{\theta})(\hat{\theta}_n - \boldsymbol{\theta}),$$

where $\dot{\Phi}_n = \partial\Phi_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\circ}$. Therefore,

$$\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta}) \approx -[\dot{\Phi}_n(\boldsymbol{\theta})]^{-1}[\sqrt{n}\Phi_n(\boldsymbol{\theta})],$$

where the first term (an average) converges in probability to the true expectation of the estimating function $\dot{G}(\boldsymbol{\theta}) = E[\dot{G}(\boldsymbol{\theta}; \mathcal{X}_i)]$, and the second term is asymptotically normal by the central

limit theorem (with possibly dependent data) with variance $V(\boldsymbol{\theta}) = \lim_{n\to\infty} \mathrm{cov}[\sqrt{n}\Phi_n(\boldsymbol{\theta})]$. So the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta})$ has a sandwich form $\dot{\mathbf{G}}^{-1}(\boldsymbol{\theta})V(\boldsymbol{\theta})[\dot{\mathbf{G}}^{-1}(\boldsymbol{\theta})]^{\mathrm{T}}$. An estimator of the asymptotic variance can be obtained by plugging the unknown $\boldsymbol{\theta}$ by its estimator $\hat{\theta}_n$.

In the OF context, we constructed unbiased EEs after deriving the bias in appendix B. The sample size n is the number of temporal points T, where each time point is treated as a cluster. The data $\mathcal{X}_t$ from cluster t contain $\mathbf{Y}_t$ and $\tilde{\mathbf{X}}_t$ in section 2a. Because of the temporal dependence, the middle matrix $\mathbf{V}(\boldsymbol{\theta})$ in the sandwich matrix is hard to estimate, which motivated our pseudo bootstrap procedure in section 2c.

APPENDIX B

Bias from Using $\tilde{\mathbf{X}}$ in Place of $\mathbf{X}$

The EE [Eq. (3)] when $\mathbf{X}$ is substituted by $\tilde{\mathbf{X}}$ is no longer unbiased. As assumption 2 and assumption 3 indicate the measurement errors are also stationary, we only need to derive the EE for one cluster. For cluster t, let $\boldsymbol{\nu}_{(t)} = (\nu_{t1}, \ldots, \nu_{tJ})$, which is an $S \times J$ matrix. We have

$$
\begin{aligned}
&E\{\tilde{\mathbf{X}}_t^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}(\mathbf{Y}_t - \tilde{\mathbf{X}}_t\boldsymbol{\beta})\}\\
&= E([\mathbf{X}_t + \boldsymbol{\nu}_{(t)}]^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}\{\mathbf{Y}_t - [\mathbf{X}_t + \boldsymbol{\nu}_{(t)}]\boldsymbol{\beta}\})\\
&= -E[\boldsymbol{\nu}_{(t)}^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\nu}_{(t)}]\boldsymbol{\beta}\\
&= -\mathrm{diag}\{E(\boldsymbol{\nu}_{t1}^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\nu}_{t1}), \ldots, E(\boldsymbol{\nu}_{tJ}^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\nu}_{tJ})\}\boldsymbol{\beta}\\
&= -\mathrm{diag}\{\mathrm{tr}[\boldsymbol{\Sigma}_t^{-1}E(\boldsymbol{\nu}_{t1}\boldsymbol{\nu}_{t1}^{\mathrm{T}})], \ldots, \mathrm{tr}[\boldsymbol{\Sigma}_t^{-1}E(\boldsymbol{\nu}_{tJ}\boldsymbol{\nu}_{tJ}^{\mathrm{T}})]\}\boldsymbol{\beta}\\
&= -\mathrm{diag}\{\mathrm{tr}(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Omega}_{t1}), \ldots, \mathrm{tr}(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Omega}_{tJ})\}\boldsymbol{\beta}\\
&= -aS\,\mathrm{diag}\left(\frac{1}{m_1}, \ldots, \frac{1}{m_J}\right)\boldsymbol{\beta}.
\end{aligned}
$$

The third equation above is because $\boldsymbol{\nu}_{tj}$ and $\boldsymbol{\nu}_{tj'}$ are uncorrelated for $j \neq j'$. In the last equation, we are able to drop the t indices because of assumption 2. This derivation is the basis for constructing unbiased EEs in terms of $\tilde{\mathbf{X}}$.

APPENDIX C

Estimation of a

Define the block residual $\mathbf{r}_t(\boldsymbol{\beta}) = \mathbf{Y}_t - \tilde{\mathbf{X}}_t\boldsymbol{\beta}$ for $t = 1, \ldots, T$; then

$$
\begin{aligned}
\mathrm{var}\{\mathbf{r}_t(\boldsymbol{\beta})\} &= \mathrm{var}(\mathbf{Y}_t - \tilde{\mathbf{X}}_t\boldsymbol{\beta}) = \mathrm{var}(\mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{X}_t\boldsymbol{\beta} - \boldsymbol{\nu}\boldsymbol{\beta})\\
&= \mathrm{var}(\boldsymbol{\epsilon}) + \mathrm{var}(\boldsymbol{\nu}\boldsymbol{\beta}) = \delta^2(\boldsymbol{\beta}, a)\boldsymbol{\Sigma}_t,
\end{aligned}
$$

where $\delta(\boldsymbol{\beta}, a) = (1 + a\sum_{j=1}^{J}\boldsymbol{\beta}_j^2/m_j)^{1/2}$. Note that

$$
\mathrm{var}\{\boldsymbol{\Psi}_t^{-1/2}\mathbf{r}_t(\boldsymbol{\beta})\} = \boldsymbol{\Psi}_t^{-1/2}\delta^2(\boldsymbol{\beta}, a)\boldsymbol{\Sigma}_t\boldsymbol{\Psi}_t^{-1/2} = \delta^2(\boldsymbol{\beta}, a)\boldsymbol{\Sigma}_t^{-1/2}\frac{\boldsymbol{\Sigma}_t}{a}\boldsymbol{\Sigma}_t^{-1/2}
$$

$$
= \frac{\delta^2(\boldsymbol{\beta}, a)}{a} = \frac{\left(1 + a\sum_{j=1}^{J}\boldsymbol{\beta}_j^2/m_j\right)}{a}.
$$

A feasible EE of a is

$$
aS_{r,T}^2 - \left(1 + a\sum_{j=1}^{J}\hat{\boldsymbol{\beta}}_{T,j}^2/m_j\right) = 0, \tag{C1}
$$

where $S_{r,T}^2$ is the sample variance of $\hat{\boldsymbol{\Psi}}_+^{-1/2}\mathbf{r}_t(\hat{\boldsymbol{\beta}})$. A closed-form estimator of a is

$$
\hat{a}_T = \frac{1}{S_{r,T}^2 - \sum_{j=1}^{J}\hat{\boldsymbol{\beta}}_{Tj}^2/m_j}. \tag{C2}
$$

From the derivation, $\hat{a}_T$ is asymptotically consistent for the true a, i.e., it converges to a as $T \to \infty$. The variance $\hat{a}_T$ can be obtained jointly with $\hat{\boldsymbol{\beta}}_T$, which can be used to construct confidence intervals and conduct a hypothesis test about a; see details in the supplemental material.

APPENDIX D

Asymptotic Normality of $\hat{\boldsymbol{\beta}}_T$

Since $\hat{\boldsymbol{\beta}}_T$ solves the EE [Eq. (6)], a Taylor expansion of the equation at the true parameter value $\boldsymbol{\beta}$ gives

$$
\begin{aligned}
0 &= T^{-1/2}\sum_{t=1}^{T}G_t(\hat{\boldsymbol{\beta}}_T; \hat{\boldsymbol{\Sigma}}_+)\\
&= T^{-1/2}\sum_{t=1}^{T}G_t(\boldsymbol{\beta}; \hat{\boldsymbol{\Sigma}}_+) + \left\{T^{-1}\sum_{t=1}^{T}\partial G_t(\boldsymbol{\beta}; \hat{\boldsymbol{\Sigma}}_+)/\partial\boldsymbol{\beta}^{\mathrm{T}}\right\}\\
&\quad \times T^{1/2}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}) + o_p(1).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
T^{1/2}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}) = -&\left\{T^{-1}\sum_{t=1}^{T}\partial G_t(\boldsymbol{\beta}; \hat{\boldsymbol{\Sigma}}_+)/\partial\boldsymbol{\beta}^{\mathrm{T}}\right\}^{-1}\\
&\times T^{-1/2}\sum_{t=1}^{T}G_t(\boldsymbol{\beta}) + o_p(1).
\end{aligned}
$$

Since G_t is stationary over time, under mild conditions, $T^{-1/2}\sum_{t=1}^{T}G_t(\boldsymbol{\beta})$ converges in distribution to a normal distribution $N(0, B)$. Further, $T^{-1}\sum_{t=1}^{T}\partial G_t(\boldsymbol{\beta}; \hat{\boldsymbol{\Sigma}}_+)/\partial\boldsymbol{\beta}^{\mathrm{T}} \to A$ by the law of large numbers. The asymptotic normality of $\hat{\boldsymbol{\beta}}_T$ then follows.

APPENDIX E

Constructing $\boldsymbol{\Sigma}$ with Block Toeplitz Structure

In the simulation, the true covariance with a block Toeplitz structure was constructed in the following way. First, we calculated the sample covariance matrix from control runs, denoting it as $\tilde{\boldsymbol{\Sigma}}$. Then, we imposed a block Toeplitz structure to $\tilde{\boldsymbol{\Sigma}}$ by taking the average to main diagonal blocks and each off-diagonal block, respectively. If necessary, we truncated some terms of off-diagonal blocks to make the temporal correlation decay as the time lag increases. Finally, the linear shrinkage method was used to estimate the true $\boldsymbol{\Sigma}$ based on $\tilde{\boldsymbol{\Sigma}}$ so that it is positive definite.

APPENDIX F

Details of the 10 Spatial Scales in Section 4

The names, coordinate ranges, spatiotemporal dimensions, and dimension of observation after removing missing values of the 10 spatial scales in section 4 are presented in Table F1.

TABLE F1. Details of the names, coordinate ranges, spatiotemporal dimensions (S and T), and dimension of observation (ST) after removing missing values of the 10 scales in section 4.

Acronym	Region	Longitude (°E)	Latitude (°N)	Grid size	S	T	ST
GL	Global	−180/180	−90/90	40 × 30	54	14	572
NH	Northern Hemisphere	−180/180	0/90	40 × 30	27	14	297
NHM	Northern Hemisphere (30°–70°N)	−180/180	30/70	40 × 10	36	14	396
EA	Eurasia	−10/180	30/70	10 × 20	38	14	418
NA	North America	−130/−50	30/60	10 × 5	48	14	512
WNA	Western North America	−130/−105	30/60	5 × 5	30	14	329
CNA	Central North America	−105/−85	30/50	5 × 5	16	14	176
ENA	Eastern North America	−85/−50	15/30	5 × 5	21	14	231
SCA	Southern Canada	−110/−10	50/70	10 × 10	20	14	220
SEU	South Europe	−10/40	35/50	5 × 5	30	14	330

REFERENCES

Allen, M. R., and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dyn.*, **15**, 419–434, https://doi.org/10.1007/s003820050291.

——, and P. A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting, part I: Theory. *Climate Dyn.*, **21**, 477–491, https://doi.org/10.1007/s00382-003-0313-9.

Bindoff, N. L., and Coauthors, 2013: Detection and attribution of climate change: From global to regional. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 867–952.

Bonan, D. B., F. Lehner, and M. M. Holland, 2021: Partitioning uncertainty in projections of Arctic Sea ice. *Environ. Res. Lett.*, **16**, 044002, https://doi.org/10.1088/1748-9326/abe0ec.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, 2006: *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. CRC Press, 488 pp.

Danabasoglu, G., 2019: NCAR CESM2 model output prepared for CMIP6 ScenarioMIP ssp245. Earth System Grid Federation, accessed 9 March 2022, https://doi.org/10.22033/ESGF/CMIP6.7748.

DelSole, T., L. Trenary, X. Yan, and M. K. Tippett, 2019: Confidence intervals in optimal fingerprinting. *Climate Dyn.*, **52**, 4111–4126, https://doi.org/10.1007/s00382-018-4356-3.

Eyring, V., and Coauthors, 2021: Human influence on the climate system. *Climate Change 2021: The Physical Science Basis*, V. Masson-Delmotte et al., Eds., Cambridge University Press, 423–552.

Gillett, N. P., and Coauthors, 2021: Constraining human contributions to observed warming since the pre-industrial period. *Nat. Climate Change*, **11**, 207–212, https://doi.org/10.1038/s41558-020-00965-9.

Giorgi, F., and R. Francisco, 2000: Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Climate Dyn.*, **16**, 169–182, https://doi.org/10.1007/PL00013733.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Godambe, V. P., 1991: *Estimating Functions*. Oxford University Press, 356 pp.

Hannart, A., 2016: Integrated optimal fingerprinting: Method description and illustration. *J. Climate*, **29**, 1977–1998, https://doi.org/10.1175/JCLI-D-14-00124.1.

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. *J. Climate*, **9**, 2281–2306, https://doi.org/10.1175/1520-0442(1996)009<2281:DGGICC>2.0.CO;2.

——, and Coauthors, 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 663–745.

Heyde, C. C., 2008: *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer Science and Business Media, 246 pp.

Huntingford, C., P. A. Stott, M. R. Allen, and F. H. Lambert, 2006: Incorporating model uncertainty into attribution of observed temperature change. *Geophys. Res. Lett.*, **33**, L05710, https://doi.org/10.1029/2005GL024831.

Katzfuss, M., D. Hammerling, and R. L. Smith, 2017: A Bayesian hierarchical model for climate change detection and attribution. *Geophys. Res. Lett.*, **44**, 5720–5728, https://doi.org/10.1002/2017GL073688.

Li, H., C. A. Calder, and N. Cressie, 2007: Beyond Moran's *I*: Testing for spatial dependence based on the spatial autoregressive model. *Geogr. Anal.*, **39**, 357–375, https://doi.org/10.1111/j.1538-4632.2007.00708.x.

Li, Y., K. Chen, J. Yan, and X. Zhang, 2021: Uncertainty in optimal fingerprinting is underestimated. *Environ. Res. Lett.*, **16**, 084043, https://doi.org/10.1088/1748-9326/ac14ee.

——, ——, ——, and ——, 2023: Regularized fingerprinting in detection and attribution of climate change with weight matrix optimizing the efficiency in scaling factor estimation. *Ann. Appl. Stat.*, **17**, 225–239, https://doi.org/10.1214/22-AOAS1624.

Liu, Y., and J. Xie, 2020: Cauchy combination test: A powerful test with analytic *p*-value calculation under arbitrary dependency structures. *J. Amer. Stat. Assoc.*, **115**, 393–402, https://doi.org/10.1080/01621459.2018.1554485.

Moran, P. A., 1950: Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23, https://doi.org/10.2307/2332142.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

Paradis, E., J. Claude, and K. Strimmer, 2004: APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290, https://doi.org/10.1093/bioinformatics/btg412.

Pešta, M., 2013: Total least squares and bootstrapping with applications in calibration. *Statistics*, **47**, 966–991, https://doi.org/10.1080/02331888.2012.658806.

Ribes, A., S. Planton, and L. Terray, 2013: Application of regularised optimal fingerprinting to attribution. Part I: Method, properties and idealised analysis. *Climate Dyn.*, **41**, 2817–2836, https://doi.org/10.1007/s00382-013-1735-7.

Swart, N. C., J. C. Fyfe, E. Hawkins, J. E. Kay, and A. Jahn, 2015: Influence of internal variability on Arctic sea-ice trends. *Nat. Climate Change*, **5**, 86–89, https://doi.org/10.1038/nclimate2483.

——, and Coauthors, 2019: The Canadian Earth System Model version 5 (CanESM5.0.3). *Geosci. Model Dev.*, **12**, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019.

Wang, Z., Y. Jiang, H. Wan, J. Yan, and X. Zhang, 2021: Toward optimal fingerprinting in detection and attribution of changes in climate extremes. *J. Amer. Stat. Assoc.*, **116** (533), 1–13, https://doi.org/10.1080/01621459.2020.1730852.

Zhang, X., F. W. Zwiers, and P. A. Stott, 2006: Multimodel multisignal climate change detection at regional scale. *J. Climate*, **19**, 4294–4307, https://doi.org/10.1175/JCLI3851.1.