Uncertainty in optimal fingerprinting is underestimated

Yan Li, Kun Chen, Jun Yan and Xuebin Zhang

1 Department of Statistics, University of Connecticut, Storrs, CT 06269, United States of America
2 Climate Data Analysis Section, Environment and Climate Change Canada, Toronto, ON M3H 5T4, Canada

E-mail: jun.yan@uconn.edu

Keywords: bootstrap calibration, climate change, confidence interval, detection and attribution, natural climate variability, uncertainty quantification

Abstract

Detection and attribution analyses of climate change are crucial in determining whether the observed changes in a climate variable are attributable to human influence. A commonly used method for these analyses is optimal fingerprinting, which regresses observed climate variables on the signals, climate model simulated responses under external forcings. The method scales the simulated response under each external forcing by a scaling factor to best match the observations. The method relies critically on the confidence intervals for the scaling factors. The coverage rate, the relative frequency a confidence interval containing the unknown true value of the corresponding scaling factor, in the prevailing practice has been noted to be lower than desired. The mechanism of this under-coverage and its impacts on detection and attribution analyses, however, have not been investigated. Here we show that the under-coverage is due to insufficient consideration of the uncertainty in estimating the natural variability when fitting the regression and making statistical inferences. The implication is that the ranges of uncertainties in important quantities such as attributable anthropogenic warming and climate sensitivity based on the optimal fingerprinting technique should be wider than what has been believed, especially when the signals are weak. As a remedy, we propose a calibration method to correct this bias in the coverage rate of the confidence levels. Its effectiveness is demonstrated in a simulation study with known ground truth. The use of a large sample of climate model simulations to estimate the natural variability helps to reduce the uncertainty of the scaling factor estimates, and the calibrated confidence intervals provide more valid uncertainty quantification than the uncalibrated. An application to detection and attribution of changes in mean temperature at the global, continental, and subcontinental scale demonstrates that weaker detection and attribution conclusions are obtained with calibrated confidence intervals.

1. Introduction

Detection and attribution of climate change refers to the process of quantifying the evidence for a causal link between external drivers of climate change and observed changes in climatic variables with certain statistical confidence. The Intergovernmental Panel on Climate Change stated in its 5th assessment that “it is extremely likely that more than half of the observed increase in global average surface temperature from 1951 to 2010 was caused by the anthropogenic increase in greenhouse gas concentration and other anthropogenic forcings together” [3]. Statistically, it means that there is more than 95% probability that the attributable anthropogenic warming is greater than half of the observed warming [3, p 876]. While this assessment is based on multiple lines of evidence, optimal fingerprinting as a major method for detection and attribution of climate change [1–3, 14, 16, 18, 30] plays a critical role in estimating the attributable anthropogenic warming and constructing its uncertainty range. Results from optimal fingerprinting have also been used to produce observationally constrained future climate projection [6, 20] and to estimate important properties of the climate system such as climate sensitivity [7, 23, 31].
Optimal fingerprinting is based on a linear regression with errors-in-variables [1] where the outcome is the observed climate variable of interest, and the predictors are the expected responses, also known as signals or fingerprints, of the climate variable to certain external forcings. The regression coefficients, also called scaling factors, scale the fingerprints which are typically estimated from climate model simulations to best match the changes in the observations. The scaling factors are the central target of the statistical inference. If the confidence interval of a scaling factor with a desired confidence level (say 90\%) is above 0 (significantly positive), then the scaling factor is significantly positive and the corresponding external forcing’s fingerprint is claimed to be ‘detected’ in the observed data at that confidence level. If the confidence interval additionally covers 1, then the corresponding fingerprint is consistent with observed changes, which is a necessary statistical evidence for attributing the observed changes to the corresponding forcing. Sufficient conditions for attribution further require careful physical reasoning that eliminates other factors [26].

The proportion of the times that a confidence interval contains the true value of the scaling factor is an important statistical property, which is called coverage probability or coverage rate. By design, a confidence interval should contain the unknown true scaling factor with a coverage rate that equals the desired confidence level, provided that all the assumptions used in deriving the confidence interval are met. If any of the assumptions is not met, however, the actual coverage rate may not be the same as the nominal confidence level. If the coverage rate is lower than the nominal confidence level, under-coverage occurs and the confidence interval is too narrow. In the context of optimal fingerprinting analysis, under-coverage can lead to false detection claims and too small uncertainty ranges for changes (e.g. warming) attributable to an external forcing. While optimal fingerprinting has been used in detection and attribution analyses for over 20 years, its coverage property is not well understood. It was not until recently that the under-coverage of the confidence intervals was explicitly reported [5].

The under-coverage of confidence intervals in optimal fingerprinting is caused by insufficient consideration of the uncertainty in the estimate of the covariance matrix \( \Sigma \) of the regression error; see model (1) in section 2.1 for details. The inference about the scaling factors involves the inverse of \( \Sigma \), or the optimal weight in statistical terms, and \( \Sigma \) represents the natural variability of the climate variable. While \( \Sigma \) needs to be estimated from climate model simulations and as such subject to sampling uncertainty, the construction of confidence intervals assumes that \( \Sigma \) can be safely substituted by an estimate. Such estimate is based on climate model simulations. Due to the limited amount of model simulations relative to the dimension of \( \Sigma \), the safe substitution assumption is often not valid, which leads to underestimation of the uncertainty in the scaling factor estimator, especially when the signal-to-noise ratio is low. To circumvent this problem, the prevailing practice is to obtain two independent estimators for \( \Sigma \) by partitioning the available simulations into two independent samples, one used in estimating the scaling factors and the other used in constructing their confidence intervals [1, 30]. It is hoped that such two-sample approach would produce proper confidence interval but its effectiveness has not be studied.

Reports on the coverage rates of the confidence intervals in the optimal fingerprinting setting are rare, casual, and under unrealistic assumptions, which have not attracted much attention. In a study that considers structural error of climate models but still assumes \( \Sigma \) known [13], the confidence intervals were found to have too low coverage rates. In a simplified case that is perhaps only of theoretical interest, it is assumed that the structure of \( \Sigma \) is partially known in the form of \( \Sigma = \sigma^2 I \), where \( \sigma^2 \) is an unknown scalar and \( I \) is the identity matrix [5]. The two-sample method in this case produces a coverage rate that is close to the nominal level when the signal-to-noise ratio is large, but is too low when the signal-to-noise ratio is small or when the fingerprints are highly correlated.

The under-coverage of the confidence intervals in a typical fingerprinting analysis where the whole structure of \( \Sigma \) is unknown has not been studied. While it was noted in passing in [12], we are not aware of any published study that cites this work in the context of under-coverage. As will be shown in our numerical study, in this case, the confidence intervals produced by the two-sample approach do not give enough coverage. This immediately affects qualitative detection statements — false claims of detection of anthropogenic fingerprint may have been made due to overly narrow confidence intervals that would otherwise have covered 0. More importantly, the uncertainty range of many estimated quantities based on such confidence intervals, such as warming attributable to greenhouse gases and transient climate sensitivity, would be overly narrow too.

The contributions of this paper are two-fold. First, we demonstrate through a simulated experiment study that there is an under-coverage issue in optimal fingerprinting and that the level of under-coverage varies with the signal strength, sample size for estimating the natural variability \( \Sigma \), and the structure of the true \( \Sigma \). Second, we propose an approach to calibrate the confidence intervals and show that it can resolve or substantially reduce the under-coverage issue. We apply our method to annual mean near-surface air temperature at the global, continental, and subcontinental scales to illustrate the impact of
the under-coverage issue in a typical optimal fingerprinting analysis. The computer codes of our method are publicly available in an open-source R package dace \cite{24}.

2. Methods

2.1. Optimal fingerprinting

The optimal fingerprinting method takes the form of a linear regression with errors-in-variables \cite{1}:

\[
\begin{align*}
Y &= \sum_{i=1}^{p} X_i \beta_i + \epsilon \\
\bar{X}_i &= X_i + \nu_i & i = 1, \ldots, p,
\end{align*}
\]

where \(Y\) is an \(n\)-dimensional outcome vector of observed climate variable; \(p\) is the number of external forcings; \(\bar{X}_i\) is the \(n\)-dimensional vector of unobserved, expected climate response to the \(i\)th external forcing; \(X_i\) is an estimate of \(\bar{X}_i\) with an \(n\)-dimensional random error \(\nu_i\); \(\beta_i\) is the unknown scaling factor of the signal of the \(i\)th external forcing; and \(\epsilon\) is a \(n\)-dimensional vector of regression error. No intercept is present because the observation and the estimated signals are both centered to their corresponding climates of a common base period. The additivity of the multiple forcing signals is typically assumed \cite[e.g. 1, 14, 15, 17, 30]{}. The assumption has been investigated for commonly used forcings in simulated climate systems \cite{9, 25, 32, 33}, and the results support its use for temperature. The errors \(\epsilon\) and \(\nu_1, \ldots, \nu_p\) are independently, normally distributed with zero mean and covariance matrices \(\Sigma, \Omega_1, \ldots, \Omega_p\), respectively. Assuming that the natural variability simulated by the climate models is the same as in the observations, if \(\bar{X}_i\) is the average of \(k_i\) runs from a single climate model under the \(i\)th forcing, then we have \(\Omega_i = \Sigma / k_i\). In practice, the runs may come from different climate models. When the numbers of independent runs under the \(i\)th forcing from different models are the same, \(k_i\) is simply the total number of runs. When the number of simulations by different models are different and when the multi-model ensemble mean is used as simulated responses, by first averaging individual model ensemble means and then averaging across the models, \(k_i\) can be adjusted to reflect the difference in the number of runs by each model because \(\Omega_i\) is proportional to \(\Sigma\).

The inference about \(\beta\) depends crucially on the natural variability represented by \(\Sigma\). In the prevailing generalized total least squares estimation, \(\Sigma\) is needed because the optimal weight matrix \(\Sigma^{-1}\) is used in the estimation of \(\beta\) \cite{35, 36}. As the true \(\Sigma\) is unknown, it is estimated from \(m\) independent replicates of the \(n\)-dimensional error vector \(\epsilon\), denoted by

\[C_m = \{\epsilon^{(1)}, \ldots, \epsilon^{(m)}\}.
\]

For ease of referencing, we call \(C_m\) replicates of ‘climate noise’, which represents natural variability. In practice, they are based on pre-industrial control simulations or intra-ensemble differences of forced simulations. Reliable estimate of the \(n \times n\) matrix \(\Sigma\) requires the sample size \(m\) to be much greater than \(n\). The reality is, however, that \(m\) is often much lower than \(n\), in which case, the sample covariance is not even of full rank and thus not invertible. In regularized optimal fingerprinting \cite{29, 30}, a linear shrinkage estimator \(\hat{\Sigma}_m\) \cite{21} of \(\Sigma\) is used, which ensures its invertibility. Constructing the confidence intervals for \(\beta\) also needs \(\Sigma\). The two-sample method partitions \(C_m\) into two halves and gets two independent estimators of \(\Sigma\): one is used to obtain point estimate of \(\beta\) while the other is used to construct confidence intervals \cite{1, 30}. Other methods investigated by \cite{5} are based on the asymptotically normal distribution of the estimator of \(\beta\) \cite{11} or bootstrap \cite{28}. These methods use the whole \(C_m\) to obtain one estimate \(\Sigma\) in both point estimation and confidence interval construction. As to be shown, confidence intervals from all three methods suffer the under-coverage issue.

To fix the under-coverage issue of the confidence intervals, the estimated standard errors needs to be enlarged. We propose a parametric bootstrap procedure to calibrate the standard errors from the normal approximation. The bootstrap procedure is parametric because bootstrap samples of \(C_m\), \(Y\), and \(\bar{X}_i, i = 1, \ldots, p\), are all generated from their fitted multivariate normal distributions with \(\Sigma\) substituted by \(\hat{\Sigma}_m\) \cite{4}. The technical details of the parametric bootstrap calibration procedure is presented in appendix.

2.2. Simulation experiment design

To demonstrate the under-coverage issue and the effectiveness of the calibrated confidence intervals, we designed a simulation experiment that mimicked real-world application but with known scaling factors. Here the detection and attribution analysis is based on the regularized optimal fingerprinting method \cite{29, 30}. The application is optimal fingerprint analysis of mean temperature at the global scale over time period 1951–2010 with a goal to detect signals from anthropogenic and natural forcings. Five-year mean temperature anomalies from observations and model simulations are averaged over 54 grid boxes of size \(40^\circ \times 30^\circ\). This resolution is commonly used for global-scale studies \cite[e.g. 37, 38]{}. The coarse resolution makes the dimension of \(\Sigma\) smaller for better estimation; it also helps reducing the noise in the observations, thereby improving the signal-to-noise ratio. Grid boxes without sufficient observational data (details in the next subsection) were removed from the analysis, resulting in \(n = 572\) space-time observational data points.

In each replicate, the following data need to be generated: (a) ‘climate noise’ \(C_m\) to represent \(m\) simulations by climate model for estimating \(\Sigma\); (b)
observations $Y_i$; (c) estimates of two signals $\hat{X}_i, \tilde{X}_i$. The data generation requires the knowledge of true $\Sigma$ and true signal $X_1$ and $X_2$ which are specified as follows. The true $\Sigma$ was set to be the linear shrinkage estimate [21] of covariance matrix, using simulations conducted with models participating in the Coupled Model Intercomparison Project Phase 5 (CMIP5) [34]. In total, 223 60-year chunks of data from the control simulations were used in this estimation; see table 2 in supplementary material (available online at stacks.iop.org/ERL/16/084043/mmedia). The true signals $X_1$ and $X_2$ were set to be the estimated fingerprints from $k_1 = 35$ and $k_2 = 46$ CMIP5 simulations under the anthropogenic (ANT) and natural (NAT) forcings, respectively. For estimating $\Sigma$, $m$ 60-year chunks of ‘climate noise’ $C_m$ were first generated from a multivariate normal distribution $N(0, \Sigma)$, with four sizes of $m \in \{50, 100, 200, 400\}$. Recall that $C_m$ contains $m$ independent copies of ‘climate noise’ $\varepsilon$, where $m = 50$ is relatively easy obtainable while $m = 400$ is close to the upper limit of what are currently available. The signal estimates $\hat{X}_i$, where $i \in \{1, 2\}$, were generated from multivariate normal distribution $N(\lambda X_i, \Sigma/k_i)$, with $(k_1, k_2) = (35, 46)$ and $\lambda \in \{1, 0.5\}$. Here the two levels of $\lambda$ were meant to consider two levels of signal strength, one is the same as model simulated response and the other signal strength is halved. Finally, the ‘observations’ vector $Y$ was generated from regression model (1) with $\varepsilon$ generated from $N(0, \Sigma)$ and $\beta_1 = \beta_2 = 1$.

For each combination of $m$ and $\lambda$, 1000 replicates were generated. For each replicate, a detection and attribution analysis was conducted with the regularized optimal fingerprinting method. Three existing methods, the commonly used two-sample (TS) approach [1, 30], asymptotic normal (N) approximation [5, 11] and nonparametric bootstrap (B) [5, 28], and the parametric bootstrap calibration (PBC), were used to construct 90% confidence intervals for the scaling factors. The bootstrap sample size was 1000 in the bootstrap methods.

2.3. Data preparation for application to mean temperature
To demonstrate the potential impact of under coverage in real-world applications, we conducted regularized optimal fingerprinting analyses on the annual mean near-surface air temperature at the global (GL), continental, and subcontinental scales during 1951–2010. The continental scale includes Northern Hemisphere (NH), NH midlatitude between 30° and 70° (NWH), Eurasia (EA), and North America (NA) [37]. The subcontinent scale includes Western North America (WNA), Central North America (CNA), Eastern North America (ENA), Southern Canada (SCA), and Southern Europe (SEU) [10]. Two external forcings were considered, ANT and NAT. See supplementary information for details of the data preparation based on HadCRUT4 dataset [27] and the CMIP5 simulations [34].

3. Results
3.1. Simulated experiments
Figure 1 shows the estimated coverage rates and lengths of the 90% confidence intervals for $\beta_1$ and $\beta_2$ summarized from the 1000 replicates. It is clear that the coverage rates of the confidence intervals from all existing methods, including the TS method, are lower than the nominal level 90% in all the cases and they can be below 60% when $m = 50$, which is consistent with the results shown in the work by Hannart [12]. As $m$ increases, the coverage rate improves but does not reach 90%; they are still below 80% even when $m = 400$; see figure 1(a). As expected, the intervals are longer for the weaker signals; they are longer for the NAT signal than for the ANT signal, and for weaker signal strength ($\lambda = 0.5$) than for stronger signal strength ($\lambda = 1$); see figure 1(b). The lengths of the confidence intervals are similar for different $m$ values, in contrast to the improved coverage rate with larger $m$. This does not mean that the intervals lengths are stable relative to $m$; rather, it indicates more severe underestimation of the uncertainty in scaling factor estimates with smaller $m$.

Next we compare the three existing methods. For both forcings and both signal strength levels, the coverage rates of the B intervals and the N intervals are similar except when $m$ is small, in which case the B intervals have slightly higher coverage rates than the N intervals. To some extent, this resembles the recent finding in the setting where $\Sigma$ is known up to a scale [5]. The results from the TS method are different. The sample size $m$ of ‘climate noise’ for estimating $\Sigma$ in the TS method is a half of those in the N and B methods. This affects both the coverage rate and the length of confidence intervals, depending on the strength of the signal. For the ANT forcing, when $m = 50$ or 100, the coverage rates from the TS method are higher and the confidence intervals are wider compared with those than those from the N and B methods. This shows the utility of the two-step method in correcting underestimation of confidence interval as intended, but the correction is not sufficient as the coverage rates are still too small. For the NAT forcing, however, the coverage rates from the TS method are even lower than those from the N and B methods while the interval lengths are similar, this indicates that the two-step approach does not improve the estimation as intended because the sample size is simply too small to provide sufficiently accurate estimate of natural variability for the detection of weak signal. These results, obtained under a completely unknown structure of $\Sigma$, are in contrast to the comparison of the TS method versus the N and B methods when $\Sigma = \sigma^2 I$ [5].

The coverage rates and lengths of the PBC intervals are also shown in figure 1. The coverage rates
Figure 1. Estimated coverage percentages (a) and lengths of 90% confidence intervals (b) constructed from TS (two-sample), N (normal approximation), B (nonparametric bootstrap), and PBC (parametric bootstrap calibration) for the ‘ANT’ and ‘NAT’ forcings based on 1000 replicates. The error bars are the 95% confidence intervals of the coverage rates. The number of ensembles for estimating the ‘ANT’ and ‘NAT’ signals are \( k_1 = 35 \) and \( k_2 = 46 \), respectively.

are much closer to the 90% nominal level than those from the existing methods. Even with \( m = 50 \), their coverage rates are above 80%. The coverage rates are close to 90% for most cases when \( m \geq 200 \) (about \( 1/3 \) of \( n = 572 \)). Seven out of 8 confidence intervals of the coverage rates cover 0.9; see figure 1(a). The PBC intervals are much longer than the existing ones. When \( \lambda = 0.5 \), the PBC intervals are 27%–53% longer for the ANT factor and 35%–70% longer for the NAT factor than those from the TS method for the four \( m \) levels; see figure 1(b). When \( \lambda = 1 \), the percentages remain about the same. These are clear indications that the severely underestimated uncertainty range of the scaling factors in the existing could be corrected to a large extent. This should have affected the estimated uncertain range of important quantities such as attributable warming or climate sensitivity that uses the results of optimal fingerprinting analyses. The PBC interval lengths decrease as \( m \) increases for both the ANT and NAT scaling factors, indicating the benefit of better \( \Sigma \) estimates.

As shown in figure 2, the widened confidence intervals lead to changes in qualitative statements about detection and attribution. A detection statement is made when the confidence interval is above zero; an attribution statement is made if the confidence interval is above zero and covers one. Obviously, it is more difficult to make a detection statement because the confidence intervals are wider (and less likely to exclude 0). The effect on attribution statement is more complex. In case a detection statement can still be made, a wider confidence interval means a higher chance for it to include 1, and thus, higher chance for making an attribution claim. When the signals are strong, the existing intervals are short enough so that the widened, PBC intervals do not change the detection rate. The decrease in detection rate using the PBC intervals is seen only in the weak-signal case of \( \lambda = 0.5 \) with the NAT scaling factor; see figure 2(a). As for attribution, the attribution rates increase except in the case of \( \lambda = 0.5 \) with the NAT scaling factor, where the rates drop slightly; see figure 2(b). These results indicate that the underestimation of the uncertainty in scaling factors does not affect the qualitative conclusion about detection and attribution for cases with strong signal-to-noise ratio such as warming in global mean temperature, but it does in cases of weak signal-to-noise ratio.

3.2. Detection and attribution of changes in mean temperature

The four methods of confidence interval construction were applied to the detection and attribution analysis of changes in mean temperature. Figure 3 shows the estimated scaling factors for the ANT and NAT forcings along with their corresponding 90% confidence intervals from the TS, N, B, and PBC approaches. The PBC intervals are much wider than those from the TS, N, and B approaches for both ANT and NAT forcings (figure 3, both panels). The TS intervals are centered differently than the others because the estimate of \( \Sigma \) uses only a half of the ‘climate noise’ replicates in \( C_m \). We focus on the comparison between the results based on PBC and those based on the prevailing TS
Figure 2. Estimated percentages of times a detection (a) and an attribution (b) can be claimed based on the 90% confidence intervals constructed from TS (two-sample), N (normal approximation), B (nonparametric bootstrap), and PBC (parametric bootstrap calibration) for the scaling factors of the 'ANT' and 'NAT' forcings, obtained from 1000 replicates. The error bars are the 95% confidence intervals for the rates. The number of ensembles for estimating the 'ANT' and 'NAT' signals are $k_1 = 35$ and $k_2 = 46$, respectively.

method. Over all the analyses considered, the percentage increase in the lengths of the intervals from TS to PBC ranges from 27% to 113% for the ANT scaling factor (figure 3, upper panel), and 19%–96% for the NAT scaling factor (figure 3, lower panel). Because the NAT signal is much weaker than the ANT signal, all the intervals are much wider for the scaling factor of the former than for that of the latter. The qualitative detection and attribution statements based on their covering zero or one are more likely to change in the analyses of the NAT signal.

For the global mean temperature, the PBC intervals are 70% and 58% wider than the TS intervals for the ANT and the NAT scaling factors, respectively. The qualitative ‘detected’ statements are the same based on the two confidence intervals for both forcings, both intervals being above zero. The qualitative attribution statements are claimed for the ANT forcing by both methods, but only the PBC method leads to the ‘attributable’ statement for the NAT forcing with a confidence interval above zero and covering one. The width of the confidence intervals of the ANT scaling factor translates directly to the range of attributable anthropogenic warming [19], observation-constrained future climate projection [6, 20], and observation-constrained climate sensitivity estimates [7, 23, 31]. It follows that the existing estimates of the uncertainty range of these important quantities for the global mean temperature also need to be widened.

In the four continental-scale analyses, the PBC intervals are 67%–113% and 27%–96% wider than the TS intervals for the ANT and the NAT scaling factors, respectively. For the ANT forcing, both methods lead to the ‘detected’ statement in all four regions with confidence intervals above zero; the PBC method supports the ‘attributable’ statement with four intervals covering one; the TS method supports the ‘attributable’ statement only for NH and NHM but not for EA and NA (confidence intervals not covering one). For the NAT forcing, both methods give confidence intervals above zero and covering one in NH, NHM, and NA; in the EA analysis, the TS interval is not above zero, which means no ‘detected’ statement, while the PBC interval is above zero, although barely, and covers one, which leads to both ‘detected’ and ‘attributable’. Although most of the qualitative statements based on the PBC intervals about the ANT signal are similar to what has been reported with data ending at 2005 [37], the longer PBC confidence intervals imply higher uncertainty in regional estimates of transient climate response to increases in greenhouse gases [22].

At the subcontinental scale, the PBC intervals are 27%–63% and 19%–43% wider than the TS intervals in the five regions for the ANT and the NAT scaling factors, respectively. For the ANT forcing, the TS method makes the ‘detected’ statement for all five regions, but the statement for SCA was reverted by the PBC interval which covers zero; both methods support the ‘attributable’ statement for WNA, CNA, and SEU, but only the PBC method supports the ‘attributable’ statement for ENA with a confidence interval
covering one. For the NAT forcing, the two methods lead to the same ‘not detected’ statement for WNA, CNA, ENA, and SEU with intervals covering zero; for the SCA analysis, the TS interval covers zero but the PBS interval is above zero and covers one, supporting both ‘detected’ and ‘attributable’ statements. Compared to the global and continental analyses, the subcontinental analyses have obviously longer confidence intervals from all methods. Between the TS and the PBS intervals, the drastically reduced lower end of the PBS confidence intervals weaken the evidence of attributable anthropogenic changes with much increased uncertainty.

4. Conclusion

The uncertainty in estimating $\Sigma$ has not been properly accounted for in detection and attribution analyses based on optimal fingerprinting. This practice leads to overly narrow confidence intervals with coverage rates below nominal levels for the scaling factors. Consequently, the ranges of uncertainty for quantities that are estimated based on the scaling factors from optimal fingerprinting, such as attributable anthropogenic warming, observation-constrained projection, and observation-constrained climate sensitivity, are too narrow as they are proportional to the scaling factor of the ANT signal [8, 31]. The prevailing two-sample approach [1] does not solve the problem as it was hoped for. The extent of the underestimation can be substantial, depending on the signal-to-noise ratio and details of the data analyses including the sample size of climate model simulations used in estimating $\Sigma$ (figure 1).

For practical purposes in detection and attribution analyses of climate change with optimal fingerprinting, a large sample size $m$ of climate model simulations to estimate $\Sigma$ helps to reduce the uncertainty. The PBC method improves the validity of the confidence intervals notably, with coverage rates close to the nominal level for $m = 200$ or higher. While it can be difficult, practical applications of optimal fingerprint method should aim at obtaining a large sample to estimate the natural variability such that
the coverage rate is close to its nominal level. For \( m = 100 \) or lower, the PBC intervals have coverage rates much better than the TS intervals, but still 5% or 10% lower than the nominal level. In our experiment, there appears to be no perfect solution that works in all situations, especially for smaller \( m \). Therefore, interpretation of detection and attribution analyses based on the optimal fingerprinting method should take this caveat that the uncertainty is underestimated into consideration.

**Data availability statement**

The data and code used to reproduce the simulation study and data analysis are available in a figshare repository for review and will be made available in a public GitHub repository. The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.14981241.v3.

**Funding**

J Y’s research was partially supported by an NSF Grant DMS 1521730.

**Author contributions**

K C and J Y supervised the methodological development. Y L implemented the methods, carried out the simulation study, and conducted the data analysis. X Z provided insight on detection and attribution of climate change. All authors designed the analyses and contributed to the manuscript preparation.

**Conflict of interest**

The authors declare that they have no conflict of interest.

**Appendix. Details of confidence interval calibration**

The essence of our calibration is to estimate a multiplicative scale of the standard errors from the normal approximation so that, after the scaling, the confidence intervals gives desired coverage rates. Assuming multivariate normality of the error terms in (1), we do this through a parametric bootstrap procedure where each parametric bootstrap version of the data \( \{C_m, Y, \tilde{X}_1, \ldots, \tilde{X}_p\} \) is generated from the fitted model.

The parametric bootstrap calibration runs as follows for a large number \( B \). For each \( b \in \{1, \ldots, B\} \):

(a) Generate a parametric bootstrap sample \( C_m^{(b)} \) of \( C_m \) from the fitted normal distribution with covariance matrix \( \Sigma_m \) and obtain the linear shrinkage estimate \( \Sigma_m^{(b)} \) based on this sample.

(b) Generate a parametric bootstrap sample of \( \gamma^{(b)} \) and \( \tilde{X}_i^{(b)} \), \( i = 1, \ldots, p \) from model (1) with \( X \) and \( \beta \) replaced with their generalized total least squares estimate \( \tilde{X} \) and \( \tilde{\beta} \), respectively, and \( \Sigma \) replaced with \( \tilde{\Sigma}_m \).

(c) For bootstrap outcome \( \gamma^{(b)} \) and predictors \( \tilde{X}_i^{(b)} \), \( i = 1, \ldots, p \), use weight matrix (\( \Sigma_m^{(b)} \))\(^{-1} \) to obtain the generalized total least squares estimator of \( \beta \), \( \tilde{\beta}^{(b)} \), and the corresponding standard error based on normal approximation method [11], denoted by \( \text{SD}(\tilde{\beta}^{(b)}) \).

Next, let \( \text{SD}_i \) be the empirical standard deviation of \( \tilde{\beta}_i^{(b)} \), \( b = 1, \ldots, B \), and \( \text{SD}_i \) be the average of the standard error \( \text{SD}(\tilde{\beta}_i^{(b)}) \), \( b = 1, \ldots, B \). The desired enlargement scale \( k_i \) to correct the coverage for \( \hat{\beta}_i \) is then estimated as \( k_i = \text{SD}_i/\text{SD}_i \), \( i = 1, \ldots, p \). Suppose that the standard errors of the regression coefficient estimates based on normal approximation or bootstrap from the observed data are \( \text{SD}(\tilde{\beta}_i) \), \( i = 1, \ldots, p \). The calibrated confidence with level \( (1 - \alpha) \) for \( \hat{\beta}_i \) is constructed as

\[
(\tilde{\beta}_i - k_i z_{\alpha/2} \text{SD}(\tilde{\beta}_i), \tilde{\beta}_i + k_i z_{\alpha/2} \text{SD}(\tilde{\beta}_i)),
\]

where \( z_{\alpha/2} \) is the \( 1 - \alpha/2 \) upper quantile of the standard normal distribution.

Our implementation of the method is publicly available in an open source R package dacc [24].

**ORCID iDs**

Yan Li @ https://orcid.org/0000-0003-2182-9048
Kun Chen @ https://orcid.org/0000-0003-3579-5467
Jun Yan @ https://orcid.org/0000-0003-4401-7296

**References**

[1] Allen M R and Stott P A 2003 Estimating signal amplitudes in optimal fingerprinting, part I: theory Clim. Dyn. 21 477–91
[2] Allen M R and Tett S F B 1999 Checking for model consistency in optimal fingerprinting Clim. Dyn. 15 419–34
[3] Bindoff N L et al 2013 Detection and attribution of climate change: from global to regional Climate Change 2013—The Physical Science Basis: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, ed T F Stocker et al (Cambridge: Cambridge University Press) pp 867–952
[4] Davison A C and Hinkley D V 1997 Bootstrap Methods and Their Application (Cambridge: Cambridge University Press)
[5] DelSole T, Trenary L, Yan X and Tippett M K 2019 Confidence intervals in optimal fingerprinting Clim. Dyn. 52 4111–26
[6] Forest C E, Allen M R, Sokolov A P and Stone P H 2001 Constraining climate model properties using optimal fingerprint detection methods Clim. Dyn. 18 277–95
[7] Forest C E, Stone P H and Sokolov A P 2006 Estimated PDFs of climate system properties including natural and anthropogenic forcings Geophys. Res. Lett. 33 L01705
[8] Frame D J, Stone D A, Stott P A and Allen M R 2006 Alternatives to stabilization scenarios Geophys. Res. Lett. 33 L14707
[9] Gillett N P, Wehner M F, Tett S F B and Weaver A J 2004 Testing the linearity of the response to combined greenhouse gas and sulfate aerosol forcing Geophys. Res. Lett. 31 L14201
[10] Giorgi F and Francisco R 2000 Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM Clim. Dyn. 16 169–82
[11] Gleser L J 1981 Estimation in a multivariate 'errors in variables' regression model: Large sample results Ann. Stat. 9 24–44
[12] Hannart A 2016 Integrated optimal fingerprinting: method description and illustration J. Clim. 29 1977–98
[13] Hannart A, Ribes A and Naveau P 2014 Optimal fingerprinting under multiple sources of uncertainty Geophys. Res. Lett. 41 1261–8
[14] Hasselmann K 1997 Multi-pattern fingerprint method for detection and attribution of climate change Clim. Dyn. 13 601–11
[15] Hasselmann K F 1979 On the signal-to-noise problem in atmospheric response studies Joint Conf. Royal Meteorological Society, American Meteorological Society, Deutsche Meteorologische Gesellschaft and the Royal Society (Royal Meteorological Society) pp 251–9
[16] Hegerl G and Zwiers F 2011 Use of models in detection and attribution of climate change Wiley Interdiscip. Rev.: Clim. Change 2 570–91
[17] Hegerl G C, Hasselmann K, Cubasch U, Mitchell J F B, Boeckner E, Voss R and Waszkewitz J 1997 Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change Clim. Dyn. 13 613–34
[18] Hegerl G C, von Storch H, Hasselmann K, Santer B D, Cubasch U and Jones P D 1996 Detecting greenhouse-gas-induced climate change with an optimal fingerprint method J. Clim. 9 2281–306
[19] Jones G S, Stott P A and Christidis N 2013 Attribution of observed historical near-surface temperature variations to anthropogenic and natural causes using CMIP5 simulations J. Geophys. Res.: Atmos. 118 4001–24
[20] Jones G S, Stott P A and Mitchell J F B 2016 Uncertainties in the attribution of greenhouse gas warming and implications for climate prediction J. Geophys. Res.: Atmos. 121 6969–92
[21] Ledoit O and Wolf M 2004 A well-conditioned estimator for large-dimensional covariance matrices J. Multivariate Anal. 88 365–411
[22] Leduc M, Matthews H D and de Elía R 2016 Regional estimates of the transient climate response to cumulative CO2 emissions Nat. Change 6 474–8
[23] Lewis N 2013 An objective Bayesian improved approach for applying optimal fingerprint techniques to estimate climate sensitivity J. Clim. 26 7411–29
[24] Li Y, Chen K and Yan J 2020 DA(C)C: Detection and Attribution of Climate Change R Package Version 0.1–13
[25] Maravel K, Schmidt G A, Shindell D Bonfils C, LeGrande A N, Nazarenko L and Toagaridz K 2015 Do responses to different anthropogenic forcings add linearly in climate models? Environ. Res. Lett. 10 104010
[26] Mitchell J F B and Karoly D J 2001 Detection of climate change and attribution of causes Climate Change 2001—The Scientific Basis: Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, ed G C Hegerl, F W Zwiers, M R Allen and J Marengo (Cambridge: Cambridge University Press) pp 695–738
[27] Morice C, P, Kennedy J J, Rayner N A and Jones P D 2012 Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set J. Geophys. Res.: Atmos. 117 D08101
[28] Pesta M 2013 Total least squares and bootstrapping with applications in calibration Statistics 47 966–91
[29] Ribes A, Azais J-M and Planton S 2009 Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate Clim. Dyn. 33 707–22
[30] Ribes A, Planton S and Terray L 2013 Application of regularised optimal fingerprinting to attribution. Part I: method, properties and idealised analysis Clim. Dyn. 41 2817–36
[31] Schurer A, Hegerl G, Ribes A, Polson D, Morice C and Tett S 2018 Estimating the transient climate response from observed warming J. Clim. 31 8645–63
[32] Schurer A P, Tett S F B and Hegerl G C 2014 Small influence of solar variability on climate over the past millennium Nat. Geosci. 7 104–8
[33] Shiogama H, Stone D A, Nagashima T, Nozawa T and Emori S 2013 On the linear additivity of climate forcing-response relationships at global and continental scales Int. J. Climatol. 33 2542–50
[34] Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design Bull. Am. Meteorol. Soc. 93 485–98
[35] van Hufiolf S 1991 The generalized total least squares problem: formulation, algorithm and properties Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, ed G H Golub and P Van Dooren (Berlin: Springer) pp 651–60
[36] van Hufiolf S and Vandewalle J 1989 Analysis and properties of the generalized total least squares problem AX ≈ B when some or all columns in A are subject to error SIAM J. Matrix Anal. Appl. 10 294–315
[37] Zhang X, Zwiers F and Stott P A 2006 Multimodel multisignal climate change detection at regional scale J. Clim. 19 4294–307
[38] Zwiers F W and Zhang X 2003 Toward regional-scale climate change detection J. Clim. 16 793–7