Clim Dyn (2013) 41:2837–2853
DOI 10.1007/s00382-013-1736-6

Application of regularised optimal fingerprinting to attribution. Part II: application to global near-surface temperature

Aurélien Ribes · Laurent Terray

Received: 27 July 2012 / Accepted: 10 March 2013 / Published online: 4 April 2013

Abstract Attribution of global near-surface temperature changes is revisited using simulations from the coupled model intercomparison project 5 and methodological improvements from the regularised optimal fingerprinting approach. The analysis of global mean temperature shows that changes can be robustly detected and attributed to anthropogenic influence. However, the differences between results from individual models and observations are found to be larger than the simulated internal variability in several cases. Discrimination between greenhouse gases and other anthropogenic forcings, based on the global mean only, is more difficult due to collinearity of temporal response patterns. Using spatio-temporal data provides less robust conclusions with respect to detection and attribution, as the results tend to deteriorate as the spatial resolution increases. More importantly, some inconsistencies between individual models and observations are found in this case. Such behaviour is not observed in a perfect model framework, where pseudo-observations and the expected response patterns are provided by the same model. However, using response patterns from a model other than the one used for pseudo-observations may lead to the same behaviour as real observations. Our results suggest that additional sources of uncertainty, such as modeling uncertainty or observational uncertainty, should not be neglected in detection and attribution.

Keywords Detection · Attribution · Climate change · Optimal fingerprints · Global temperature

1 Introduction

The assessment of the anthropogenic contribution to twentieth century near surface air temperature (SAT) changes is one of the key issues in climate science. This question has been discussed intensively in past Intergovernmental panel on climate change (IPCC) assessment reports, particularly within the framework of detection and attribution studies (D&A). These analyses seek to quantify the individual contributions of greenhouse gases, other anthropogenic and natural factors to past climate change. Many formal detection and attribution studies have used the optimal fingerprinting (OF) approach pioneered by Hasselmann (1979, 1993, 1997). This approach can be framed as a statistical multi-linear regression model in which the observations of a given climate variable are regressed against model-based estimates of response patterns to different external (anthropogenic and natural) forcings (Hegerl et al. 1996; Allen and Tett 1999; Allen and Stott 2003). It also requires two independent estimates of internal variability, which are usually provided by pre-industrial climate model simulations with constant external forcings and intra-ensemble variability derived from the residuals of historical simulations (using one or several observed forcings) after subtracting the ensemble mean.

Most recent global SAT detection and attribution studies have used a spatio-temporal analysis over the 1901–2000

Electronic supplementary material The online version of this article (doi:10.1007/s00382-013-1736-6) contains supplementary material, which is available to authorized users.

A. Ribes (✉)
CNRM-GAME, Météo France, CNRS, 42 avenue G. Coriolis, Toulouse 31057, France
e-mail: aurelien.ribes@cnrm.meteo.fr

L. Terray
SUC, CERFACS-CNRS URA1875, 42 avenue G. Coriolis, Toulouse 31057, France
e-mail: terray@cerfacs.fr

period (Stott et al. 2006; Huntingford et al. 2006). Response patterns to various external forcings were provided by historical simulations performed with four coupled general circulation models (CGCMs) constrained by observed estimates of anthropogenic and/or natural forcings. These models contributed to the coupled model intercomparison exercise version three (CMIP3) and were the only ones in CMIP3 perform attribution experiments (i.e. with only one source of varying external forcing, the others being held constant).

The main objective of this paper is to update the results of these previous multi-model studies using two new elements. First, we use an improved implementation of the optimal fingerprinting technique, as described in Part I (Ribes et al. 2013, R13 thereafter; see also Ribes et al. 2009). This variant, termed Regularised Optimal Fingerprinting (ROF), proposes a new, well-conditioned estimator of the internal variability covariance matrix. This estimator alleviates the frequently discussed truncation sensitivity issue related to EOF projection in the standard OF approach (Allen and Tett 1999; Allen et al. 2006). It also allows for improved discrimination of different forcing response patterns thanks to a better use of spatio-temporal information. A second new element is the availability of the new CMIP5 data sets, with many more CGCMs (10 currently, see Table 1) having performed attribution experiments than in previous intercomparison projects.

We ask the following questions: using the three elements detailed above, can we attribute some of the recent changes in global SAT to anthropogenic (ANT) influence? Can we separate the relative influence of greenhouse gases (GHG) from that of other anthropogenic (AER) and natural (NAT) forcings? At which spatial scales? Do the main assumptions behind the standard OF statistical method still hold when we use spatial information?

The remainder of the paper is outlined as follows. The observed and simulated datasets and methods are described in Sect. 2. Results are presented in Sect. 3, structured according to the sequence of questions listed above. A discussion focusing on a comparison with previous studies and remaining limitations as well as a summary are provided in Sect. 4.

Table 1 Ensembles of D&A simulations used

Climate model	Experiment	RIP	External forcings	Nb runs
CNRM-CM5[a]	HistoricalMisc	r*i1p1	ANT	10
	HistoricalNat	r*i1p1	NAT	6
	HistoricalGHG	r*i1p1	GHG	6
CanESM2	Historical	r*i1p1	ALL	5
	HistoricalNat	r*i1p1	NAT	5
	HistoricalGHG	r*i1p1	GHG	5
HadGEM2-ES	Historical	r*i1p1	ALL	4
	HistoricalNat	r*i1p1	NAT	4
	HistoricalGHG	r*i1p1	GHG	4
GISS-E2-R	HistoricalMisc	r*i1p109	ANT	5
	HistoricalNat	r*i1p1	NAT	5
	HistoricalGHG	r*i1p1	GHG	5
GISS-E2-H	HistoricalMisc	r*i1p109	ANT	5
	HistoricalNat	r*i1p1	NAT	5
	HistoricalGHG	r*i1p1	GHG	5
CSIRO-Mk3-6-0	Historical	r*i1p1	ALL	10
	HistoricalAnt	r*i1p1	ANT	5
	HistoricalNat	r*i1p1	NAT	5
	HistoricalGHG	r*i1p1	GHG	5
IPSL-CM5A-LR	Historical	r*i1p1	ALL	4
	HistoricalNat	r*i1p1	NAT	3
	HistoricalGHG	r*i1p1	GHG	3
bcc-csm1-1	Historical	r*i1p1	ALL	3
	HistoricalNat	r*i1p1	NAT	1
	HistoricalGHG	r*i1p1	GHG	1
NorESM1-M	Historical	r*i1p1	ALL	3
	HistoricalNat	r*i1p1	NAT	1
	HistoricalGHG	r*i1p1	GHG	1
FGOALS-g2	Historical	r*i1p1	ALL	1
	HistoricalNat	r*i1p1	NAT	2
	HistoricalGHG	r*i1p1	GHG	1

All outputs have been downloaded from the PCMDI website (http://pcmdi3.llnl.gov)

[a] For CNRM-CM5, some additional simulations have been used with respect to the ones available on the PCMDI website

2 Data and method

The observed temperature data is based on the median realisation of the HadCRUT4 merged land/sea temperature data set (Morice et al. 2012). The HadCRUT3 observed dataset, which is the previous version of the HadCRUT dataset (Brohan et al. 2006), is also used in some cases to assess the robustness of the results to slight changes in the data used. The simulated temperature data used to estimate the response patterns are provided by results from the CMIP5 archive (available at: http://cmip-pcmdi.llnl.gov/cmip5/) arising from different sets of historical simulations performed with ten different models. The type of historical simulations differs among the ten models (see Table 1). For instance, only a few models have performed AER-only simulations. For this reason, in this study, AER refers to all ANT forcings except GHG, with the AER scaling factor estimated similarly to that of R13. The size of the ensembles considered also varies substantially, from one single simulation up to 10 members. Note that the analysis is done over the 1901–2010 period, so we only consider models which provide sets of historical simulations (in particular, GHG-only and NAT-only) up to 2010. We use the same estimates of internal variability as in

R13. These are based on intra-ensemble variability from the above CMIP5 experiments as well as pre-industrial simulations from both the CMIP3 and CMIP5 archives, leading to a much larger sample than previously used (see R13 for details about ensembles). We then implicitly assume that the multi-model internal variability estimate is reliable. Further, no analyses are performed here based on individual model internal variability (i.e. the internal variability as estimated from one single model). This is mainly because the set of independent segments that can be derived from one single model is much smaller, leading to a less accurate estimate of internal variability (see e.g. Fig. 1 in R13).

Standard D&A pre-processing is applied to all observed and simulated data. This pre-processing is described in detail in R13 and only summarised here. Model output is first interpolated on the $5° \times 5°$ observational grid, and then the spatio-temporal observational mask is applied. The dimension of the data set is then further reduced by computing decadal means and projecting the resulting spatial patterns onto spherical harmonics. Results with resolutions T0 (i.e. global mean only), T1, T2 and T4 will be used to investigate the sensitivity to resolved spatial scales.

The statistical analysis is based on ROF as described by R13. The use of the same implementation as in R13 allows a direct comparison with results from the idealised analysis performed in R13. The method assumes that observed decadal mean near-surface temperature changes can be expressed as the linear sum of simulated changes due to various forcings, where the unknown quantities are the scaling factors estimated in the regression. Here we apply it to both two-forcing (ANT and NAT) and three-forcing (GHG, AER, NAT) cases. We use only the total least square (TLS) algorithm, thus accounting for the statistical uncertainty introduced by taking the model response from a finite ensemble. A residual consistency test (RCT) is also used to test whether the regression residuals are consistent with internal variability. The RCT implementation uses a non-parametric estimation of the null distribution through Monte Carlo simulations (see R13 for details). Note that we find that this null distribution is not very sensitive to the choice of response patterns used in the Monte Carlo simulations. All RCT results presented here are based on the simulated null-distribution with the CNRM-CM5 patterns, and corresponding ensemble sizes.

3 Results

3.1 Global mean temperature only

3.1.1 Two-forcing analysis (ANT + NAT)

We first apply ROF to the global mean near-surface temperature using the ANT and NAT response patterns as predictors. Detailed results, including the estimated scaling factors, the p value from the residual consistency test, as well as the time-series of the simulated and reconstructed response to each forcing, are shown in Fig. 1 for the ten models considered. Corresponding results, obtained with the HadCRUT3 observed dataset (instead of the HadCRUT4 median dataset), and a slightly smaller set of control segments to estimate internal variability, are shown in Online Resource 1 (OR1). Note that the use of another observed dataset impacts the computation of model data, as another observational mask is then applied.

Detection (i.e. scaling factor inconsistent with zero) of the ANT response pattern occurs in nine cases. The only exception is the FGOALS-g2 response pattern. In this case, the regression appears to be degenerate, because the NAT response pattern (which only accounts for variations of the solar activity in FGOALS-g2) is close to zero throughout, and the size of the ensembles used is very small (1 or 2 members). The detection of the anthropogenic influence is then very robust based on this diagnosis. Attribution (i.e. scaling factor inconsistent with zero and consistent with one) of the ANT response pattern occurs in four cases, and fails by only a small margin in four other cases (HadGEM2-ES slightly underestimates, while CanESM2, IPSL-CM5A-LR and bcc-csm1-1 slightly overestimate the ANT response). One model (CSIRO-Mk3-6-0) seems to underestimate the response to the ANT forcing substantially.

The NAT response pattern can be detected and attributed using four models (CNRM-CM5, GISS-E2-H, GISS-E2-R and bcc-csm1-1), and comes quite close when using one additional model (IPSL-CM5A-LR). The NAT response appears to be primarily overestimated by the models, because the best estimates of the corresponding scaling factor are all smaller than unity (except for FGOALS-g2). Note that two models (HadGEM2-ES and CSIRO-Mk3-6-0) have a negative (though consistent with zero) NAT scaling factor best-estimate. The response to a change in solar activity has opposite sign to that expected.

Discrimination between ANT and NAT forcings appears to be robust as a result of the weak collinearity of the corresponding response patterns. Most of these results are also robust to the use of the previous version of the HadCRUT dataset (see OR1).

Furthermore, five models among the eight providing constrained scaling factors, do not pass the RCT at the 10 % level. Rejection is even quite strong for three models (CNRM-CM5, CanESM2 and HadGEM2-ES). This result suggests, e.g., a possible error in the global mean forced response or an underestimation of internal variability. It may be noted in particular that this rejection was, if not removed, weaker based on HadCRUT3 observations (OR1). In both cases, the models seem to have difficulty in simulating a warming as large as observed for the first part

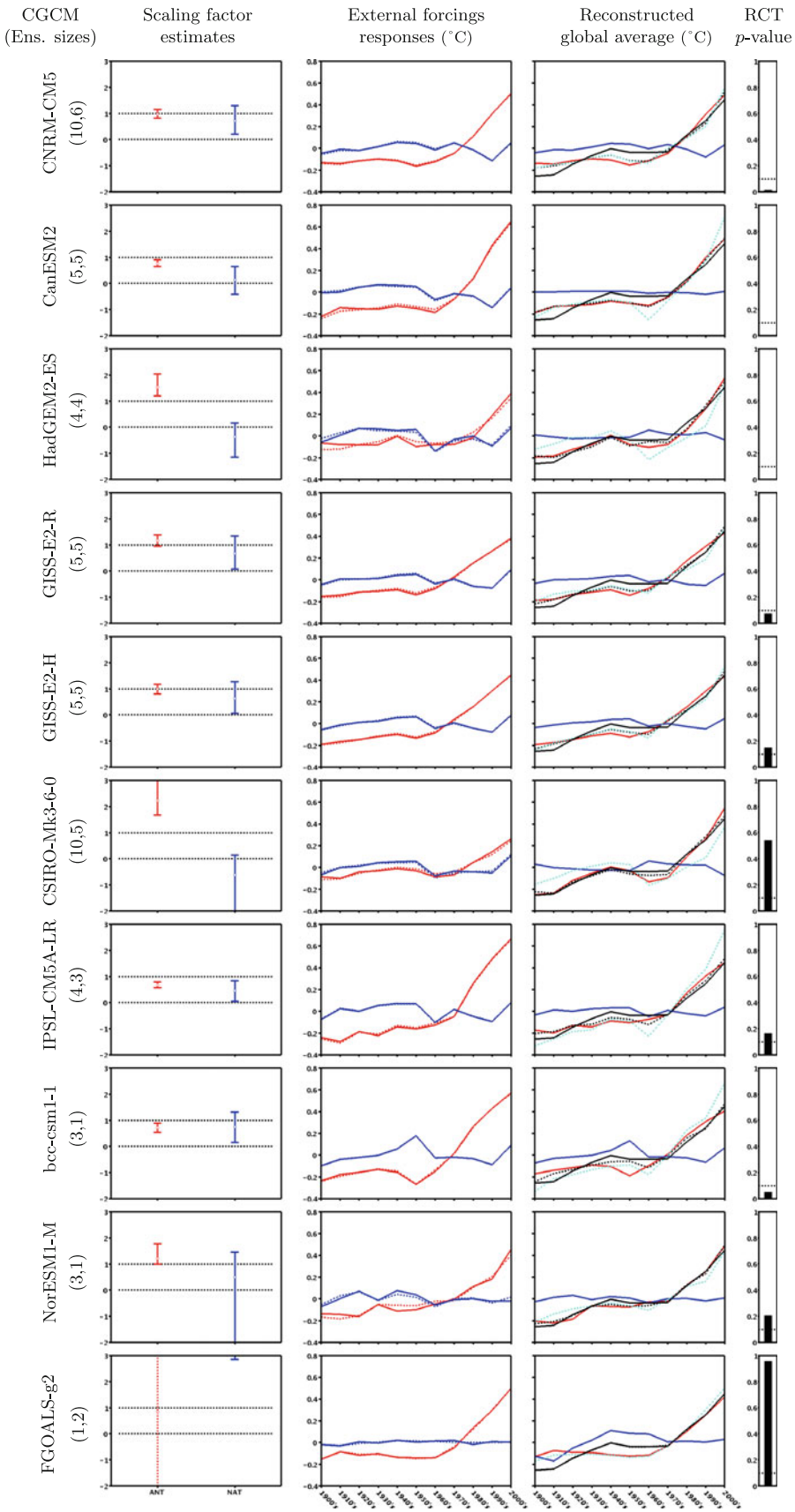


Fig. 1 Attribution analysis based on global average time-series with two external forcings. Results are shown for ten climate models from the CMIP5 database, in two-forcing analysis: ANT (*red*) + NAT (*blue*). *Left* scaling factors best-estimate (*diamond*) and confidence interval; *middle left* model response to each forcing, as computed from the CGCM outputs (*solid line*), and as reconstructed by the TLS algorithm (*dotted line*); *middle right* global temperature time-series, as measured in the HadCRUT3 observations (*solid black line*), as estimated directly by the CGCM (*dotted cyan line*), and as reconstructed by the TLS algorithm (*dotted black line*), with the scaled contributions from each forcing (*other solid lines*); *right* result from the residual consistency test in terms of *p* value. *Dotted* confidence intervals are unbounded. The global temperature time-series estimated by the CGCM is computed as the addition of the estimated responses (i.e. *solid lines* in *middle left*). In *middle right*, the TLS reconstruction of global temperature is the addition of the scaled responses. On the x-axis, 1900s denotes the 1901–910 decade, 1910s denotes the 1911–1920 decade, etc

of the twentieth century (cyan dotted line), even after rescaling the forcing responses (black dotted line). However, the discrepancy between the best fitted reconstruction and the observations seems higher in the case of the HadCRUT4 data over the middle of the last century (from the 40s to the 60s), with the reconstructed global mean temperature often more severely underestimated over this period (e.g. CNRM-CM5, CanESM2, GISS-E2-R, GISS-E2-H). This discrepancy, though small, suggests that accounting for observation uncertainty could be of interest in such a study. This could be done in another study using the different HadCRUT4 ensemble members.

3.1.2 Three-forcing analysis (GHG + AER + NAT)

We then perform the same analysis using the GHG, AER and NAT response patterns (similar as the Stott et al. (2006) study). Detailed results are shown in Fig. 2. As in the previous section, the results obtained with the HadCRUT3 observed dataset are shown in Online Resource 2 (OR2).

We first analyse results from the CNRM-CM5, CanESM2, HadGEM2-ES and IPSL-CM5A-LR models. For these models, the scaling factors are well constrained for all three forcings, meaning that none of the confidence intervals is unbounded. This suggests that ROF is able to discriminate between the three forcings. GHG scaling factors are always inconsistent with 0 and include 1 in two cases (it comes very close for IPSL-CM5A-LR) with their best estimate being smaller than one. This means that we detect the GHG influence and that the models slightly (i.e. non significantly in two cases out of four) overestimate the response compared to the observations. This contrasts with the AER forcing, which is not detectable in three out of four cases (and comes very close in the fourth case), as zero is included in the 5–95 % confidence interval. The AER response pattern, however, is found to be consistent with observations in three out of four cases. The influence of the NAT forcing is detected in three cases and the best estimates of the scaling factors, as in the AER and GHG cases, are smaller than one (although very close in two cases). However, these models differ significantly in two aspects. HadGEM2-ES and CNRM-CM5 fail the RCT at the 10 % level, while the two others barely pass. Among other discrepancies with observations, all these models underestimate, to different degrees, early twentieth century warming. As in the 2-forcing case, rejection of the RCT is more pronounced with the HadCRUT4 dataset than with the HadCRUT3 dataset (OR2). Again, this seems primarily due the model underestimation of the global mean temperature over the middle of the twentieth century, which is even stronger based on HadCRUT4. Finally, the CNRM-CM5 model is the only model in which all scaling factors are roughly consistent with one, though the AER response is not detected.

It is worth suggesting a possible explanation for the results obtained for both the GHG and AER scaling factors. In the case of the CNRM-CM5, CanESM2 and HadGEM2-ES, the discrimination between these two forcings occurs essentially as a result of the last two decades, during which the AER cooling response has been stabilising or weakening (unlike many other models). Indeed, no constraint is found on the scaling factors if the analysis is performed over the 1901–1990 period, consistent with highly correlated GHG and AER response patterns. Over the last two decades, the stabilisation of AER response has tended to reinforce the total ANT warming. This warming is partly compensated over the 1990s due to a cooling from the NAT response, partially linked to the Pinatubo eruption. Such compensation no longer occurs over the last decade, as the NAT contribution leads to a significant warming. Consequently, the simulated warming accelerates in the three models, in contrast with the observations. Low frequency internal variability may contribute to offseting or reinforcing the warming over a few decades. However, this discrepancy between models and observations constrains the AER scaling factor to be very small and, if not inconsistent, barely consistent with one, for all three models. As the AER response is decreased, it also leads to values of the GHG scaling factors being smaller than one. Note also that the case of the IPSL-CM5A-LR model is slightly different. As simulated by this model, the main cooling from the AER forcing occurs between the 1950s and the 1970s, with a stabilisation afterwards. Therefore, the AER forcing offsets the GHG-induced warming before the 1970s, which results in a small warming over this period, in line with the HadCRUT4 observations. Then, stabilisation allows a more pronounced warming after the 1970s, also in agreement with the observations. This leads to good overall results: each external forcing is detected (with AER in particular, in contrast with the three other models), with each scaling factor being relatively close to one, and no significant inconsistencies arise from the RCT.

The other models (Figs. 2, OR2) all exhibit unbounded values for the scaling factors for two principal reasons. The GISS-E2-R, GISS-E2-H and CSIRO-Mk3-6-0 models GHG and AER response patterns exhibit strong collinearity (correlations are less than -0.96). The cooling linked to the AER response shows no sign of weakening at the end of the period, in line with the GHG-induced warming. Note that this collinearity is even clearer in the reconstructed response (e.g. GISS-E2-R in Fig. 2, the correlation being then less than -0.98), suggesting that ROF is not able to discriminate between these two forcings in such a case. This degeneracy leads to very large confidence intervals with even positive and negative values in some cases.

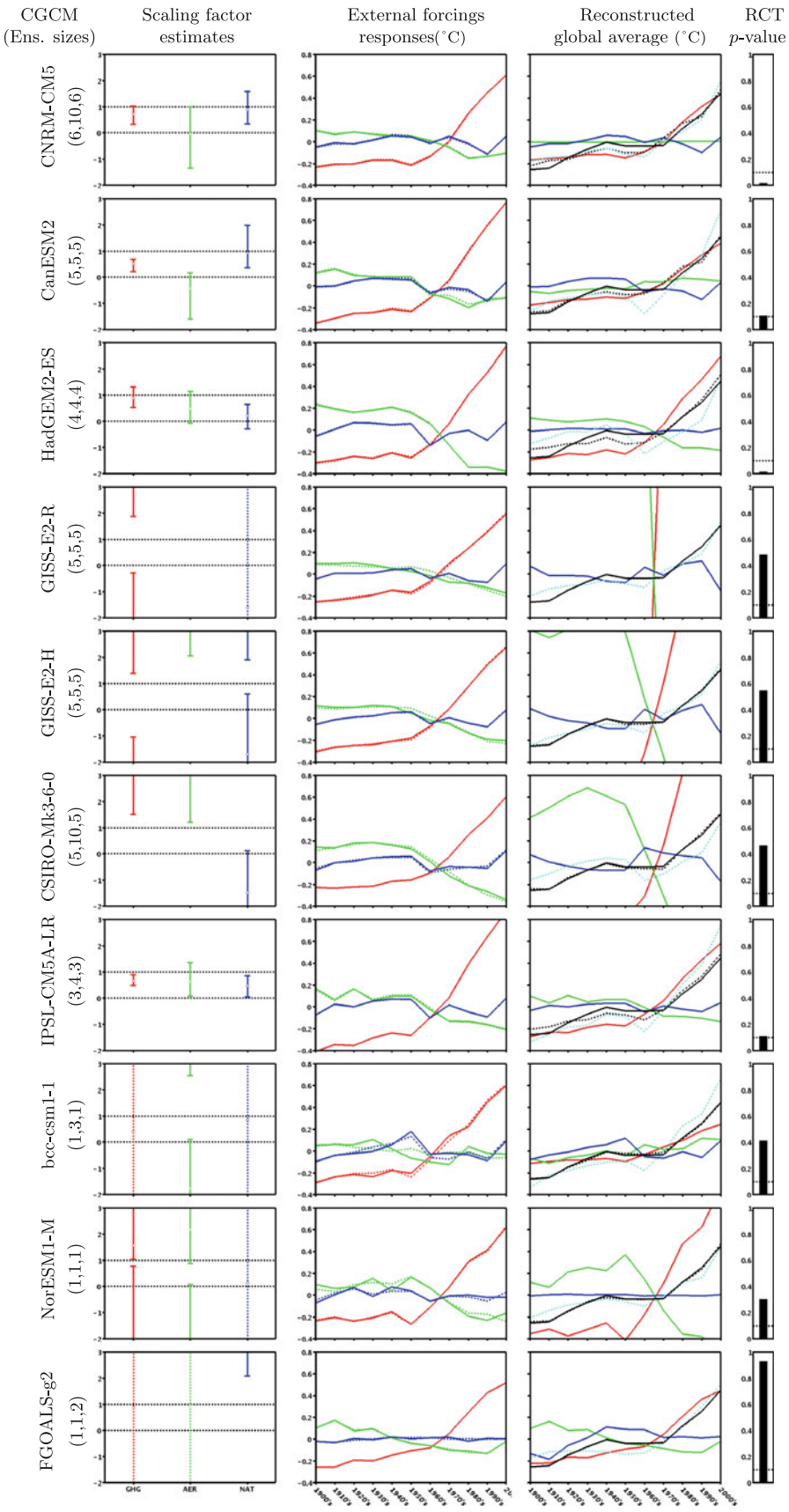


Fig. 2 Attribution analysis based on global average time-series with three external forcings. Same as Fig. 1, in three-forcing analysis: GHG (*red*) + AER (*green*) + NAT (*blue*)

It thus leads to the impossibility of assessing the contribution of various forcings. Other models show a weak response to at least one forcing and/or have very small ensemble sizes, leading directly to a degeneracy in the regression. The FGOALS-g2 model in particular has no volcanic forcing, leading to a very weak NAT response pattern (which can then be made collinear to any linear combination of the GHG and AER patterns). Despite the presence of volcanic forcing, bcc-csm1-1 and NorESM1-M show roughly the same behaviour, with a weak response to AER. Whatever the reason for the degeneracy, it usually leads to a perfect fit with the observations (the response patterns may be easily modified by the TLS algorithm so that a close fit observations is obtained). As a result, the RCT p value is usually high. The model and the observations are thus perfectly consistent, but the analysis yields no constraint on the contributions of the various forcings and no conclusion about detection and attribution.

These results suggest that, based on the global mean time-series only, discrimination between GHG, AER and NAT forcing is much harder than between ANT and NAT. Such a behaviour is expected here, as the time-only response patterns to GHG and AER have been previously shown to be highly correlated (Gillett et al. 2002; Allen et al. 2006).

3.2 Spatio-temporal analysis

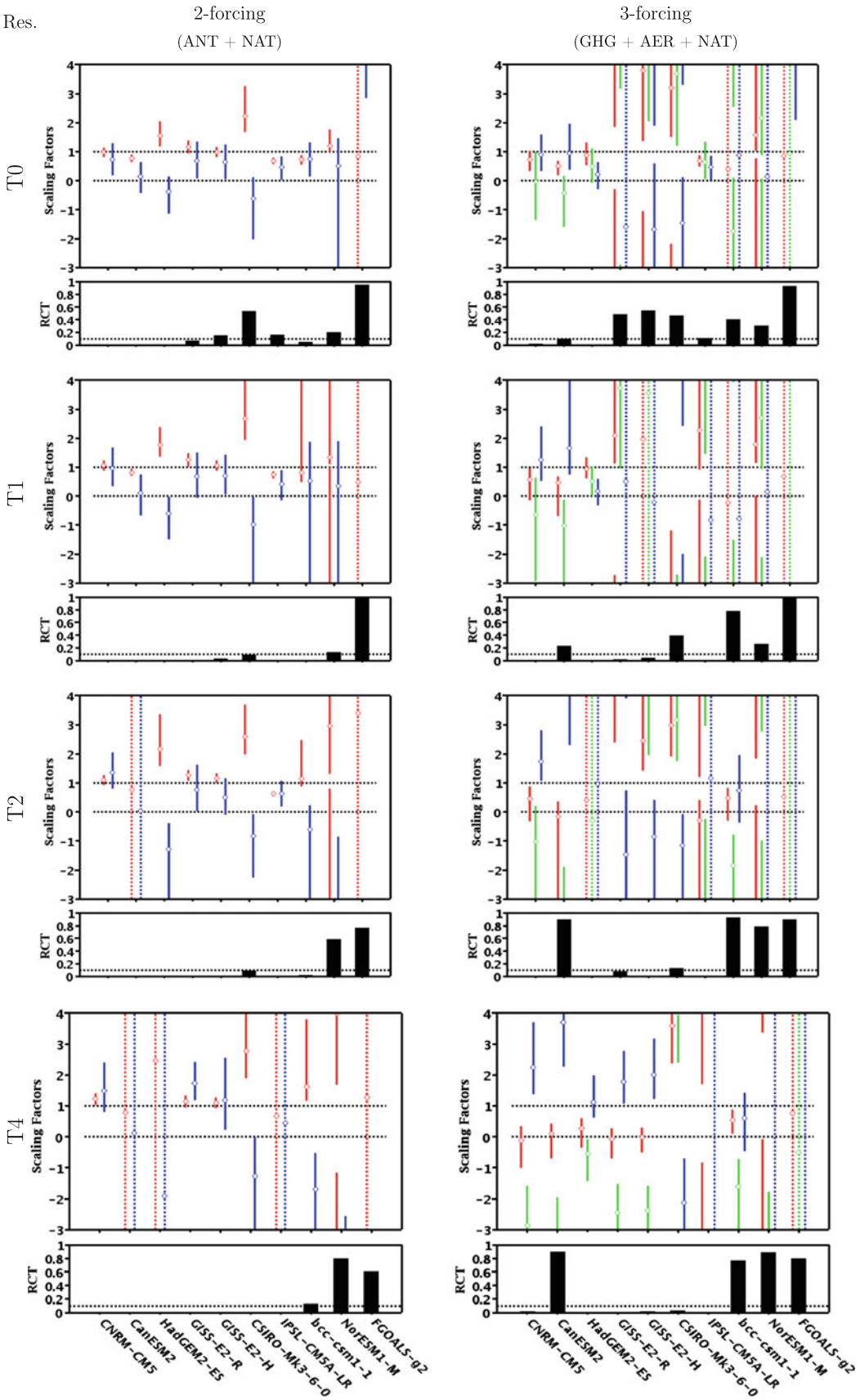
3.2.1 ROF using HadCRUT4

Previous studies have suggested that OF algorithms using information on both space and time scales have the potential ability to better distinguish between the GHG and AER fingerprints and overcome the degeneracy issue. Consequently, we now apply ROF to both two-forcing (ANT and NAT) and three-forcing (GHG, AER, NAT) cases using a detection vector with increasing amounts of spatial information. Three additional resolutions (T1, T2, T4) are used in the projection step. The results in terms of scaling factors and RCT p value are shown in Fig. 3.

Increasing spatial information in the detection vector first leads to very low RCT p values for most models, whatever the number of external forcings taken into account. Most cases with a p value higher than 10 % correspond to unconstrained scaling factors. Such a phenomenon is not surprising: if some degeneracy occurs in the regression, then the residuals are usually very small and consistent with the assumed internal variability. The only two exceptions (i.e. RCT p value higher than 10 % and constrained scaling factors) are CanESM2 and bcc-csm1-1, in the 3-forcing analysis, at resolutions T2 and T4. Note that in the case of CanESM2, this no longer holds if the historical AER-only simulation is used instead of the ALL forcings simulations. One possible explanation for this could be that additional external forcings such as ozone or land use are important. These results from the RCT must be interpreted with caution, as it was suggested in R13 that the null-distribution of this test is difficult to evaluate accurately, even based on Monte-Carlo simulations. However, the perfect model framework results shown in R13 did suggest that the RCT was too conservative. Therefore, such rejections must be taken into account, as they suggest that at least one assumption of the method is not satisfied.

Furthermore, when increasing spatial information, the stability of scaling factors best-estimates and confidence intervals is not observed. This sensitivity of the results is striking in the 3-forcing analysis, at resolutions T1 and T2, with a large variation of scaling factor best-estimates on the one hand, and much larger confidence intervals on the other hand. In particular, the four models with constrained scaling factors at T0 resolution show roughly unconstrained values at these resolutions. This also occurs to some extent in the two-forcing case, at all resolutions, though it is less pronounced at both T1 and T2 resolutions. In particular, at the T4 resolution, discrimination between ANT and NAT is sometimes not achieved. These results strongly contrast with those from the perfect model framework reported by R13 (R13 reported consistent results and narrower confidence intervals as the resolution increases).

Then, even more unexpected results are obtained in the 3-forcing analysis at T4 resolution, as several models provide relatively constrained scaling factors. First, these results are not robust to small variations in the analysis, such as using HadCRUT3 observations instead of HadCRUT4, or changes in the construction of samples Z_1 and Z_2. This is illustrated, for instance, in Online Resource 3 (OR3), which reproduces the same analysis as in Fig. 3, based on observations from the HadCRUT3 dataset. The results from ROF at T4 resolution are then much closer to those obtained at the T1 or T2 resolution, with mainly unconstrained results. Secondly, as discussed above, the RCT is not passed in most cases. Third, the scaling factors estimated here are in many cases unphysical. This occurs in particular for the AER forcing, which is found to be significantly negative in several cases. Some of these results are also inconsistent with those found at other resolutions, e.g. with disjoint confidence intervals. Note that a biased estimate of internal variability can lead to a too frequent rejection of the RCT (underestimation of the true internal variability) or an unrealistically large confidence interval (overestimation). However, it can hardly explain unphysical scaling factor estimates as observed when ROF is applied using the T2 and T4 resolutions. Fourth, similarly to Figs. 1 and 2, Online Resource 4 (OR4) investigates how well the global mean temperature changes are reproduced in this case. Note that, unlike in Sect. 3.1.2, the fit is not


Res.
2-forcing
(ANT + NAT)
3-forcing
(GHG + AER + NAT)
T0
T1
T2
T4
Scaling Factors
RCT
CNRM-CM5
CanESM2
HadGEM2-ES
GISS-E2-R
GISS-E2-H
CSIRO-Mk3-6-0
IPSL-CM5A-LR
bcc-csm1-1
NorESM1-M
FGOALS-g2

◀ **Fig. 3** Results from the spatio-temporal analysis, as a function of the spatial resolution. ROF is applied to both observed (HadCRUT4) and simulated (CMIP5 models) data after projection onto T0 (or global mean only, *top*), T1 (*middle top*), T2 (*middle bottom*) or T4 (*bottom*) spherical harmonics. The analysis is performed under a two-forcing (ANT + NAT), or a three-forcing analysis (GHG + AER + NAT), and applied to 10 CGCMs from the CMIP5 database. Shown are the scaling factors best-estimates (*diamond*) and confidence-intervals (*colored bars*), together with the *p* value from the residual consistency test (*black bars*). Dotted confidence intervals are unbounded. Note that *black bars* cannot be seen in case of too small *p* values

only based on global mean time series here. This figure suggests that, in many cases, the AER response pattern is widely redrawn by the method in order to best fit the data, while the GHG response pattern is virtually removed (with a small scaling factor). This could be understood as follows. In the TLS algorithm, both the observations y and the response patterns x_i (see R13 for notation) are noisy and may be modified to obtain the best fit. Then, as a rescaling of the amplitude is allowed, only the shape of the signal matters. In our case, the AER and NAT signals are relatively small compared to the GHG signal, and at the same time, the internal variability (related to the size of the corresponding ensemble) is roughly the same, if not higher. As a consequence, the shape of these response patterns is more uncertain than the shape of the GHG response pattern, and the best fit is obtained by modifying these shapes first. Given the relatively small size of the ensembles used, the signal-to-noise ratio on these signals seems even lower than in the observations. The OR4 illustrates this phenomenon, as the NAT and, above all, the AER signals are widely redrawn by the algorithm. The GHG response, conversely, is not modified, but is virtually removed, as the corresponding scaling factor is very small. Finally, we are led to question whether the confidence intervals computed are suitable in such a case, as the high signal-to-noise assumption mentioned by AS03 for the uncertainty analysis may not be satisfied.

In order to investigate what might have happened with a higher signal-to-noise ratio (i.e. larger ensembles), we performed the same analysis with the size of the ensembles arbitrarily increased by a factor of ten. More precisely, the nominal ensemble size—which is one input of the method—is increased, whereas the actual ensemble size is not modified. The other input data, i.e. the observations y and the noisy response patterns $\widetilde{x_i}$, are identical. This protocol is somewhat unrealistic because the response patterns will then contain more noise than the algorithm assumes. In this respect, this is similar to using the OLS approach (which assumes no noise in $\widetilde{x_i}$), with noisy signals. Figure 4 shows the results provided by this analysis at T4 resolution. It may be compared to the fourth row of Fig. 3. The uncertainty is then strongly reduced, with all confidence intervals being well-constrained, and no significantly negative values. Note that the OLS estimates based on noisy signals are known to be biased towards zero (see e.g. AS03), which is very consistent with our results. This suggests that an increase of the ensemble size could help to better discriminate between different forcings (at T4 resolution). In particular, it could prevent us from finding a best fit where the response patterns are substantially redrawn, as described above. Note that the use of

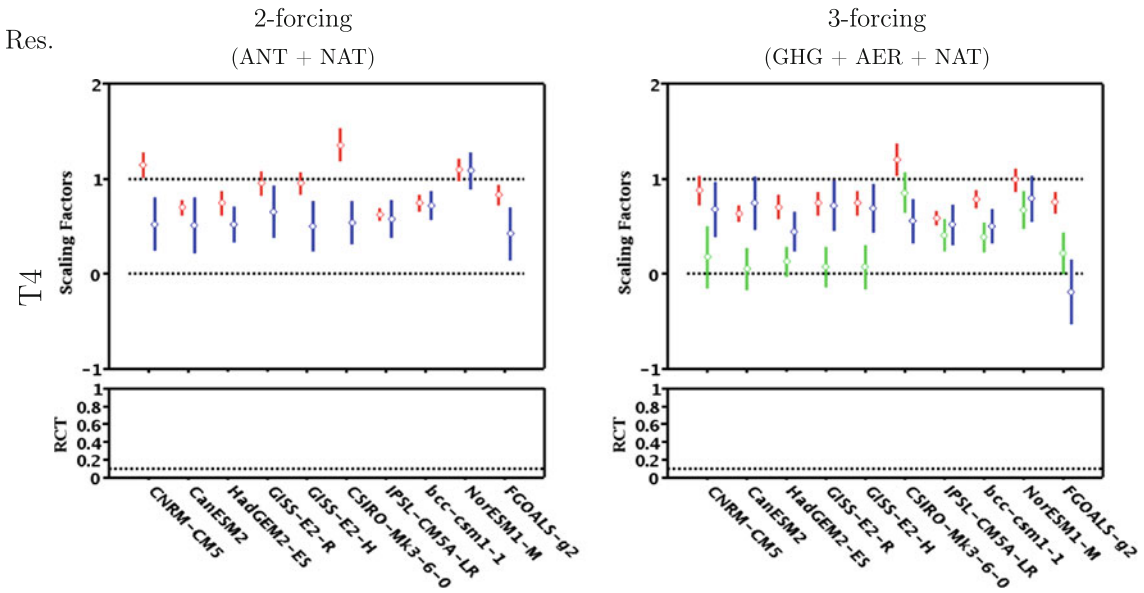


Fig. 4 Results from the spatio-temporal analysis, based on 10 times larger ensembles. Same as in Fig. 3 with T4 spherical harmonics, but with the size of each ensemble of simulations arbitrarily increased by a factor 10 (see text). Note the different *y*-axis scale from that of Fig. 3

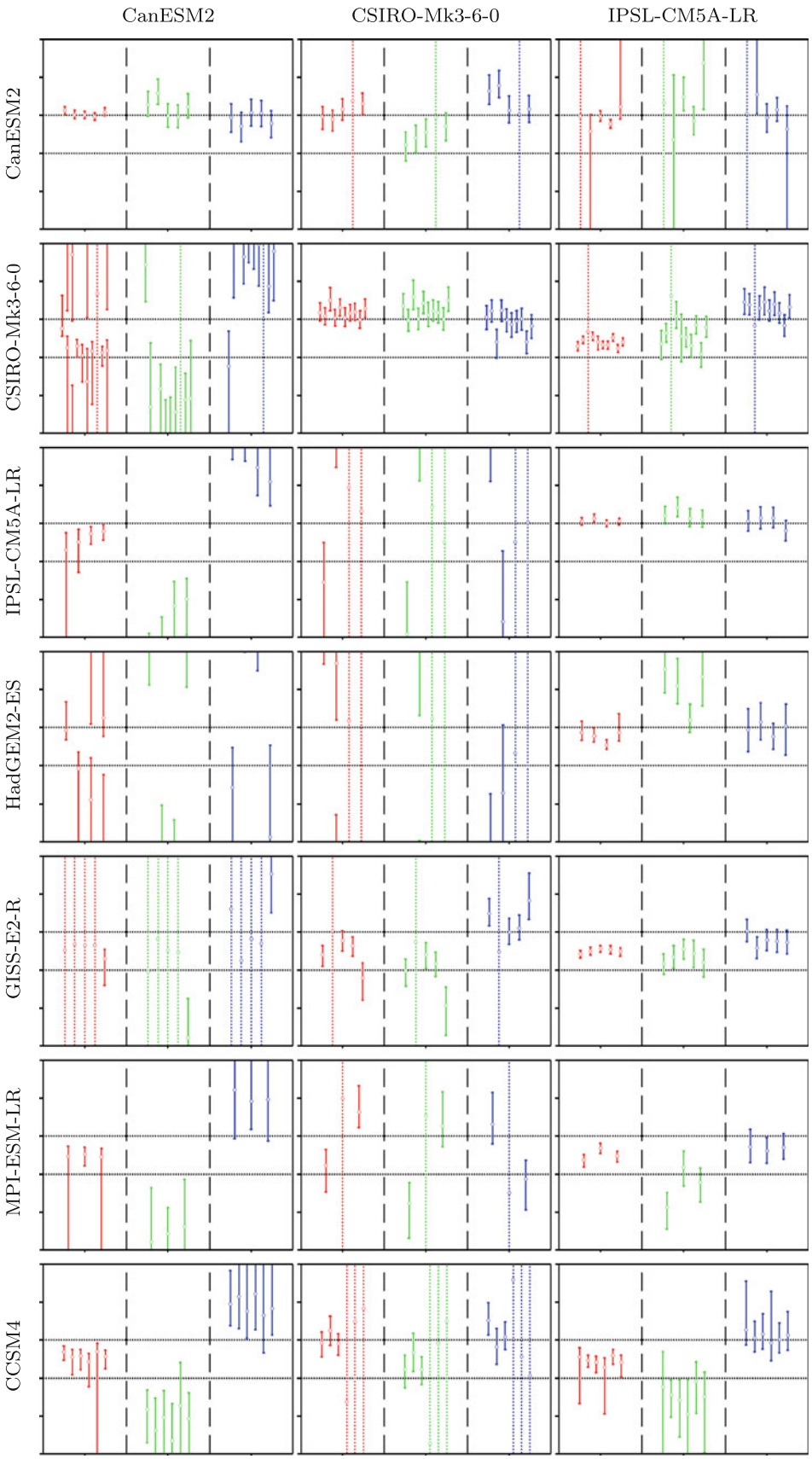


Fig. 5 Cohort framework analysis, Part I. Scaling factors estimates derived within the cohort framework analysis: historical simulations (sometimes extended with RCP8.5 simulations over the 2006–2012 period) from 7 CMIP5 models are used as pseudo-observations (*rows*), response patterns are taken from 3 CMIP5 models (*columns*). Scaling factors are estimated in a 3-forcing (GHG + AER + NAT) analysis with ROF, at resolution T4, over the 1901–2010 period. The panels on the diagonal correspond to perfect model framework, as the same model is used to provide pseudo-observations as well as response patterns. Additional results are shown in Online Resource 7

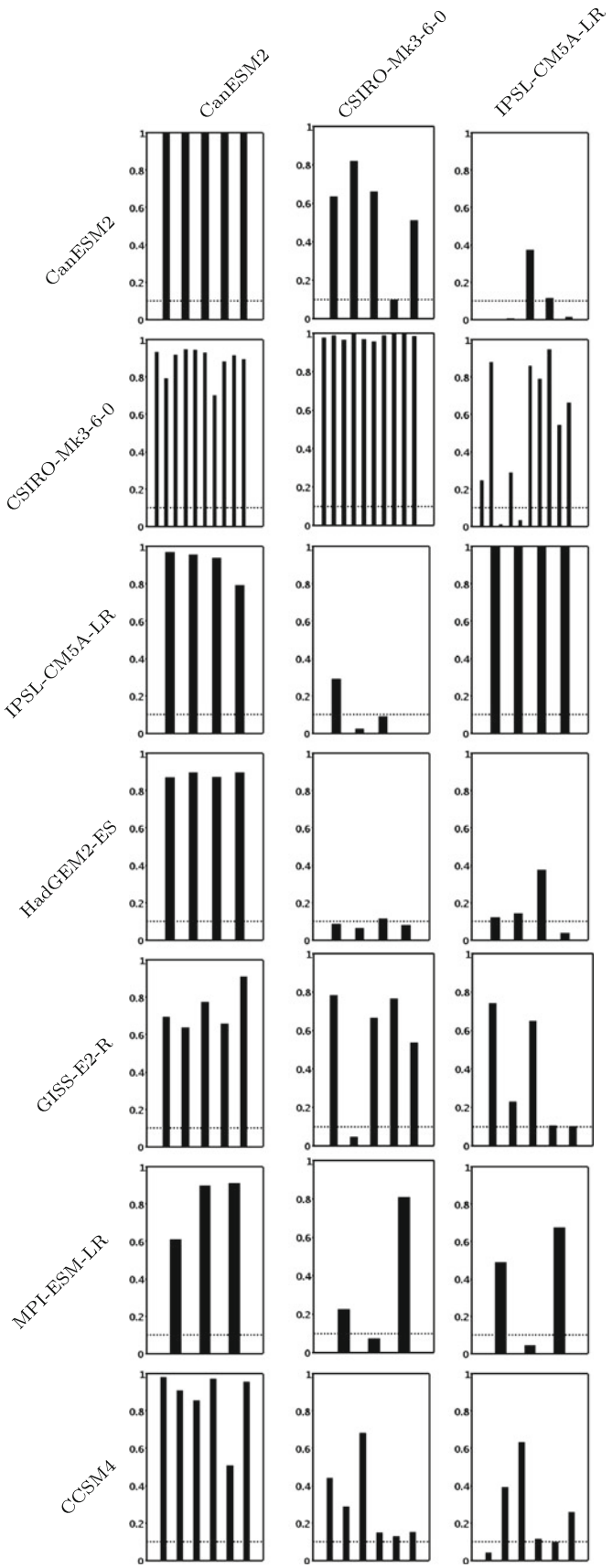


Fig. 6 Cohort framework analysis, Part II. RCT p values derived within the cohort framework analysis (similar to Fig. 5). Additional results are shown in online Resource 8

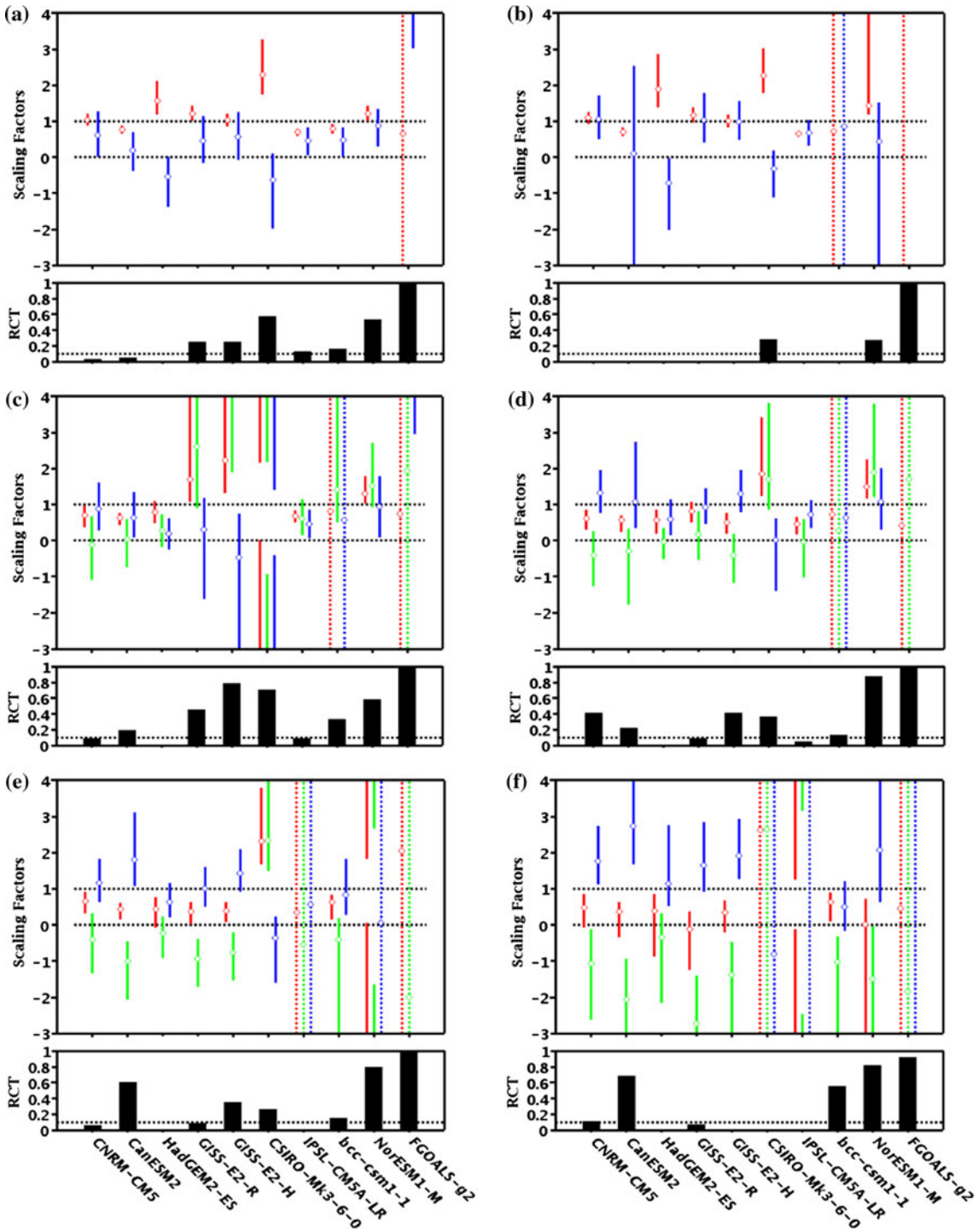


Fig. 7 EOF projection results. Scaling factor estimates and RCT p value from EOF projection, for several values of the number of EOFs retained k, for 10 CMIP5 models. The standard EOF projection is applied to T4 resolution data. Results are shown in a two-forcing analysis with $k = 20$ and $k = 60$ (*top*, respectively **a**, **b**) and in a three-forcing analysis with $k = 20$, 40, 60, 120 (*middle* and *bottom*, respectively **c**, **d**, **e** and **f**)

multi-model response patterns may provide one method of obtaining larger ensemble, as has been done in other studies (e.g. Gillett et al. 2012). Note also that this artificial increase of the ensemble sizes leads to even smaller RCT p values. This is expected here, as the response patterns do contain more noise than the algorithm assumes.

3.2.2 Cohort framework analysis

Another useful diagnosis comes from applying ROF at T4 resolution to pseudo-observations from historical climate model simulations (as done in R13 based on CNRM-CM5 outputs), with the estimated response patterns coming from another model. Here, we refer to this protocol as a cohort framework analysis. It extends the perfect model framework used in R13. Figures 5 and 6 show the results obtained at T4 resolution, in a 3-forcing analysis, with historical simulations from 7 climate models taken as pseudo-observations, while 3 climate models are used to provide response patterns. A more complete overview, based on 14 climate models taken as pseudo-observations, and 7 climate models providing response patterns, is provided in Online Resource 5, 6 (2-forcing analysis), 7 and 8 (3-forcing analysis). Note that OR5 and OR7 show scaling factor estimates, while RCT p values are shown in OR6 and OR8.

First, the diagonal panels correspond to the perfect model framework, as the same model provides both pseudo-observations and response patterns. Note that in some cases, these results must be interpreted with some caution, as the response pattern estimates are not strictly independent from the pseudo-observations (meaning that the ensemble mean of the ALL-forcing historical simulations is used to derive at least one response pattern). These diagonal results are very consistent with those obtained with CNRM-CM5 (see R13, Figs. 4, 5), as scaling factor confidence intervals are very narrowly distributed around unity. Note that each of the 7 models considered presumably has its own specific internal variability (i.e. the features of internal variability may differ from one model to another). However, satisfying results are obtained here when the same estimate of the internal variability covariance matrix (derived from a multi-model ensemble) is used. So the use of a common, multi-model estimate of internal variability does not seem to deteriorate the accuracy of ROF substantially. As for the RCT, similar behaviour is found as in R13, with p values clearly larger than expected, resulting in strong acceptance of the test. As mentioned in R13, this may be due to a poor estimation of the null-distribution. These results, however, tend to confirm the conclusion that this leads to a too conservative test.

Second, the off-diagonal panels provide some indication as to how the results can be impacted by some error in the expected response patterns. Here again, different models presumably lead to different response patterns, although many common features certainly appear. Figure 5 shows that several scaling factor confidence intervals are still constrained, suggesting that this does not always lead to a huge deterioration of the results. This is clearer if the whole set of climate models is considered (OR7), or in a 2-forcing analysis (OR5). Several confidence intervals exclude one, which is consistent with a discrepancy in the sensitivity to some forcing. However, these confidence intervals are often larger than those observed in the perfect model framework (Fig. 5 and, e.g., second and third rows in OR5), meaning that an inappropriate assumption on the response patterns impacts the accuracy of the method. This increase of the confidence interval ranges can be very pronounced and may even lead to unconstrained scaling factors (many cases in Fig. 5 and, e.g. in OR5, if simulations from IPSL-CM5A-LR are used as pseudo-observations). This suggests that using model response patterns significantly different from the true patterns may lead to unconstrained scaling factors. Moreover, several off-diagonal panels show small RCT p values, sometimes leading to a rejection of the test, although this test has been shown to be too conservative. This behaviour is somewhat clearer in the 2-forcing analysis (OR6). In this case, some of the pseudo-observations used even lead to rejection of the RCT with almost all response patterns (see e.g. bcc-csm1-1, MPI-ESR-LM, IPSL-CM5A-LR or, to a lesser extent, CCSM4). Such behaviour is similar to that obtained with real observations.

Finally, the cohort framework analysis suggests that the scaling factor and RCT results obtained with true observations are not consistent with what might have been expected assuming that the models simulate the external forcing response patterns correctly. It also suggests that biased response patterns may at least partially explain the deterioration of the results obtained when using real observations.

3.2.3 Comparison to EOF projection

We now compare the results provided by ROF for the HadCRUT4 dataset to the results provided by the standard OF implementation for the same data. Standard OF implementation means projection onto k leading EOFs, as discussed in R13 and several previous papers (e.g. Allen and Tett 1999). The analysis is performed with the same models and pre-processing, at T4 resolution, to allow a direct comparison. Results are shown in Fig. 7 for different choices of the number of EOF retained k (namely the truncation).

In a two-forcing analysis, the cases $k = 20$, 60 illustrate the relative robustness of the results, at least for the ANT forcing. This robustness is less marked for much higher values of k (e.g. $k > 100$, not shown). As in R13, the results obtained with $k = 20$ are very close to those provided by ROF at T0 resolution. This similarity occurs in particular for scaling factor estimates, but also, to a lesser extent, for RCT p values. Note that in both approaches, the dimensionality of the initial T4 data is strongly reduced (either by

projection on EOFs or spherical harmonics), leading to an effective dimension of 20 and 10, respectively. Then, some additional consistency may be found with the ROF results at resolutions higher than T0, as the RCT is often rejected (very sharply for several models) if $k = 60$.

In a three-forcing analysis, the results provided by EOF projection ($k = 20, 40, 60, 120$) seem more sensitive to the truncation, with the detection or the attribution to some forcing dependent on the value chosen. The response to the GHG forcing, for instance, is found to be sometimes overestimated ($k = 20$), mainly overestimated ($k = 40$, 60), or not clearly detectable ($k = 120$). Sensitivity seems even higher for the AER and NAT scaling factors. Again, the results obtained with $k = 20$ share many common features with those obtained with ROF at T0 resolution. Some similarity may also be found between the case with $k = 120$ and ROF results at T4 resolution. Note that the choice of $k = 120$ is relatively close to the optimal value of k found in R13 based on Monte-Carlo simulations.

Figure 7 suggests that standard EOF projection results (in terms of scaling factor best estimate and confidence intervals) do show a significant sensitivity to k, leading to ambiguity in interpretation of results. The RCT doesn't clearly allow removal of this ambiguity, as some models do not pass the RCT at any of the four truncations considered here, while others pass the RCT with $k = 120$, etc. Furthermore, the sensitivity to truncation, although not widely investigated here, seems somewhat higher than that reported in R13 when the same algorithm was applied in a perfect model framework. So EOF projection results, as well as ROF results, do not seem consistent with what might have been expected.

4 Discussion and conclusions

4.1 ROF

This paper was primarily intended to provide a first application of ROF to observed global near-surface temperature, based on CMIP5 simulations. Improvements in observationally-constraining the response to external forcings might have been expected given the potential of ROF illustrated in R13. Strictly speaking, these improvements were not found. The results provided by ROF at T4 resolution differ from those obtained in the perfect model framework in, at least, two aspects. First, the RCT is almost always rejected. Second, the estimated scaling factors seem relatively unstable, and do not allow clear discrimination of the external forcing considered (confidence intervals are very large). Both aspects suggest some lack of consistency between models and observations, as the analysis of observations do not lead to the expected results.

These poor results seem to highly depend on spatial resolution. As long as global mean signals are investigated (i.e. T0 resolution), scaling factors are reasonably well constrained (the 2-forcing case in particular), or understandably unconstrained (3-forcing case). Rejection of the RCT is also less pronounced at such a coarse resolution. So, the deterioration of the results seems related to adding more detailed spatio-temporal information into the analysis. In the case of ROF, this is done by increasing the spatial resolution. It is important to note that a similar phenomenon occurs with EOF projection when the truncation is substantially increased (what may also be seen as adding spatio-temporal details). For instance, EOF projection based on $k = 120$ seems to provide larger confidence intervals than obtained at smaller truncation, in disagreement with the Monte-Carlo simulations reported in R13, which showed a smaller mean quadratic error at this truncation. Similar conclusion may be drawn from RCT results. It is also important to note that this investigation of the sensitivity of the results to the amount of spatio-temporal detail accounted for is heavily related to the use of a set of control segments much larger than previously used. Indeed, many studies have been based on a few tens of such segments (while we are currently considering a few hundred), which makes the investigation of $k = 120$ largely unattainable. Yet, at very modest truncations, rejection of the RCT was quite often reported when increasing k. The underestimation of the smallest eigenvalues was then invoked to explain this rejection. Our results suggest that the rejection occurs even without such underestimation. Note that the number of segments used also matters in the case of ROF. A smaller number of segments would have meant stronger regularisation in the LW estimate, and much lower weight given to high-order EOFs. This would presumably lead to a smaller sensitivity to the spatial resolution (i.e. results much closer to T0 resolution or identically to EOF projection with a low truncation). Finally, it seems that some discrepancy appears between observations and models, as the resolution increases, preventing the expected OF results improvement.

We now discuss possible explanations of these results, as well as possible ways to improve the method. The very frequent rejection of the RCT suggests that at least one of the basic assumptions behind the statistical model used is not satisfied. In this study, we investigate, in particular, to what extent this phenomenon may be due to an imperfect estimation of the response patterns by climate models. The cohort framework analysis suggests that the discrepancies between model response patterns may have a substantial impact on the ROF results. It even leads in some cases to a deterioration of the results comparable to that observed when using the HadCRUT4 dataset. In our view, this creates a first, important concern about the statistical method

used, as model uncertainty is not taken into account in the TLS statistical model. The inclusion of this uncertainty in an EIV approach, as proposed by Huntingford et al. (2006), will be a natural continuation of this work. This will require adapting ROF to the EIV statistical model. Further, note that significant discrepancies between model spatial patterns together as well as between simulated and observed spatial patterns have been reported e.g. by Shin and Sardeshmukh (2011) in the case of recent sea surface temperature trends. It may also be noted that, from an historical perspective, the assumption that the response to the forcing is known up to a certain scaling factor has been introduced in order to account for a potentially unknown feedback that might change the amplitude of the response, but not the spatio-temporal pattern. Because many feedbacks impact both the amplitude and the pattern of the response, this assumption may be regarded as somewhat simplistic.

Other plausible explanations could also be mentioned. First, we have seen that using HadCRUT3 instead of HadCRUT4 may impact some results, suggesting that accounting for observational as well as model uncertainty is clearly of interest. HadCRUT4 ensemble members will allow further investigation into this question. Secondly, we do not investigate the use of alternative estimates of internal variability. Alternatives may come from using one single model instead of a multi-model ensemble (although this would presumably strongly decrease the number of independent segments involved), using other estimators or other regularisation techniques. Using an inaccurate estimate of internal variability may, of course, have substantial implications. However, one important result is that the multi-model, regularised estimate used in this study provided successful results in the perfect model framework, with all models considered. Then, while poor estimation of internal variability may explain the rejection of the RCT, it can hardly be used to explain the large changes in scaling factors best-estimate displayed in Fig. 3. So this issue does not seem to be the dominant uncertainty. Third, other issues may arise regarding, for instance, the validity of the additivity assumption. While this study provides some useful information with respect to the impact of model errors, future investigations will be required to confirm our interpretation and develop solutions.

We also suggest that some thought may be needed regarding the pre-processing step usually performed in OF analyses. The choice of spherical harmonics is quite naive in order to focus on large spatial scales. But, given the lack of consistency between models and observations found here, the question becomes, to what extent do models and observations match? In particular, can we find spatio-temporal pre-filtering that ensures better consistency? Other simple, physically-based approaches such as those suggested by Karoly and Braganza (2001); Drost and Karoly (2012), based on simple climate indices, could lead to improved results in this respect.

4.2 Detection and attribution results

Although some discrepancy between models and observations is certainly found when refined spatio-temporal information is taken into account, the results obtained at coarser resolution, or equivalently, at low truncation, deserve several comments.

The results obtained in assessing the contributions of anthropogenic versus natural forcings to changes in global mean temperature (Sect. 3.1.1) supports the conclusion that most of the warming observed over the last 110 years can be attributed to human influence. This is still widely supported by the results shown at T1 resolution, or with EOF projection, at low truncation. This conclusion is very consistent with many previous studies (see e.g. Hegerl et al. 2007). Nevertheless, it is further reinforced here, as it is supported by a set of climate models much wider than previously used. One additional, less common conclusion that may be drawn from the two-forcing analysis pertains to the RCT, which is not passed in several cases. This seems to be related to the difficulty involved in climate models to simulate the early twentieth century accurately. Further investigations could be useful in achieving a better understanding of this result, which may be related to, e.g., observational and model uncertainty.

The discussion about the respective contributions of the GHG, AER and NAT forcings must be undertaken with caution. Several previous studies have addressed this issue for global near-surface temperatures, in particular Stott et al. (2006), Hegerl et al. (2007) and a more recent study by Gillett et al. (2012). Results from the first two sources show well-constrained scaling factors using results from four CGCMs. Direct quantitative comparison with Stott et al. (2006) is difficult, however, as all the components (the statistical method, the observed data, the estimate response patterns and the estimate of internal variability) of the analysis differ. However, the results shown in Fig. 3 at T0 resolution and with the four models with constrained scaling factors are qualitatively coherent with those presented by Stott et al. (2006) and Hegerl et al. (2007). A more meaningful comparison can be made with the study by Gillett et al. (2012) as they apply the standard OF approach to two CMIP5 models also used here, with very similar data and pre-processing. Scaling factor results from CanESM2 over the 1901–2010 period (their Fig. 3a, fourth set of bars) are rather similar to those obtained here using T0 resolution: the GHG and NAT responses are detected, with some overestimation of the GHG response by CanESM2. The agreement is less marked when examining

the CNRM-CM5 results [see auxiliary materiel from Gillett et al. (2012)]. In both studies, however, detection of the GHG and AER responses with CNRM-CM5 is seen to be very sensitive to methodological details (truncation and use of global mean only in Gillett et al. (2012), use of T0 or T1 resolutions here).

Some discrepancies may also be found, however, between our results and these studies. They reported constrained scaling factors with all response patterns considered. Conversely, at T0-resolution, we obtain comparable results for only about a half of the models involved. This means that, if we believe that all models provide a plausible estimate of the response patterns, then the discrimination between GHG and AER based on global average observational constraint is not as clear. This was related, in particular, to the stabilisation of the AER response over the last decades, which still seems uncertain. This discrepancy may be related to the number of models involved here, which is larger than previously used. Obviously, this discrepancy may also be explained partly by differences in the statistical method (ROF at T0-resolution, versus EOF projection). Figure 7 suggests that EOF projection with $k = 40$ potentially allows better discrimination between these forcings based on observations. Even there, however, we consider that the results should be viewed cautiously. First, these results seem sensitive to the choice of k, with no clear, a priori reason to chose $k = 40$. To our knowledge, physically-based reasons for such a choice have not been proposed. The use of the same method with a different, but also acceptable choice of truncation may lead to very different conclusions in assessing the respective GHG and AER contributions. Secondly, attempts to make this choice more objective (such as ROF, or the choice of k by optimising the accuracy of the scaling factor estimate, as suggested in R13), do not lead to confirmation of these results. Third, the consistency between models and observations, in this space of the 40 leading EOFs, is still debatable based on the RCT. In general, these results suggest that the observational constraint on magnitudes of the GHG and AER contributions is still relatively weak.

Finally, we suggest that better discriminating the GHG and AER responses from observations will require further methodological improvements. In particular, we suggest that improvements on the statistical methods and models used, e.g. by taking into account a wider spectrum of uncertainties than considered in this paper, could help to distinguish between these two forcings. It could increase the strength of the separation of the two signals, and make the estimation of both more accurate. This is an important challenge, as an improved estimation of the GHG response in person is directly related to the estimation of climate sensitivity.

Acknowledgments We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1) for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We acknowledge Sophie Tyteca for great technical help on the data pre-processing.

References

Allen M, Stott P (2003) Estimating signal amplitudes in optimal fingerprinting, Part I: theory. Clim Dyn 21:477–491. doi:10.1007/s00382-003-0313-9

Allen M, Tett S (1999) Checking for model consistency in optimal fingerprinting. Clim Dyn 15(6):419–434

Allen M, Gillett N, Kettleborough J, Hegerl G, Schnur R, Stott P, Boer G, Covey C, Delworth T, Jones G, Mitchell J, Barnett T (2006) Quantifying anthropogenic influence on recent near-surface temperature change. Surv Geophys 27(5):491–544. doi:10.1007/s10712-006-9011-6

Brohan P, Kennedy J, Harris I, Tett S, Jones P (2006) Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. J Geophys Res 111:D12106. doi:10.1029/2005JD006548

Drost F, Karoly D (2012) Evaluating global climate responses to different forcings using simple indices. Geophys Res Lett. doi:10.1029/2012GL052667

Gillett N, Hegerl G, Allen M, Stott P, Schnur R (2002) Reconciling two approaches to the detection of anthropogenic influence on climate. J Clim 15:326–329

Gillett N, Arora V, Flato G, Scinocca J, von Salzen K (2012) Improved constraints on 21st-century warming derived using 160 years of temperature observations. Geophys Res Lett 39(L01704). doi:10.1029/2011GL050226

Hasselmann K (1979) On the signal-to-noise problem in atmospheric response studies. In: Shaw DB (ed) Meteorology over the tropical oceans. Royal Meteorological Society, pp 251–259

Hasselmann K (1993) Optimal fingerprints for the detection of time-dependent climate change. J Clim 6(10):1957–1971

Hasselmann K (1997) Multi-pattern fingerprint method for detection and attribution of climate change. Clim Dyn 13(9):601–611

Hegerl G, Von Storch H, Santer B, Cubash U, Jones P (1996) Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. J Clim 9(10):2281–2306

Hegerl G, Zwiers F, Braconnot P, Gillet N, Luo Y, Marengo Orsini J, Nicholls N, Penner J, Stott P (2007) Understanding and attributing climate change. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Huntingford C, Stott P, Allen M, Lambert F (2006) Incorporating model uncertainty into attribution of observed temperature change. Geophys Res Lett 33:L05710. doi:10.1029/2005GL024831

Karoly D, Braganza K (2001) Identifying global climate change using simple indices. Geophys Res Lett 28(11):2205–2208. doi:10.1029/2000GL011925

Morice C, Kennedy J, Rayner N, Jones PD (2012) Quantifying uncertainties in global and regional temperature change using an

ensemble of observational estimates: the hadcrut4 data set. J Geophys Res 117(D8). doi:10.1029/2011JD017187

Ribes A, Azaïs J-M, Planton S (2009) Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. Clim Dyn 33(5):707–722. doi:10.1007/s00382-009-0561-4

Ribes A, Terray L, Planton S (2013) Application of regularised optimal fingerprinting to attribution. Part I: method, properties and idealised analysis. doi:10.1007/s00382-013-1735-7

Shin SIS, Sardeshmukh D (2011) Critical influence of the pattern of tropical ocean warming on remote climate trends. Clim Dyn 36(7–8):1577–1591

Stott P, Mitchell J, Allen M, Delworth D, Gregory J, Meehl G, Santer B (2006) Observational constraints on past attributable warming and predictions of future global warming. J Clim 19(13): 3055–3069