Combining temperature and precipitation to constrain the aerosol contribution to observed climate change

Carla M. Roesch\textsuperscript{a}, Andrew P. Ballinger\textsuperscript{a}, Andrew P. Schurer\textsuperscript{a}, Gabriele C. Hegerl\textsuperscript{a}

\textsuperscript{a}School of GeoSciences, University of Edinburgh, Edinburgh, EH9 3FF, United Kingdom

\textbf{Corresponding author:} Carla M. Roesch, carla.roesch@ed.ac.uk

\textbf{Early Online Release:} This preliminary version has been accepted for publication in \textit{Journal of Climate}, may be fully cited, and has been assigned DOI 10.1175/JCLI-D-23-0347.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2024 The Author(s). Published by the American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License.
ABSTRACT: Using the past to improve future predictions requires an understanding and quantification of the individual climate contributions to the observed climate change by aerosols and greenhouse gases (GHG), which is hindered by large uncertainties in aerosol forcings and responses across climate models. To estimate historical aerosol responses, we apply detection and attribution methods to attribute a joint change in temperature and precipitation to forcings by combining signals of observed changes in tropical wet and dry regions, the interhemispheric temperature asymmetry, global mean temperature (GMT) and global mean land precipitation (GMLP). Fingerprints representing the climate response to aerosols (AER) and the remaining external forcings (noAER; mostly GHG) are derived from large-ensembles of historical single- and ALL-forcing simulations from three models in phase 6 of the Coupled Model Intercomparison Project and selected using a perfect model study. Results from an imperfect model study and a hydrological sensitivity analysis support combining our choice of temperature and precipitation fingerprints into a joint study. We find that diagnostics including temperature and precipitation slightly better constrain the noAER signal than diagnostics based purely on temperature or GMT-only and allow for the attribution of AER cooling (even when GMT is not included in the fingerprint). These results are robust across, using fingerprints from different climate models. Estimated contributions for AER and noAER agree with estimates from the most recent IPCC report. Finally, we attribute a best estimate of 0.46 K (0.05-0.86 K) of aerosol-induced cooling and of 1.63 K (1.26-2.00 K) of noAER warming in 2010-2019 relative to 1850-1900 using the combined signals of GMT and GMLP.
SIGNIFICANCE STATEMENT: Aerosols are small liquid or solid airborne particles. They are predominantly the secondary result of emissions of aerosol precursor gases emitted via industrial or natural processes. While greenhouse gases warm the climate, aerosols can have a cooling effect on the climate system, thus offsetting some of the greenhouse gas-related warming. We expect greenhouse gas concentrations in the atmosphere to continue to increase, while aerosol concentrations are likely going to decline due to their impacts on human health. Our study uses observed temperature and precipitation changes to quantify how much aerosols have offset warming from past greenhouse gas emissions. This can help to constrain future predictions of global warming.

1. Introduction

Anthropogenic aerosols are small liquid or solid airborne particles. They are predominantly the secondary result of emissions of aerosol precursor gases emitted via industrial (e.g., sulphur dioxide) processes, and are found to have a net negative effective radiative forcing (Bellouin et al. 2020). By cooling the climate they have offset some of the past greenhouse gas (GHG) induced warming (Foote 1856; IPCC 2021a). It is becoming increasingly important to understand and quantify the individual impacts these two forcings have had on the climate system over the historical period, to improve constraints on observed GHG effects and thus future predictions, given GHG and aerosol emissions will likely follow opposing pathways (Andreae et al. 2005; Lund et al. 2019; Larson and Portmann 2019; Persad et al. 2022).

While GHGs trap (outgoing) longwave (LW) radiation, heating the climate system, aerosol-associated cooling results from aerosols’ interactions with (incoming solar) radiation and clouds, referred to as aerosol-radiative (ARI) and aerosol-cloud interactions (ACI; Bellouin et al. 2020). Aerosols directly absorb and scatter incoming solar radiation, causing changes to the energy budget at Earth’s surface. Some aerosols, such as sulfate, can further serve as cloud condensation nuclei (CNN), affecting cloud albedo (first indirect or Twomey effect; Twomey 1977) and cloud lifetime (second indirect effect; Albrecht 1989). The relatively short lifetimes of aerosols, along with a shift in the main emission regions from North America and Europe to Southeast Asia, have led to spatial and temporal heterogeneous aerosol distributions, with their effects being most dominant over the main emission regions and downstream thereof (e.g., Undorf et al. 2018). Thus, regional...
aerosol-induced cooling has been associated with shifts in large-scale circulation patterns resulting in, e.g., suppressed summer monsoon precipitation in West Africa, the Sahel, and Asia (e.g., Bollasina et al. 2011; Polson et al. 2014).

The most recent IPCC report (IPCC 2021a) assesses “the likely range of total human-caused global surface temperature increase” since the mid-twentieth century as 0.8°C to 1.3°C where “well-mixed GHG contributed a warming of 1.0°C to 2.0°C” while “other human drivers (principally aerosols) contributed a cooling of 0.0°C to 0.8°C” (IPCC 2021b). This shows that there is more confidence in the influence of combined human emissions to the observed warming, than in separating the effects of GHG and aerosols which causes much larger uncertainties. To disentangle the contributions to an observed change of the two different climate drivers from those of internal climate variability, detection and attribution methods are used. These methods commonly assume that observations are comprised of a linear combination of model responses to different forcings, usually referred to as fingerprints plus noise. Using optimized linear regression, as introduced by Hasselmann (1993) and further developed by Hegerl et al. (1997), Allen and Tett (1999), Allen and Stott (2003), and Ribes et al. (2013), the combination best fitting the observations can then be derived by estimating scaling factors for the individual fingerprints. Finally, these estimated magnitudes from observations can be used to constrain future impacts.

Constrained attributable GHG warming (e.g., Bindoff et al. 2013; Gillett et al. 2021) has been estimated by averaging across response patterns from multiple models (multimodel mean; MMM), which has been found to provide a more robust assessment by smoothing model errors and biases (Hegerl and Zwiers 2011). However, while a range of studies have detected the contribution of GHG, the response to non-GHG anthropogenic forcings (predominantly aerosols) is not as robust, and results using time-space patterns from different climate model simulations yield quite different results compared to each other, and the multi-model mean (e.g., Jones et al. 2013; Gillett et al. 2013; Jones et al. 2016; Schurer et al. 2020). The use of the multi-model mean averages across aerosol uncertainty and is likely to underestimates the true uncertainty Schurer et al. (2018). This is due to a variety of representations of aerosol processes in models (Wilcox et al. 2015) leading to uncertainties in the simulated climate response to historical aerosol forcing in addition to uncertainty in the forcing itself. Studies have shown that these differences are mostly resulting
from varying representations of aerosol-cloud interactions (Wilcox et al. 2015; Zelinka et al. 2014, 2020), that are essential for reconstructing historical global warming (Wang et al. 2021).

In this study, we focus on observed changes where aerosol influences have been previously detected in the literature, and using large-ensembles from the 6th version of the Climate Model Intercomparison Project (CMIP6; Eyring et al. 2016), we investigate whether combining time series of temperature and precipitation changes can improve the detection of and constraint on aerosol contributions. Recent studies, by Bonfils et al. (2020) looking at a joint change in temperature, precipitation, and continental aridity, and Schurer et al. (2018) investigating the transient climate response by merging various temperature indices, show that adding physically motivated spatial and temporal information can improve results of a detection and attribution analysis, and improve the constraint on individual forcing contributions. Based on the success of these previous studies, we apply scaling factors derived for joint changes in temperature and precipitation to estimate aerosol contributions to the observed warming since the mid-twentieth century.

The remainder of this paper is organised in the following order. Section 2 describes the framework of detection and attribution methods applied in this study, our selection of fingerprints, and the observed and modeled temperature and precipitation data. Section 3 is split into two parts. First, using the CMIP6 large-ensembles a hydrological sensitivity study, and a perfect and imperfect model study are conducted for both individual models and the multimodel mean to test our methodology. In the second part of Section 3, we estimate scaling factors for single and joint fingerprints of the response to aerosols and other forcings and apply them to estimate aerosol contributions to observed global warming. The paper concludes with a discussion of our approach and results in section 4.

2. Methods and data

Detection and attribution methods

We use optimal detection methods based on regularised optimal fingerprinting (Ribes et al. 2013), following the total least squares (TLS) methodology described in Allen and Stott (2003) to detect and attribute long-term observed changes (implemented in Kirchmeier-Young et al. 2017; Gillett et al. 2021). Here, model-derived climate response patterns (fingerprints) to a set of $m$ forcings $X_i$ are regressed against observations $y$ to retrieve scaling factors $\beta_i$ giving the magnitude
of each fingerprint in the observations. The TLS model can then be described as a set of three equations (Ribes et al. 2013):

\[ y = y^* + \epsilon_y, \]  
\[ X_i = X_i^* + \epsilon_i, \text{ and} \]  
\[ y^* = \sum_{i=1}^{m} \beta_i X_i^* \]

where \( y^* \) represents the true climate response to all acting forcings, \( X_i^* \) the noise-free climate model response to forcing \( i \), and \( \epsilon_y \) describes the internal climate variability. \( \epsilon_i \) denotes the noise on \( X_i \) and represents both internal variability and the finite size of the ensemble used to calculate the fingerprints. However, note that \( \epsilon_i \) does not account for climate model uncertainty. We evaluate the latter by comparing results across large-ensembles from different climate models. Including \( \epsilon_i \) differentiates TLS from an ordinary least squares approach. Scaling factors \( \beta_i \) are estimated following Van Huffel and Vandewalle (1991) and the response of the forcing \( i \) is assumed detected if the estimate of its scaling factor is significantly larger than zero. If \( 0 < \beta_i < 1 \) given uncertainty then the climate model significantly overestimates the observed change and needs to be scaled down to be consistent with observations, whereas if \( \beta_i > 1 \) then the climate model underestimates the observations. To sample climate variability, unforced anomalies from control time series from the combined set of available CMIP6 piControl simulations, in which all anthropogenic forcings are kept constant at pre-industrial levels, are randomly selected. Of those, half are used to estimate the covariance matrix of internal variability \( \epsilon \) for optimization. The remaining half is used to estimate scaling factor uncertainties and to conduct a residual consistency test, which evaluates if the observational residuals are consistent by estimating uncertainties of internal variability (Kirchmeier-Young et al. 2017). It has to be noted, that recent studies, e.g., DelSole et al. (2019) and Li et al. (2021), have found that scaling factor uncertainty estimates obtained via optimal fingerprinting can be underestimated leading to coverage rates below nominal levels, especially for small and noisy signals. Alternative approaches have, therefore, been proposed, such as based on bootstrapping (DelSole et al. 2019) and estimating equations (Ma et al. 2023). In our approach, the
joint use of temperature and precipitation fingerprints should improve the signal-to-noise ratios, but the potential for under-coverage has to still be considered when interpreting our results.

To identify the climate response to individual forcings, simulations from the Detection and Attribution Model Intercomparison Project (DAMIP; Gillett et al. 2016) are used, which is a component of the 6th phase of the Coupled Model Intercomparison Project (CMIP6; Eyring et al. 2016). These contain historical (1850-2020) simulations of the model response to individual anthropogenic and natural forcings, such as aerosols-only (AER), along with model runs forced with all external forcings (ALL; 1850-2014) from a variety of Earth System Models. To distinguish the role of aerosols we need to derive a fingerprint containing the signal for all forcings except aerosols, which is primarily driven by GHG. Assuming a linear combination of responses to forcings (Shiogama et al. 2013), we derive this fingerprint by subtracting the AER from the ALL-forcing simulations:

\[ X_{\text{noAER}} = X_{\text{ALL}} - X_{\text{AER}} \]  

(4)

where \( X_i \) represents the fingerprint for forcing \( i \). Then, in a 2-signal analysis, both the noAER and AER fingerprints are simultaneously regressed onto the observations to estimate the role of aerosols. Gillett et al. (2021) found that following this approach where a noAER fingerprint is derived, which is mostly driven by the stronger GHG signal, provides more robust results than estimating the weaker aerosol contribution from the fingerprint with lower signal-to-noise ratio (i.e., noGHG; Ribes et al. 2015).

In the joint study, we conduct a single detection and attribution analysis on the joint fingerprint. Joint fingerprints are obtained by concatenating the single-variable fingerprints along the time axis. This means the joint fingerprint \([A,B]\) of the single variables A and B is given by concatenating the time series of A and B where \( A \in \mathbb{R}^n \), \( B \in \mathbb{R}^m \) and \( A-B \in \mathbb{R}^{n+m} \). By combining the fingerprints in a single scaling factor, we get the benefit of bringing two variables together, which are closely physically related. Each of the individual climate variables is affected by noise much of which is independent and will, therefore, be reduced in the joint study.

Finally, to evaluate the skill of the joint study, we reconstruct AER and noAER contributions to increases in global mean temperature (GMT) and compare these estimates to results from recent literature, i.e., Gillett et al. (2021). Forcing contributions are reconstructed by first scaling (multi)model means of AER and noAER GMT anomalies \( X_i \) (relative to 1961-1990) using scaling
factors $\beta_i$ from the joint analysis:

$$\text{GMT}_{\text{AER}} = \beta_{\text{AER}} \cdot X_{\text{AER}}, \text{ and}$$
$$\text{GMT}_{\text{noAER}} = \beta_{\text{noAER}} \cdot X_{\text{noAER}}$$

(5)

where $X_i$ covers the period 1850-2019. Then, forcing contributions are estimated as follows:

$$\text{AER} = \text{GMT}_{\text{AER}}^{2010-2019} - \text{GMT}_{\text{AER}}^{1850-1900}, \text{ and}$$
$$\text{noAER} = \text{GMT}_{\text{noAER}}^{2010-2019} - \text{GMT}_{\text{noAER}}^{1850-1900}$$

(6)

Uncertainties in the attributed warming arise from both the uncertainty in scaling factor estimates and the internal variability of the simulated GMT for each forcing. Thus, confidence intervals are calculated by adding uncertainties of the scaling factors and the internal variability in quadrature. Estimates of internal variability are obtained by subtracting the model mean from the ensemble and multiplying the temperature anomalies by $\sqrt{\frac{N}{N-1}}$ where N is the number of ensemble members.

To maintain the skewness of the scaling factor distribution, we estimate positive and negative deviations from the best estimate separately, as described in Gillett et al. (2021).

Choice of fingerprints

To separate aerosol from GHG influences, we choose diagnostics that have been physically motivated in the literature to differentiate between the two. These are based on both the individual climate impacts these two forcings have and the spatially and temporally unique emission, and response pathways they started to follow in the late twentieth century. While GHG concentrations are globally homogeneously distributed and have continued to increase globally since the 1850s, aerosol emissions show unique spatial concentration patterns due to their relatively short lifetimes in the atmosphere and shifts in their macro emission regions. Aerosol emissions started to decline in Europe and the US following the implementation of national air quality legislation in the 1980s. At the same time, emissions in Asia experienced a rapid increase as a result of economic development leading to a shift of emissions from Europe/North America to South East Asia (Bellouin et al. 2020; Hoesly et al. 2018; McDuffie et al. 2020). Nevertheless, simulated responses to aerosol forcings vary strongly between models (Wilcox et al. 2015), which leads to differences in fingerprints for
AER from different models and thus model sensitivity of detection and attribution results. Finally, as aerosols impact both temperature and precipitation through direct and indirect effects, with aerosol responses expected to be substantial (Allen and Ingram 2002), we are considering both temperature- and precipitation-based variables to investigate aerosol impacts.

Fig. 1. Time series of standardised historical all forcing (ALL; green), aerosol-only (AER; blue), and historical all minus aerosol forcing (noAER; red) multimodel mean annual anomalies, along with standardised observed anomalies (solid black) of a) the global mean temperature (GMT; HadCRUT5), b) the interhemispheric temperature asymmetry (ITA; HadCRUT5) and c) global mean land precipitation (GMLP; CRUTS) for the period 1955-2020, and d) tropical wet- and e) dry-region precipitation (GPCPv3.2) for 1988-2020. The noAER signal is derived by subtracting the AER-only from the ALL-forcing signal. Faint black dotted lines show the $2\sigma$ range of a set of 500 anomalies from the respective piControl simulations, which are randomly sampled from the CMIP6 model cohort. The same set of piControl simulations is used to standardise the displayed observational and model time series. Anomalies are relative to the entire analysis period (1955-2020 for GMT, ITA, and GMLP; 1988-2020 for wet and dry) and a 5-year running mean is applied for better visualisation. Numbers in brackets show the number of models used to calculate the multimodel mean. (Single-)Forcing fingerprints (AER and noAER) are shown in dashed lines while the ALL fingerprint and observations are plotted in solid lines.

The first fingerprint we investigate is the GMT due to its large signal-to-noise ratio (see Figure 1a) which has been shown to work well to detect anthropogenic climate change signals, but shows large uncertainties when trying to use it alone to distinguish between aerosol and GHG impacts (Wilcox
et al. 2013; Schurer et al. 2018; Gillett et al. 2021). Comparing the climate response to AER and noAER forcing in Figure 1a shows that noAER (largely GHG) warming has likely been offset to some extent by AER-induced cooling. The strongest aerosol signal is found until the 1980s when global aerosol emissions were still increasing.

Aerosol emissions being predominantly located in the Northern Hemisphere (NH) have induced a change in the interhemispheric temperature asymmetry (ITA; Friedman et al. 2013; Wilcox et al. 2013; Schurer et al. 2018) defined as the difference in temperature between NH and Southern Hemisphere (SH). Model studies suggest that until the 1980s the ITA response to aerosol and GHG forcing offset each other, then, GHG effects start to dominate due to declining global aerosol emissions (Friedman et al. 2013). This shift is represented by a sharp increase of the ITA due to a disproportional GHG-warming of the Arctic and NH landmasses which had previously been compensated by aerosol-induced NH cooling (Wilcox et al. 2013; Friedman et al. 2013, 2020; Shindell et al. 2015). The change in the trend of the ITA in the CMIP6 large-ensembles and observations (Figure 1b) reflects aerosols’ emission history clearly showing a reduction in the ITA in the AER and ALL signals and the observations until the mid-1980s followed by a sharp increase. The spread between CMIP6 models is larger for the ITA than GMT, highlighting model uncertainties in simulating (regional) aerosol impacts (Wilcox et al. 2015).

Changes in global mean temperatures are considered to drive thermodynamic changes in the hydrological cycle described through the Clausius-Clapeyron relationship (Allen and Ingram 2002; Hegerl et al. 2015) and dynamic changes by shifting large-scale atmospheric circulation, such as the southward shift of the Intertropical Convergence Zone (ITCZ; Schneider et al. 2014). Opposing the increase in precipitation associated with GHG warming, aerosol cooling has, therefore, been found to result in a reduction of evaporation and precipitation (Ming and Ramaswamy 2009; Bony et al. 2013; Polson et al. 2013b), especially over land where sulfate particles both cool the surface and serve as CNN (Ramanathan et al. 2001; Zhang et al. 2021). Given that instrumental records on land go back further than the satellite era (1980s) we investigate global mean land precipitation (GMLP) as a diagnostic. ALL, noAER, and AER signals for the GMLP are shown in Figure 1c and highlight an aerosol-induced drying until around 1990 and GHG-related wetting. Again the ALL fingerprint and observation suggest that both forcings worked against each other in the past.
In the tropics, the intensification of the water cycle due to global warming has led to a ‘wet-get-wetter, dry-get-drier’ pattern when tracking the motion of wet and dry regions (e.g., Held and Soden 2006; Polson et al. 2013a). Schurer et al. (2020) observed that the opposite is the case for aerosols. By cooling the climate system, models suggest a drying of the wet regions and a wetting of the dry regions associated with increasing global aerosol emissions. Thus, following a method developed by Polson et al. (2013a) and Polson and Hegerl (2017) for each monthly precipitation field we rank tropical (30°S-30°N) grid boxes according to their absolute monthly total precipitation in ascending order. The upper 30% are designated as wet regions and the bottom 30% as dry regions. Time series of average monthly precipitation from area-weighted means of the respective regions are shown in Figure 1d and e and show clear changes in wet regions, while the impact from AER and noAER is less clear for dry regions. As changes in both regions are linked through the transport of water we will treat tropical wet and dry regions as a single fingerprint ‘wet dry’ (WD), where the individual time series are concatenated along the time axis to form a single fingerprint (Schurer et al. 2020).

Other potential climate variables were investigated before choosing those mentioned above, such as the diurnal temperature range (DTR; Stjern et al. 2020) defined as the difference between daily maximum (Tmax) and daily minimum temperature (Tmin). Due to GHG and aerosols’ unique impact on radiation, the DTR has been suggested as a diagnostic to separate the response to the two forcings (e.g., Wild et al. 2007; Stjern et al. 2020; Undorf et al. 2018). However, as we found large differences between models and a low signal-to-noise ratio, we decided against including DTR as a potential fingerprint in this study. Referring to shifts in aerosol macroemission regions from the Western to the Eastern Hemisphere, we also conducted a meridional analysis of aerosol impacts on ITA and tropical wet-dry regions but our investigation did not return a robust spatial signal of the shift to more south-eastern emission regions in either temperature or precipitation reemphasizing the impact of model uncertainty.

**Observations and climate model data**

Full coverage of the tropics is required to track wet and dry regions, thus, following Schurer et al. (2020) and the IPCC (2021a), the satellite-gauge merged data set of monthly precipitation data from the Global Precipitation Climatology Project v3.2 (GPCPv3.2; Huffman et al. 2023)
is used. Although observations are available from January 1979, the analysis is limited to the period January 1988 to December 2020, where measurements from the Special Sensor Microwave Imager are available (following Polson and Hegerl 2017). A sensitivity test and a comparison of different data sets for tropical wet and dry regions can be found in Schurer et al. (2020).

To calculate GMLP for the period 1955-2020, the Climate Research Unit’s gridded Time Series (CRUTSv4.05; Harris et al. 2020) data set, for which monthly observations are available starting in January 1901, is used. Other observational precipitation data sets are available, which differ in their distribution of station input, homogenization, area averaging, and quality control sampling. However, it has been argued that the homogeneous coverage of CRUTS provides a more reliable time series (Lorenz and Kunstmann 2012). Finally, observed changes in the ITA and GMT are calculated from monthly temperature anomalies in the fifth Hadley Centre/Climatic Research Unit Temperature (HadCRUT5; Morice et al. 2021) non-infilled data set, where data is available from January 1850 onwards.

We limit our detailed analysis to models with at least 6 ensemble members, i.e., large-ensembles. Thus, large-ensembles of three CMIP6 models are analysed (CanESM5, CNRM-CM6-1, IPSL-CM6A-LR) for monthly near-surface air temperature and precipitation for which CMIP6 historical, Scenario Model Intercomparison Project (ScenarioMIP) SSP2-4.5 and Detection Attribution Model Intercomparison Project (DAMIP) hist-aer simulations are available. As SSP2-4.5 forcings are used in DAMIP simulations for the period 2015-2020, we merge CMIP6 historical simulations with SSP2-4.5 simulations for 2015-2020 to be able to extend the analysis to 2020. Note that because HadCRUT5 contains temperature anomalies relative to 1961-1990, for consistency we take monthly anomalies of each grid box in reference to 1961-1990 in the temperature simulations.

**Table 1.** Fingerprints investigated in this analysis. Variable names with the respective abbreviations, the dataset used to derive the time series, the analysis periods, and the observational mask applied to calculate the time series are listed.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Data set</th>
<th>Period</th>
<th>Mask</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wet Dry</td>
<td>WD</td>
<td>GPCPv3.2</td>
<td>1988-2020</td>
<td>-</td>
</tr>
<tr>
<td>Global Mean Land Precipitation</td>
<td>GMLP</td>
<td>CRUTS</td>
<td>1955-2020</td>
<td>CRUTS seasonal mask</td>
</tr>
<tr>
<td>Global Mean Temperature</td>
<td>GMT</td>
<td>HadCRUT5</td>
<td>1955-2020</td>
<td>HadCRUT5 seasonal mask</td>
</tr>
<tr>
<td>Interhemispheric Temperature Asymmetry</td>
<td>ITA</td>
<td>HadCRUT5</td>
<td>1955-2020</td>
<td>HadCRUT5 seasonal mask</td>
</tr>
</tbody>
</table>
Table 2. Models used in this analysis. CMIP6 models are selected from the CMIP6 cohort if CMIP6 historical, SSP2-4.5 and DAMIP single forcing simulation for temperature and precipitation are available. Aerosol schemes are separated into specified (externally simulated fields of aerosol optical depth) and prognostic (propagating physical aerosol properties) schemes. Numbers marked # show the number of ensemble members included in each ensemble. The horizontal divides highlight the separation between the CMIP6 large-ensembles used in Section 3a (top), and the remaining CMIP6 models with at least 3 ensemble members analysed as an additional reference fingerprint in Section 3b (bottom). Values for the equilibrium climate sensitivity (ECS) are taken from Gettelman et al. (2019) and Meehl et al. (2020), and for aerosol effective radiative forcing of the net aerosol effect \( \text{ERF}_{\text{AERnet}} = \text{ERF}_{\text{ACI}} + \text{ERF}_{\text{ARI}} \) from Zelinka et al. (2014) and Smith et al. (2020). IPCC estimates a very likely range of 2 to 5 \( ^\circ \text{C} \) for ECS (Forster et al. 2021). (*) Aerosol ERF was calculated using a prognostic aerosol scheme. While all models include the first indirect effect, Twomey effect (Twomey 1977), models also simulating the second indirect effect are marked with (**).

<table>
<thead>
<tr>
<th>Model name</th>
<th>#</th>
<th>Scheme</th>
<th>ECS [(^{\circ}\text{C})]</th>
<th>ERF(_{\text{AERnet}}) [Wm(^{-2})]</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>CanESM5**</td>
<td>30</td>
<td>Prognostic</td>
<td>5.6</td>
<td>-1.06</td>
<td>Swart et al. (2019); Cole et al. (2023)</td>
</tr>
<tr>
<td>CNRM-CM6-1</td>
<td>10</td>
<td>Specified</td>
<td>4.8</td>
<td>-1.43*</td>
<td>Voldoire et al. (2019); Roehrig et al. (2020)</td>
</tr>
<tr>
<td>IPSL-CM6A-LR</td>
<td>6</td>
<td>Specified</td>
<td>4.6</td>
<td>0.65</td>
<td>Boucher et al. (2020); Lurton et al. (2020)</td>
</tr>
<tr>
<td>MIROC6**</td>
<td>3</td>
<td>Prognostic</td>
<td>2.6</td>
<td>-0.94</td>
<td>Tatebe et al. (2019)</td>
</tr>
<tr>
<td>HadGEM3-GC31-LL**</td>
<td>4</td>
<td>Prognostic</td>
<td>5.6</td>
<td>-1.11</td>
<td>Williams et al. (2018); Walters et al. (2019)</td>
</tr>
<tr>
<td>MRI-ESM2-0**</td>
<td>3</td>
<td>Prognostic</td>
<td>3.2</td>
<td>-1.16</td>
<td>Yukimoto et al. (2019)</td>
</tr>
</tbody>
</table>

This is done before GMT and ITA fingerprints are derived. While testing of the methodology in Section 3a is purely based on these large-ensembles, in order to compare the large-ensembles to the remaining CMIP6 model ensemble, we additionally download model ensembles for which at least 3 historical, historical SSP2-4.5 and DAMIP simulations are available (see Table 2). The multimodel mean of the in total six CMIP6 models will be analysed as a reference fingerprint, referred to as ‘CMIP6 MMM’, in the observational analysis in Section 3b.

For consistency, all observational datasets and the model data are regridded to a 2.5° x 2.5° grid. For precipitation (i.e., CRUTS, GPCPv3.2, and precipitation model simulations) a conservative regridding method is chosen (preservation of the integral of the precipitation field; Jones 1999), while temperature fields (HadCRUT5 and temperature model simulations) are bi-linearly regrid-ded, i.e., linear interpolation in two directions using the 4 nearest grid boxes. Because CRUTS and HadCRUT5 only contain values for grid points where observations are available, fewer mea-
measurements exist for earlier decades. To compensate for this, we derive observational masks from the regridded observational datasets to only use grid cells for which at least one observation in each season in each year (i.e., 4 per year) is provided over the period 1955-2020. The CRUTS and HadCRUT5 seasonal masks are then applied to both the respective observations and model data to calculate the fingerprints (see Table 1). Then, 3-year means of the fingerprints are taken, and both the observations and the model simulations (including the control samples) are standardised. We standardise the time series by dividing the fingerprints by the average standard deviation of the 500 control simulations samples from the full CMIP6 model ensemble for each respective climate variable. These are the same control simulations as used during the detection and attribution analysis (see methods in Section 2). The average standard deviation is obtained by taking the mean of the 500 standard deviations of each individual control simulation. Standardising the time series accounts for the relative difference in magnitude of the change in the different temperature- and precipitation-based fingerprints and this allows us to avoid the larger magnitude of, for example, GMT, dominating the analysis (Schurer et al. 2020). In the final step, the fingerprints are centered, i.e., anomalies are taken over the whole analysis period.

**Aerosol forcing in CMIP6**

Aerosol schemes have been updated in this newest generation of climate models compared to previous versions. This concerns especially aerosol microphysics and aerosol-cloud interactions with all models in this study containing direct effects and at least the first indirect effect, i.e., cloud albedo effect or Twomey effect (Twomey 1977). Studies have found that CMIP5 models that simulate some form of the Twomey effect reproduce representations of interdecadal variability in historical GMT, land-only precipitation, and zonal temperature trends better than models that only include the direct effect (Wilcox et al. 2013; Ekman 2014). However, radiative feedbacks from clouds, especially aerosol-cloud interactions, have also been associated with causing a large spread in predictions of future global warming (Boucher et al. 2013; Zelinka et al. 2020; Smith et al. 2020). Stronger cloud feedbacks have further been linked to causing an increase in spread and magnitude of Equilibrium Climate Sensitivity (a standard metric to measure model sensitivity to CO$_2$ increase; ECS) among CMIP6 models, although some disagreement remains on the correlation between effective radiative forcing from aerosols (ERF$_{AERnet}$) and ECS in CMIP6 models (Meehl et al.
Similarly, no significant association between stronger aerosol forcing and higher ECS is found for CMIP5 models when considering the range of models (Forster et al. 2013; Chylek et al. 2016). However, this is different when only considering CMIP5 models that include aerosol indirect effects (Chylek et al. 2016). While models of both high and low ECS are able to simulate the observed evolution of GMT, inconsistencies arise when investigating the representation of observed interhemispheric temperature records, where high ECS models produce a strong ACI cooling in the NH (Wang et al. 2021). Nevertheless, the models considered in this study have been found to be consistent with recent observational estimates of aerosol effective radiative forcing ($\text{ERF}_{\text{AER_{net}}};$ -2.0 to -0.4 Wm$^{-2}$ as a 90% confidence interval; Bellouin et al. 2020).

Further differences between models arise from varying aerosol schemes with some models having specified schemes, where 3D aerosol fields are simulated externally before being used as aerosol forcing in the CMIP6 simulations, while others include prognostic aerosol schemes propagating physical aerosol properties within the model (see Table 2). These differences between aerosol schemes have shown to have the highest impacts in regions remote from aerosol emission sources due to differing dynamical responses partially driven by deviating aerosol optical depths especially over oceans. This is likely due to relative humidity-related aerosol swelling in prognostic models leading to increased AOD and additional cooling (Randles et al. 2013).

3. Results and discussion

This result section is separated into two subsections using different methodologies and data. A schematic describing the content of section 3 is displayed in Figure 2. In the first subsection (section 3a) we test our approach of joining temperature and precipitation fingerprints using the CMIP6 large-ensemble model simulations. The method validation is conducted in 4 steps, analysing the models’ temperature and precipitation sensitivities to external forcings and the ability of our choice of fingerprints to separate AER and noAER contributions in model data (“imperfect model study”). In the remainder of this section (section 3b), we include observations and the CMIP6 MMM reference fingerprint to detect and attribute AER and noAER influences on observed climate changes. Finally, to evaluate the implications of combining temperature and precipitation fingerprints, we plot the [2010-2019] minus [1850-1900] temperature anomalies of scaled model
Fig. 2. Schematic describing the data and methods used, and the objective of the different analyses conducted in section 3. \( y \) and \( X \) describe the time series serving as dependent and independent variables in the TLS analyses and the linear fit (only for hydrological sensitivity study) and \( \beta \) stands for the scaling factors estimated in the respective analyses. Note that, green \( \beta \) highlight analyses where single variable fingerprints are used, and purple \( \beta \) show where a joint study is conducted. Blue boxes highlight model simulations and dark green represents the application of observational data.

GMT fingerprints for AER and noAER. The fingerprints are scaled using scaling factors from the joint study.

**a. Testing of the methodology using CMIP6 large-ensembles**

Should we combine temperature and precipitation fingerprints?

If the sensitivity to an external forcing in the model is incorrect and the model fingerprints need to be scaled to match the observations then we are assuming that both the temperature and precipitation fingerprints need to be scaled by the same amount. Several reasons exist why this is not always true, such as the model having the incorrect sensitivity to external forcings, a wrong forcing (which can also be due to the observations serving as forcing input), or simulating the
Fig. 3. Anomalies of global mean precipitation (relative to climatology) plotted against anomalies of the model mean global mean temperature change for IPSL-CM6A-LR (purple), CanESM5 (orange) and CNRM-CM6-1 (green), both smoothed by a running 3-year average. These displayed changes (% K$^{-1}$) are calculated for the period 1988-2020, with brackets showing the change over the longer period, 1955-2020. Anomalies are taken over 1988-2020 and 1955-2020, respectively. Change predicted by the Clausius-Clapeyron relationship and energy constraint (2%K$^{-1}$) is shown as a dotted line with the grey shaded area depicting the observed range of hydrological sensitivity (2$\pm$0.5 %K$^{-1}$; Allan et al. 2020).

Global changes in precipitation are linked to changes in temperature referred to as hydrological sensitivity (HS) which is defined as the relative change in global mean precipitation (GMP) as a function of degree change in global mean temperature [%K$^{-1}$]. The HS is constrained by the energetics of the atmosphere to 2 %K$^{-1}$ (e.g., Allen and Ingram 2002; Hegerl et al. 2015).

Using similar observational datasets to those used in this study, Allan et al. (2020) derive an observational HS of 2 $\pm$ 0.5 %K$^{-1}$ for the period 1979-2019, while model-based studies (e.g., Norris et al. 2022; Fläshner et al. 2016) have found a wide range of HS values with large scatter arising from the choice of the studied period and the chosen dataset. Norris et al. (2022) calculate...
a model spread in HS of 1.1-2.2 °C using 150 years of CMIP6 1pctCO2 simulations (increase of CO2 by 1 %yr\(^{-1}\) from preindustrial levels), while comparing CMIP6 abrupt carbon dioxide quadrupling (abrupt-4xCO2) experiments to piControl simulations Pendergrass (2020) find a range of HS of 2.1-3.1 °C in the CMIP6 ensemble. Using historical ALL simulations we estimate a range of best estimate HS values for the models used here of 1.5-1.71 °C for the period 1988-2020 and slightly lower values ranging from 1.21-1.38 °C when extending the period to 1955-2020 (see Figure 3). Lower HS values could be due to the stronger aerosol forcing in the longer period, which overcompensates the GHG-related precipitation increase potentially due to aerosol-cloud interactions, thus decreasing the HS. Previous studies have noted that aerosols might impact precipitation more than GHG (e.g., Allen and Ingram 2002; Hegerl et al. 2015). Thus, our model results (1988-2020) lie in the lower bounds of the observed HS (1979-2019) and align with model-based studies although we find lower values also likely due to historical aerosol forcing. As our models lie in the uncertainty range of the observed HS, we assume that the fingerprints can be combined.

A PERFECT MODEL STUDY OF OUR CHOICE OF FINGERPRINTS

To evaluate our choice of fingerprints in their ability to detect aerosol influences, a perfect model study is conducted, i.e., analyse the signal-to-noise ratio of our choice of single-variable fingerprints (Ribes and Terray 2013; Schurer et al. 2018; Gillett et al. 2021). Here, one of the ALL ensemble members is used as a pseudo-observation and all other ALL and AER members from the same model are used to derive fingerprints to determine scaling factors \(\beta_{\text{ALL}}\) for a 1-signal, and \(\beta_{\text{noAER}}\) and \(\beta_{\text{AER}}\) for a 2-signal study. This process is repeated for all historical model simulations for all models. By definition, this analysis should give a 5-95 % confidence interval where on average 1 is included in the scaling factor range in 90 % of cases allowing us to evaluate variables in their ability to reconstruct observed changes as well as separate noAER and AER contributions. To rate the individual variables, we calculate the percentage where the scaling factors deviate from 1 (fail to include 1) referred to as deviation rate. Note that cases where uncertainties cannot be constrained, i.e., infinite confidence intervals, are also assumed to include 1.

Results from the perfect model study are shown in Figure 4. Across variables, we find the lowest uncertainties for \(\beta_{\text{ALL}}\), while separating the noAER (\(\beta_{\text{noAER}}\)) and AER (\(\beta_{\text{AER}}\) ) contribution
leads to less constrained scaling factors with the largest uncertainties observed for $\beta_{AER}$. We find the smallest deviation rates for GMT and GMLP depending on the model. However, our analysis shows that across models GMT is the best single diagnostic to attempt to separate AER and noAER contributions, as it leads to the tightest constraint. There is more uncertainty and variability between models in separating AER and noAER contributions for WD and ITA likely due to varying aerosol schemes between models. Among the three large-ensembles, CanESM5 is the only model with a prognostic aerosol scheme, simulating a strong cooling induced by aerosol impacts on clouds in the Northern Hemisphere which results in a clear AER trend in the ITA, GMT,

![Image](image_url)

**Fig. 4.** Results of perfect model study: Estimates of $\beta_{ALL}$ (1-signal analysis), $\beta_{AER}$, and $\beta_{noAER}$ (2-signal analysis) for single-variable fingerprints (a) WD, (b) GMLP, (c) ITA and (d) GMT. Colours show individual models as pseudo-observations. Crosses depict cases where scaling factors cannot be constrained. Numbers indicate the percentage of scaling factors where the 5%-95% uncertainty range does not include 1 (unconstrained cases are considered to include 1). Note that this percentage is statistically expected to be 10%. The linestyle of the whiskers indicates P values from a residual consistency test. Dashed whiskers show P values of less than 0.05, indicating residuals that are significantly larger than expected. Dotted whiskers represent P values greater than 0.95 indicating residuals significantly smaller than expected based on the simulated internal variability and solid lines P values between 0.05 and 0.95. The x-axis shows results for each ensemble member in the respective model, ordered by scaling factors for visual clarity.
and GMLP while less of an aerosol signal is visible for CNRM-CM6-1 and IPSL-CM6A-LR (see Figure S1). The weaker aerosol signal might, therefore, be due to the different aerosol schemes, although the reduced ensemble size of CNRM-CM6-1 and IPSL-CM6A-LR (10 and 6 respectively) compared to CanESM5 (30) could also be a factor. While CNRM-CM6-1 has a lower deviation rate when separating noAER and AER impacts for WD than ITA in the perfect model study, the opposite is the case for CanESM5. Note that the unsuccessful attribution of the WD response to noAER and AER forcing in IPSL-CM6A-LR is likely due to a low signal-to-noise ratio (compare Figure S1). Although ideally, we would expect to find deviation rates of around 10% (i.e., 90% coverage of the confidence intervals) this is not the case for all models, especially CNRM-CM6-1 when separating AER and noAER contributions (20-30% in Figure 4). Studies mentioned above (e.g., Li et al. 2021) have noted this problem before. This could be a combination of underestimated uncertainty ranges when deriving them using TLS when the signal-to-noise ratio is weak, such as in the precipitation fingerprints (DelSole et al. 2019; Li et al. 2021; Ma et al. 2023), and because we sample control simulations from the range of CMIP6 models. Therefore, scaling factors of models with larger internal variability than the sampled control variability (CNRM-CM6-1 and IPSL-CM6A-LR; Parsons et al. 2020) might return overconfident results (Schurer et al. 2018), and reversed.

An imperfect model study using single variables

An imperfect model study further highlights the varying model sensitivities to external forcings (Figure 5; note that colours depict the model from which pseudo-observations are taken) as it differs from a perfect model study by using one of the ALL simulations from one model as pseudo-observations while fingerprints are derived as the multimodel mean of the remaining models. Thus, we compare model responses to each other, and scaling factors in the imperfect model framework, therefore, account for model differences, including in the aerosol response, and highlight the varying sensitivities to external forcings between the different models (Zelinka et al. 2020). For this reason, we also do not expect to find deviation rates of 10%, i.e., the scaling factors should be different to 1, and uncertainty in the aerosol forcing itself and model uncertainty contribute to the scaling factor uncertainties (e.g., Wilcox et al. 2015). As for the perfect model study, we find the lowest uncertainties for GMT across all models where among the different scaling factors
Fig. 5. Results from imperfect model study: Similar to Figure 4, but showing results estimating signal magnitudes using data from different models as pseudo-observations (note that consistency with “1” is not expected where model sensitivities to external forcings are different). Estimates of $\beta_{\text{ALL}}$ (1-signal analysis), $\beta_{\text{AER}}$, and $\beta_{\text{noAER}}$ (2-signal analysis) from an imperfect model study for single-variable fingerprints (a) WD, (b) GMLP, (c) ITA and (d) GMT. Colours indicate individual models used as pseudo-observations. Remaining notations are the same as for Figure 4.

$\beta_{\text{ALL}}$ is constrained the most and $\beta_{\text{AER}}$ the least. As expected, the range of scaling factors widens to account for model differences (see Figure 5). Overall, consistency in scaling factors for GMT, ITA, and GMLP across models indicates that the models show a comparable response to external forcings in both temperature and precipitation, further supporting our approach of combining the two fields in a joint study. As for the perfect model study, we observe an under-coverage of the estimated confidence intervals (e.g., Li et al. 2021).
An imperfect model study of the joint fingerprints

An imperfect model study of combined fingerprints evaluates the reliability of our approach and its sensitivity to model differences in fingerprints. By combining the time series in a joint study we add physically motivated temporal and spatial information which helps the separation of aerosol cooling from GHG (noAER) warming. We investigate 7 different combinations of diagnostics: two fingerprints combining precipitation ([WD; GMLP]) and temperature ([ITA; GMT]) variables, and fingerprints summarising the global impact ([GMT; GMLP]), dynamical shifts ([ITA; GMLP]) and thermodynamic changes ([WD; GMT]). We then add information to the dynamical fingerprint ([ITA; GMT; GMLP] and [WD; ITA; GMT]), before joining all variables ([WD; ITA; GMT; GMLP]). The joint study is conducted by concatenating the respective fingerprints and conducting a single analysis in the joint fingerprint (see Section 2).

In Figure 6, estimates of the pseudo-observed AER and noAER contributions to changes in GMT ([2010-2019] minus [1850-1900]) from model simulations are shown (y-axis) to diagnose performance in the imperfect model study. Scaled estimates are constructed by multiplying scaling factors from the imperfect model study (see Figure S3) with the respective AER and noAER GMT multimodel mean fingerprints (from 2 models) and then subtracting the 1850-1900 mean from the mean for 2010-2019 (see methods in Section 2). These attributed contributions from the imperfect model study are then compared to the (true) simulated AER and noAER GMT contributions derived from the GMT model mean of the model serving as pseudo-observations (x-axis). Finally, to evaluate the skill of the various diagnostics the percentage agreement between attributed and true contributions is calculated, referred to as success rate. As we find that GMT is the best diagnostic to separate AER and noAER contributions (see Figure 4 and Figure S2d) among the single variables, we are verifying our approach by comparing it to results obtained for GMT-only shown in Figure 6a (same as in Figure S2). Figure 6b-h display attributed contributions for an imperfect model study of the combined diagnostics. Here, the colour of the scatter points depicts the model which serves as pseudo-observation from which the (true) simulated contribution is estimated.

Our assumption of similar model sensitivities to external forcings for both temperature and precipitation seems to hold as we still observe stronger simulated GMT contributions of AER and noAER when CanESM5 is used as pseudo-observations across combined diagnostics (orange circles). We find that our approach shows some skill in improving the attribution of AER con-
Fig. 6. Results of imperfect model study comparing estimated responses from detection and attribution (‘attributed contribution’) with model simulated contributions (contr.): AER (left, negative) and noAER (right, positive) contributions to changes in GMT ([2010-2019] minus [1850-1900] temperature anomalies) reconstructed from $\beta_{AER}$ and $\beta_{noAER}$ estimates for an imperfect model study of the combined diagnostics (y-axis) plotted against the (true) AER and noAER contributions of the GMT model mean of the model serving as pseudo-observation (x-axis). Imperfect model contributions (y-axis) are reconstructed by multiplying scaling factors from 2-signal detection and attribution analyses using the following diagnostics a) GMT, b) [WD; GMLP], c) [ITA; GMT], d) [ITA; GMLP], e) [GMT; GMLP], f) [ITA; GMT; GMLP], g) [WD; ITA; GMT], and h) [WD; ITA; GMT; GMLP] with the respective GMT multimodel mean fingerprints for AER and noAER contributions from the 2 models not serving as pseudo-observations. Note that the results for GMT in a) are the same as in Figure S3d). Colours depict the model serving as pseudo-observations. Numbers are the percentage of cases where the 5%-95% range of the reconstructed contributions includes the (true) simulated contribution for AER and noAER. Crosses show cases where scaling factors cannot be constrained.
tributions. Across models and combined diagnostics the fraction of unconstrained scaling factors decreases compared to individual variables (only a single unconstrained case for [WD; GMLP] and [WD; ITA; GMT; GMLP]). Further, for the majority of combined diagnostics, the percentage of cases including the true simulated value of the AER contribution similar to for GMT on its own (success rate AER ≥ 86%). Although attributed contributions are less constrained in some cases (e.g., [WD; GMLP]), similar uncertainties to GMT-only are observed for AER and noAER contributions when including both global means ([GMT; GMLP] and [ITA; GMT; GMLP]). In general, whether the addition of the wet-dry contrast or the interhemispheric temperature asymmetry to the combined fingerprints increases or decreases the uncertainty of estimated AER and noAER contributions is model-dependent.

However, our study also suggests that combining fields introduces a tendency to overestimate noAER contributions, i.e., estimates lie above the 1 line. As pseudo-observations with higher variability than the fingerprints (and the piControl simulations to calculate the covariance matrix) will have overconfident results, i.e., small confidence intervals, the combination of overestimation and small uncertainty ranges leads to success rates of less than 90% (especially for noAER). Among models in this study, CNRM-CM6-1 and IPSL-CM6A-LR have very high internal variability compared to most of the remaining CMIP6 models and CanESM5 (Parsons et al. 2020). Finally, as mentioned before, model uncertainty and under-coverage of the confidence intervals arising from the TLS analysis can also play a role in observing success rates of less than 90% in an imperfect model study (Schurer et al. 2018; DelSole et al. 2019), particularly where noisier diagnostics are used (for example [WD; GMLP]). However, we note that the signal to noise ratio tends to increase when we combine variables.

b. Detection and attribution of observed climate changes

Regressing the ALL fingerprint onto the observations (1-signal analysis; \( \beta_{\text{ALL}} \)) confirms that external forcings play a role in driving observed changes across variables (\( \beta_{\text{ALL}} > 0 \)). Scaling factors for the 3 individual CMIP6 large-ensemble models are compared in Figure 7, along with their multimodel mean (LE MMM; black). For additional comparison, a larger CMIP6 multimodel mean (CMIP6 MMM; grey) is also derived containing the 6 CMIP6 models listed in Table 2. The respective observational data sets for individual single variables are listed in Table 1. Note that
while in previous plots colours depict the model serving as source for the pseudo-observations, they now indicate the model from which fingerprints are derived.

**Fig. 7.** Detection and attribution analysis from observations, using fingerprints from different climate model simulations (colours) and variables (panels). Vertical bars show the estimated magnitude (90% confidence intervals) of ALL forcing $\beta_{\text{ALL}}$, and the contribution of all-except aerosols $\beta_{\text{noAER}}$ and aerosols $\beta_{\text{AER}}$ to observed changes for a) WD (GPCPv3.2; 1988-2020), b) GMLP (CRUTS; 1955-2020), c) ITA (HadCRUT5; 1955-2020) and d) GMT (HadCRUT5; 1955-2020) from a 1-signal (top row) and 2-signal (bottom 2 rows) total least squares regression analysis. Vertical lines separate scaling factors for individual models and multimodel means (MMM). Use of crosses, dashed, dotted and solid lines same as in Figure 4 and Figure 5.

Comparing $\beta_{\text{ALL}}$ for temperature- and precipitation-based variables, it becomes clear that across models precipitation-based variables show a larger scaling factor, i.e., smaller change compared to the observations, than those for temperature. For CNRM-CM6-1 and IPSL-LM6A-LR, ITA and GMT scale around 1 (Figure 7c and d) and are thus consistent with the observations, while scaling factors for WD are well constrained and lie around 2 (Figure 7a). For GMLP scaling factors are higher (between 2 and 3) and less constrained. CanESM5, which exhibits a stronger response to external forcings (see Figure S1), overestimates the ITA and GMT responses but shows scaling factors around 1 for WD and GMLP. The large-ensemble multimodel mean matches the
mean responses of the 3 individual models and is consistent with the CMIP6 multimodel mean. Uncertainties are higher for CNRM-CM6-1 and IPSL-CM6A-LR, as they are smaller ensembles.

AER and noAER contributions are detectable in a 2-signal analysis for the interhemispheric temperature asymmetry and global mean land precipitation ($\beta_{\text{noAER}}$ and $\beta_{\text{AER}} > 0$; Figure 7) but a residual consistency test shows that for CanESM5 and CNRM-CM6-1 GMLP residuals are larger than the 95th percentile of internal variability estimates for GMLP indicating large unexplained variability, while they are surprisingly small for ITA ($< 5$th percentile). Thus, the scaled noAER and AER fingerprints for GMLP explain less observational variability than expected indicating a possible underestimation of precipitation variability in the sampled control simulations. At the same time, smaller residuals for the ITA suggest a possible overestimation in the hemispheric temperature variability in the control simulations. GMT responses to AER and noAER forcings are attributable to IPSL-CM6A-LR only.

**Why do we observe different scaling factors for temperature and precipitation when using observations?**

By comparing scaling factors for temperature and precipitation in the imperfect model study and from the hydrological sensitivity analysis, we would assume that the analysed models exhibit the same sensitivity in temperature and precipitation to external forcings. However, scaling factors from the 1-signal analysis ($\beta_{\text{ALL}}$) in Figure 7 suggest that temperature and precipitation scale differently. To analyse if this is true or if the differences are consistent with uncertainty, we conduct an observational detection and attribution study for the same 2-model mean fingerprints as in the imperfect model study. The resulting scaling factors for wet-dry (WD; y-axis) are plotted against temperature-based (x-axis) variables for the observational study (black whiskers) and the imperfect model results (coloured whiskers) in Figure 8. Since this difference in scaling factors is especially true for WD, only results for WD are shown here. Scaling factors for GMLP are more consistent with those from GMT and ITA, and are displayed in Figure S6. Comparing scaling factors for temperature and WD, indicates that the models do not exhibit the same response in temperature and precipitation to historical forcings, as results do not fall on the diagonal (dotted line). In general, best estimates of the precipitation-based variables (compare GMLP in Figure S6) are closer to 1, tighter constrained and show more consistency with ITA and GMT when CanESM5 is included in
Fig. 8. Scaling factors $\beta_{\text{ALL}}$ for temperature-based variables (x-axis) plotted against scaling factors $\beta_{\text{ALL}}$ for wet-dry (WD; y-axis). Model fingerprints are taken as the mean of two model simulations, these are regressed against simulations from the third model in the top row (circles, coloured whiskers) and observations in the bottom row (squares, black whiskers). Colours of the coloured whiskers indicate the model used as pseudo observations, thus, not included in the multimodel mean fingerprints. Colour of the scatter points with black whiskers indicates the model missing from the multimodel fingerprints. Scaling factor ranges including the 1:1 line suggest the same model response in temperature and precipitation to external forcings. For the imperfect model study, results using individual simulations as pseudo-observations are plotted in transparent, and the mean of all results are shown in bold. Thin black whiskers highlight scaling factor uncertainties for double variance in the observational studies.

the fingerprint (purple and green scatter points) confirming that a stronger precipitation response to external forcings is needed to reconstruct observed changes. This inconsistency between obser-
vational and imperfect model results raises the question of whether these differences are within the range of internal model variability. Previous studies have found discrepancies between simulated and observed precipitation variability in climate models (e.g., Zhang et al. 2007; Hegerl et al. 2015), particularly in the tropics. Thus, given that we observe an underestimation in precipitation variability when separating noAER and AER influences on GMLP in the previous section (see Figure 7b), residuals that are larger than expected and our models are consistent with observational hydrological sensitivity, we continue this analysis with doubled estimates of model precipitation variance (as it has been done in, e.g., Polson et al. 2013a; Schurer et al. 2020) to address the failure of variance. This also leads to a decrease in residuals for WD and GMLP (see Figure S4).

Overall, results from the imperfect model study suggest that it is helpful to combine precipitation and temperature fingerprints for a clearer estimate of the climate response to external forcings.

**CAN WE IMPROVE THE DETECTION AND ATTRIBUTION OF AEROSOL INFLUENCES IN OBSERVATIONS?**

Based on imperfect model results, the combination of global means in [GMT; GMLP] could be considered the most reliable to estimate aerosol contributions to observed changes as the joint diagnostic estimates the ‘true’ AER contribution with a success rate of close to 90% (similar to GMT; see Figure 6) and shows the tightest constrain. To confirm this, forcing contributions to the observed warming in 2010-2019 relative to 1850-1900 using scaling factors from an observational joint study (see Figure S5) are derived, and for the large-ensemble multimodel mean plotted in Figure 9 (for individual models and the CMIP6 multimodel mean see Figure S7). As for the joint study (Figure 6), contributions are estimated by scaling GMT (multi)model mean fingerprints (covering the period 1850-2019) for AER and noAER with the respective scaling factors (Figure S5) before subtracting the [1850-1900] mean from the [2010-2019] mean. Although the scaling factors are calculated on a shorter period (1988-2020/1955-2020) estimating GMT this way allows us to directly compare and evaluate our approach with recent results from Gillett et al. (2021).

In the case of the large-ensemble multimodel mean, findings from the imperfect model study are supported. The most constrained estimates of AER and noAER contributions of the 7 diagnostics analysed are obtained for [GMT; GMLP] with an aerosol cooling of 0.46 K (0.05-0.86 K) offsetting 1.63 K (1.26-2.00 K) of noAER/GHG warming. [GMT; GMLP] is slightly improving the detection
Fig. 9. Attributable warming [2010-2019] minus [1850-1900] (orange bar, $\beta_{\text{ALL}}$) and contributions of aerosols (blue bar, $\beta_{\text{AER}}$) and noAER (red bar, $\beta_{\text{noAER}}$) to observed warming for temperature and precipitation combined diagnostics for the large-ensemble multimodel mean (LE MMM). Shading highlights the 90% confidence ranges of contributions of GHG (red) and aerosols (blue) to the observed warming (grey) from Gillett et al. (2021).

relative to GMT alone by reducing the uncertainty range of the noAER (mostly GHG) contribution and attributing an AER cooling ($<0$). We further do not constrain the AER impact to cooling based on GMT alone. The addition of precipitation in the fingerprint, therefore, allows us to attribute a response to aerosol forcing of the sign expected from models in observations. This is consistent with results from literature that show AER-induced surface cooling triggers a stronger response in precipitation compared to GHG (Allen and Ingram 2002).

Similar results to those for the large-ensemble multimodel mean are obtained for the individual models and the CMIP6 multimodel mean. As previously noted for the imperfect model study and also here for observations, diagnostics including global means behave best across models, while it is model specific which non-global/dynamical variables cause an increase in uncertainty when added to the combined fingerprint. For example, the addition of the interhemispheric temperature asymmetry to [GMT; GMLP] decreases the uncertainty of the AER contribution for CNRM-CM6-
Table 3. Table showing best estimate contributions and respective 5%-95% uncertainty ranges in squared brackets of AER and noAER to the observed warming using scaling factors for the LE MMM analysis (multimodel mean of the 3 CMIP6 large-ensembles).

<table>
<thead>
<tr>
<th>Diagnostic</th>
<th>AER contribution [K]</th>
<th>noAER contribution [K]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[WD; GMLP]</td>
<td>-0.71 [-2.05; 0.60]</td>
<td>2.55 [1.06; 4.10]</td>
</tr>
<tr>
<td>[ITA; GMT]</td>
<td>-0.51 [-0.96; -0.07]</td>
<td>1.65 [1.25; 2.05]</td>
</tr>
<tr>
<td>[ITA; GMLP]</td>
<td>-0.84 [-1.44; -0.26]</td>
<td>1.59 [0.64; 2.56]</td>
</tr>
<tr>
<td>[GMT; GMLP]</td>
<td>-0.46 [-0.86; -0.05]</td>
<td>1.63 [1.26; 2.00]</td>
</tr>
<tr>
<td>[ITA; GMT; GMLP]</td>
<td>-0.56 [-0.97; -0.14]</td>
<td>1.67 [1.28; 2.07]</td>
</tr>
<tr>
<td>[WD; ITA; GMT]</td>
<td>-0.45 [-0.90; 0.00]</td>
<td>1.65 [1.25; 2.05]</td>
</tr>
<tr>
<td>[WD; ITA; GMLP; GMT]</td>
<td>-0.50 [-0.93; -0.07]</td>
<td>1.67 [1.28; 2.07]</td>
</tr>
</tbody>
</table>

1 and IPSL-CM6A-LR, while it increases the confidence interval for CanESM5 due to its stronger AER response (Figure S7). As for the large-ensemble model mean, AER and noAER contributions can only be clearly attributed for the single-model analyses when using combined fingerprints. We again note that aerosol pattern uncertainty and the under-coverage of uncertainties when using TLS possibly renders our results slightly overconfident (Schurer et al. 2018; DelSole et al. 2019).

Apart from [WD; GMLP] and [WD; ITA; GMT] fingerprints, large-ensemble estimates for noAER and AER agree with recent findings from a 3-signal study in Gillett et al. (2021) which attributes contributions of GHG+, NAT and AER as [1.2;1.9], [-0.01;0.06] and [-0.7;-0.1] °C, respectively. Although we are using a similar methodology, with the main difference being that we do not consider observational uncertainty from the ensemble spread of HadCRUT4 or uncertainty arising from the ratio of global surface air temperature and global mean surface temperature, we find slightly larger uncertainties than Gillett et al. (2021). This is likely due to the fact that attributions in Gillett et al. (2021) are estimated based on the multimodel mean of 13 models and temperature data for the period 1850-2019. For individual models, Gillett et al. (2021) cannot clearly attribute the AER signal when looking at GMT alone, which is consistent with our results. Thus, while the skill of combining temperature and precipitation fingerprints is not so clear for the multimodel mean, it allows for the attribution of aerosol impacts in individual models.

In summary, this study shows that diagnostics combining both precipitation and temperature outperform or perform as well as diagnostics where individual fields of temperature and precipitation are combined ([WD; GMLP] and [ITA; GMT]). Further, our results suggest that a joint
study improves the separation of anthropogenic forcing contributions to observed changes relative to GMT on its own. The addition of a precipitation-based fingerprint increases the detectability of the AER forcing since the probability of no contribution (scaling factor equal to zero) from this fingerprint decreases (except for [WD; GMLP] and [WD; ITA; GMT]). Thus, for most models and diagnostics, we attribute an AER-induced cooling with consistent best estimates of around 0.5K to 0.8K and GHG warming of around 1.7K.

4. Conclusions

Using large-ensembles from three CMIP6 models and observational datasets for temperature (HadCRUT5) and precipitation (CRUTS and GPCP) we comprehensively investigated whether combining physically motivated information from temperature and precipitation into a joint study can reduce the confidence intervals of aerosol contributions to observed changes in GMT. Both a perfect and imperfect model study were conducted to evaluate model uncertainties and to rate our choice of diagnostics. We then compared model sensitivities in temperature and precipitation to evaluate our approach before reconstructing aerosol influences and contributions from the remaining forcings (noAER, mostly GHG) to the observed warming. Finally, we compared our results with estimates from a detection and attribution study using GMT-only and values from recent literature, i.e., Gillett et al. (2021), which is based on a similar methods, but using multimodel mean fingerprints and temperature-only diagnostics. The main findings from the present study are:

- A hydrological sensitivity analysis shows that the 3 large-ensembles simulate the lower bound of the observational energetic relationship between GMP and GMT of $2 \pm 0.5 \text{ } \% \text{ K}^{-1}$ (Allan et al. 2020) and the assumed consistency in model sensitivities in temperature and precipitation to external forcings is further supported by results from an imperfect model study. However, large residuals from a 2-signal analysis for GMLP and inconsistency between findings from the imperfect model and observational study raise the question of whether our models underestimate (tropical) precipitation variability, as has been found in previous studies (e.g., Zhang et al. 2007; Schurer et al. 2020). To address this we, thus, double precipitation variance for the observational detection and attribution analysis.
Apart from the purely precipitation-based diagnostic [WD; GMLP], estimates of noAER and AER contributions agree with the findings from the literature (Gillett et al. 2021). Our best estimate using the multimodel mean of the 3 large-ensembles is an AER cooling of 0.46 K (0.05-0.86 K 90% confidence interval) and a warming associated with noAER of 1.63 K (1.26-2.00 K) using the concatenated time series of GMT and GMLP. We detect negative AER contributions for a range of diagnostics even if GMT is not included and can slightly reduce the uncertainty range of the AER and noAER contributions compared to our study of using GMT alone (AER: -0.38 [-0.81; 0.06] K; noAER: 1.59 [1.21; 1.97] K). Highlighting model sensitivity, we find quite consistent estimates of AER cooling from observations across the three large ensembles: for CNRM-CM6-1 -0.57 [-1.16;0.01] K, IPSL-CM6A-LR: -0.5 [-0.94; -0.09] K and CanESM5 AER -0.42 [-0.8; -0.04] K. This cooling is counteracted by similarly consistent estimates of noAER across the respective same models of 2.2 [1.67; 2.74] K, 1.95 [1.46; 2.46] K and 1.18 [0.9; 1.45] K. These estimates are more robust across models than using GMT only.

We find model dependency in the ability of non-global mean fingerprints to improve the detectability of the AER signal, i.e., the separation from 0, which is likely due to varying aerosol schemes and aerosol sensitivities across models. This highlights the need to better constrain the spatial patterns of aerosol response in climate models.

Although statistically we would expect confidence intervals to deviate from 1 in only 10% of cases for a perfect model study, we find deviation rates range from 0 to 40%. There is a number of factors contributing to deviations from 10%, such as a low signal-to-noise ratio (especially for the precipitation-based fingerprints alone) and because we sample control simulations from the range of CMIP6 models, which do not represent the true internal variability of the models investigated in this study. This causes both unexpectedly low and high failure rates depending on a model’s true internal variability. In the imperfect model study, we do not expect to find deviation rates of 10% as we are calculating scaling factors for model simulations with different sensitivities to forcings, thus, the scaling factors should be different from 1 (deviation rates range from 0 to 100%). Additionally, model uncertainties due to differences in aerosol forcing among the different models,
contribute to the uncertainties (Wilcox et al. 2015; Schurer et al. 2018). A final factor to consider is the potential under-coverage of the confidence intervals derived using TLS, which is an intrinsic limitation of the method that becomes apparent for low signal-to-noise ratios (e.g., single precipitation or ITA fingerprints). This has been noted by DelSole et al. (2019); Li et al. (2021) and Ma et al. (2023), and our study, further, contributes to highlighting these methodological constrains and the need for future improvements.
Acknowledgments. This work was supported by the European Union’s Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No. 860100 (iMIRACLI). GH, AB, and AS were also funded by the Natural Environmental Research Council (NERC) project GloSAT (NE/S015698/1). AS further received funding from a Chancellor’s fellowship at the University of Edinburgh. 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 for producing and making available their model output (listed in Table 2). We thank the iMIRACLI cohort for their guidance and general support throughout the project. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Data availability statement. Observational data from the Global Precipitation Project v3.2 (GPCPv3.2; Huffman et al. 2023), the Climate Research Unit’s gridded Time Series v4.05 (CRUTSv4.05; Harris et al. 2020) and the fifth version of the Hadley Centre/Climate Research Unit Temperature (HadCRUT5; Morice et al. 2021) are freely available online (see https://disc.gsfc.nasa.gov/datasets/GPCPMON_3.2/summary, https://catalogue. ceda.ac.uk/uuid/c26a65020a5e4b80b20018f148556681 and https://www.metoffice. gov.uk/hadobs/hadcrut5/, respectively). CMIP6 (Eyring et al. 2016) model data is freely available through the Earth System Grid (see https://esgf-node.llnl.gov/ projects/cmip6/). The detection and attribution analysis code used in this study is based on ESMValTool (see https://github.com/ESMValGroup/ESMValTool/blob/gillett20/ esmvaltool/; Kirchmeier-Young et al. 2017; Gillett et al. 2021). Code for calculating the time series and reproducing the figures can be found at https://doi.org/10.5281/zenodo.10381597.

References


Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, Eds., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.


Accepted for publication in *Journal of Climate.* DOI 10.1175/JCLI-D-23-0347.1.


