

# Detection and attribution of climate change using a neural network

Constantin Bône<sup>1,2</sup>, Guillaume Gastineau<sup>1</sup>, Sylvie Thiria<sup>1</sup>, Patrick  
Gallinari<sup>2,3</sup> and Carlos Mejia<sup>1</sup>

<sup>1</sup>UMR LOCEAN, IPSL, Sorbonne Université, IRD, CNRS, MNHN

<sup>2</sup>UMR ISIR, Sorbonne Université, CNRS, INSERM

<sup>3</sup>Criteo AI Lab

## Key Points:

- We present a non linear method based on neural network to attribute the global mean surface air temperature variability to different forcings.
- We use a CNN associated with a backward optimization to estimate the climate response to the different external forcings.
- The attributable forcings are consistent with those obtained using another state-of-the-art method.

---

Corresponding author: Constantin Bône, [constantin.bone@sorbonne-universite.fr](mailto:constantin.bone@sorbonne-universite.fr)

**Abstract**

15 A new detection and attribution method is presented and applied to the global mean sur-  
16 face air temperature (GSAT) from 1900 to 2014. The method aims at attributing the  
17 climate changes to the variations of greenhouse gases, anthropogenic aerosols, and nat-  
18 ural forcings. A convolutional neural network (CNN) is trained using the simulated GSAT  
19 from historical and single-forcing simulations of twelve climate models. Then, we per-  
20 form a backward optimization with the CNN to estimate the attributable GSAT changes.  
21 Such a method does not assume additivity in the effects of the forcings. The uncertainty  
22 in the attributable GSAT is estimated by sampling different starting points from single-  
23 forcing simulations and repeating the backward optimization. To evaluate this new method,  
24 the attributable GSAT changes are also calculated using the regularized optimal finger-  
25 printing (ROF) method. Using synthetic non-additive data, we first find that the neu-  
26 ral network-based method estimates attributable changes better than ROF. When using  
27 GSAT data from climate model, the attributable anomalies are similar for both meth-  
28 ods, which might reflect that the influence of forcing is mainly additive for the GSAT.  
29 However, we found that the uncertainties given both methods are different. The new method  
30 presented here can be adapted and extended in future work, to investigate the non-additive  
31 changes found at the local scale or on other physical variables.  
32

**Plain Language Summary**

33 In order to design effective adaptation policies, it is essential to have reliable es-  
34 timates of the effect of anthropogenic activities on the climate. For that purpose a new  
35 attribution method based on a neural network is designed and evaluated. The method  
36 estimates the past global mean surface air temperatures anomalies caused by the changes  
37 in the greenhouse gases concentration, the variation of anthropogenic aerosols, and the  
38 variations driven by naturally occurring phenomena. To build this estimation, the data  
39 from observations and climate models are used. This methodology is compared with an-  
40 other state-of-the-art method. The results of both methods are evaluated and discussed.  
41 The proposed method provide better estimations in the case of large non-additivity of  
42 the causes of climate change and can be applied to other physical variables or at the re-  
43 gional scale. In the case of the global mean surface air temperature, the method presented  
44 provides estimation similar to other methods.  
45

## 1 Introduction

Detection and attribution of climate change is key to understanding past climate change and devising adaptation policies. This problem is an important part of IPCC reports (Eyring et al., 2021) as it directly inquires about the impact of anthropogenic activities on the climate system. Detection aims to compare climate change with internal variability. A change is detected if it exceeds the anomalies generated by the internal climate variability. Internal variability refers to climate variations resulting from processes intrinsic to the climate system, occurring in the absence of external forcing. Internal variability may arise from processes within each of the climate system components (atmosphere, ocean, land surface, cryosphere) or may emerge from their interactions (Cassou et al., 2018). For instance, the global mean surface air temperature (GSAT) varies by a few tenths of degrees during the El Niño or La Niña phases of the El Niño Southern Oscillation (Neelin et al., 1998). Similarly, the Pacific decadal variability and the Atlantic multi-decadal variability can also influence the GSAT (Meehl et al., 2016; Z. Li et al., 2020). Forcing agents external to the climate system, known as external forcings, can also cause climate changes. The dominant forcings in the historical period (i.e. 1850 to present-day) are the increase in the concentration of greenhouse gases, the variations of the aerosol concentrations, the variations of incoming solar radiation, the changes in land use and stratospheric ozone concentration (Masson-Delmotte et al., 2021). Attribution then aims to explain and quantify the impacts of the different forcings. Anthropogenically driven and naturally occurring forcings are often considered separately to understand the impact of human activities. Natural forcings include the effects of natural sources of aerosols and solar activity. The anthropogenic effects include the contributions of other effects. Hasselmann (1993) defined a method called “optimal fingerprinting” for detection and attribution relying on climate model simulations and observations. This method has been improved to build more reliable uncertainties and to check for the consistency between models and observations (Allen & Tett, 1999), or to account for the residual internal variability in ensembles of climate model simulations (Allen & Stott, 2003). To better account for the uncertainty in the estimation of forcings Ribes et al. (2013) proposed to use a regularized estimator of the covariance matrix of internal variability. A review, based, among other, on regularized optimal forcing estimates, concluded that the likely range (5-95% range) of the attributable anthropogenic GSAT anomaly in 2010-2019 relative to 1850-1900 is between +0.8 to +1.3°C (Eyring et al., 2021). The anomaly

79 attributable to greenhouse gases reported is  $+1.0^{\circ}\text{C}$  to  $+2.0^{\circ}\text{C}$ , while it is from  $-0.8^{\circ}\text{C}$   
80 to  $0.0^{\circ}\text{C}$  for other anthropogenic forcings, and from  $-0.1^{\circ}\text{C}$  to  $+0.1^{\circ}\text{C}$  for natural forc-  
81 ings.

82 However, the optimal fingerprinting has several limitations such as the loss of in-  
83 formation due to the reduction of the temporal and spatial dimensionality of data, needed  
84 to make a proper approximation of the covariance matrix of internal variability. Another  
85 problem is the additivity assumption where the individual forcing effects are summed  
86 together to estimate the climate response to the sum of forcings even if it is verified for  
87 the attribution of historical GSAT (Marvel et al., 2015; Shiogama et al., 2013). This ad-  
88 ditivity assumption also found to be invalid for precipitation (Marvel et al., 2015), the  
89 surface air temperature changes driven by greenhouse gases and aerosols can be non-additive  
90 over the extra-tropical regions such as the Arctic (Deng et al., 2020) or the Southern Hemi-  
91 sphere (Pope et al., 2020).

92 To take account of non-additive changes, we present here a new method for attribut-  
93 ing past climate using machine learning. A neural network is a machine learning method  
94 consisting of consecutive hidden layers of nonlinear transformations and adjustable weights  
95 and biases which are determined by applying gradient descent using backpropagation  
96 (Goodfellow et al., 2016). It is a statistical tool increasingly used in recent years in many  
97 scientific fields (Choudhary et al., 2022). Convolutional neural networks (CNN, Yamashita  
98 et al. (2018)) are a class of non-linear neural networks used notably in imagery problems  
99 (O’Shea & Nash, 2015). Their main characteristic is the use of a learnable kernel that  
100 slides along the input data. The CNNs have also shown their great capacity to analyze  
101 time series and other one-dimensional patterns (Kiranyaz et al., 2021) and have become  
102 common machine learning tools. For instance, without being exhaustive, neural networks  
103 have been used in climate science to predict the evolution of El Nino Southern Oscilla-  
104 tion (Ham et al., 2019), to identify storm structures (Gagne II et al., 2019), for weather  
105 prediction (Lam et al., 2022; Gagne II et al., 2019), or for detection studies (Labe & Barnes,  
106 2021; Barnes et al., 2019). However, they are still emerging in large parts of the geosciences.

107 Here, we propose an alternative attribution framework based on a CNN to account  
108 for interactions between the forcings. To the best of our knowledge, this is the first at-  
109 tempt to apply a neural network to the problem of detection and attribution of climate  
110 change. We compare the results obtained with the neural-network based attribution method  
111 with those resulting from regularized optimal fingerprinting. We chose to study the GSAT

112 as it is widely studied in the detection and attribution literature in order to properly in-  
113 troduce our methodology. We investigate the effects of greenhouse gases, anthropogenic  
114 aerosols and natural forcings. In the future, this attribution method based on a neural  
115 network could be applied to other physical variables such as precipitation, or changes  
116 at the regional scale where non-additivity are expected to be more important (Good et  
117 al., 2015).

118 To evaluate our neural network based attribution method and compare it to reg-  
119 ularized optimal fingerprinting, we first build synthetic data to assess the ability of meth-  
120 ods to take non-addivities into account. Then we use a perfect model approach. This  
121 consists of removing data coming from one climate model and treating its simulations  
122 as pseudo-observations. The estimated effect of each forcing is then compared to their  
123 actual simulated effects.

124 The article is organized as follows. In section 2, we present the data and the pre-  
125 processing applied and how we built up synthetic data. In section 3, we present the neu-  
126 ral network and its direct performance. We also introduce the two attribution methods  
127 used in this paper : backward optimization and regularized optimal fingerprinting (ROF).  
128 In section 4, we present the results obtained by the two attribution methods. Finally in  
129 section 5, we conclude and discuss the limitations as well as future perspectives.

## 130 **2 Model and Data**

### 131 **2.1 Climate models simulations**

132 In this section, we present the climate model data used in this study. We use the  
133 monthly surface air temperature from the outputs of the Coupled Model Intercompar-  
134 ison Project 6 phase (CMIP6; Eyring et al. (2016)) and of the Detection and Attribu-  
135 tion Model Intercomparison Project (DAMIP; Gillett et al. (2016)) panel of CMIP6. All  
136 simulations from CMIP6 use the same experimental protocol with identical boundary  
137 conditions based on reconstructions and observations.

138 We use the historical simulations, called HIST, to obtain estimation of the com-  
139 bined effect of the forcings. These simulations use as variable boundary conditions all  
140 external forcings from 1850 to 2014. This includes the reconstructed concentrations of  
141 greenhouse gases, anthropogenic aerosols and ozone, and the estimated past variations  
142 of solar incoming radiation and land-use.

143 We also use single-forcing simulations to obtain estimation of the individual effect  
144 of the forcings. These simulations use as variable boundary conditions only one of the  
145 external forcings, all the other external forcings being fixed at their value from 1850. We  
146 use the single-forcing simulations hist-GHG denoted later GHG, hist-aer denoted AER,  
147 and hist-nat denoted NAT dedicated respectively to greenhouse gas concentrations, an-  
148 thropogenic aerosols, and natural forcings (i.e. volcanic aerosol and solar variations) as  
149 variable forcings for the same period (1850-2014). The effect of stratospheric ozone and  
150 land use was not investigated as only a few simulations have been performed in CMIP6,  
151 and because their effective radiative forcings are much smaller than the ones of green-  
152 house gases, aerosols or natural forcings (Smith et al., 2020).

153 We also use the preindustrial control simulations, called PI, to estimate of the ef-  
154 fects of internal variability. These control simulations use fixed forcings from their es-  
155 timated pre-industrial levels corresponding that of 1850. The PI simulations are multi-  
156 centennial with usually a single realization for each climate model. These simulations  
157 show a small drift due to incomplete spin-up or nonclosure of the energy budget (Hobbs  
158 et al., 2016). Hereafter such small long-term drift (Irving et al., 2021) is deleted from  
159 each PI simulations by removing a quadratic trend (Gupta et al., 2013) of the simulated  
160 GSAT before analysis in all simulations.

161 All simulations but PI includes multiple realizations called ensemble members and  
162 denoted later as members. The members use different initial conditions which are sam-  
163 pled from the PI simulation. We use 12 atmosphere-ocean general circulation models (AOGCMs,  
164 see Tab. 1 for details) where at least two members are available for the simulations HIST,  
165 GHG, AER and NAT.

**Table 1.** Presentation of the climate models used.  $n_{GHG}$ ,  $n_{AER}$ ,  $n_{NAT}$  and  $n_{HIST}$  denote the number of members used for GHG, AER, NAT and HIST. The duration of the PI simulation is indicated, in yr.  $\sigma_{PI}$  denotes the year to year standard deviation of the GSAT from PI, in °C.

Model	$n_{GHG}$	$n_{AER}$	$n_{NAT}$	$n_{HIST}$	PI (yr)	$\sigma_{PI}$ (°C)	Reference
CanESM5	50	30	30	65	1000	0.10	Swart et al. (2019)
CESM2	3	3	2	11	500	0.13	Danabasoglu et al. (2020)
IPSL-CM6-LR	10	10	10	32	1000	0.15	Boucher et al. (2020)
ACCESS-ESM1-5	3	3	3	30	500	0.11	Ziehn et al. (2020)
BCC-CSM2-MR	3	3	3	3	600	0.17	Wu et al. (2019)
CNRM-CM6-1	9	10	10	30	500	0.13	Voldoire et al. (2019)
FGOALS-g3	3	3	3	6	700	0.10	Li et al. (2020)
HadGEM3	4	4	4	5	500	0.11	Roberts et al. (2019)
MIROC6	3	3	3	50	500	0.13	Tatebe et al. (2019)
MRI-ESM2.0	5	5	5	7	500	0.10	Yukimoto et al. (2019)
NorESM2-LM	3	3	3	3	500	0.15	Seland et al. (2020)
GISS-E2-1-G	5	7	15	19	500	0.15	Kelley et al. (2020)

166

## 2.2 Observations

167

168

169

170

171

172

173

174

175

We use observations of the 2m air temperature from HadCRUT5 (Morice et al., 2021). The gridded data is a blend of the CRUTEM5 (Osborn et al., 2021) land-surface air temperature dataset and the HadSST4 (Kennedy et al., 2019) sea-surface temperature (SST) dataset. Such a blending is necessary because there are few observations of temperature at 2 meters over the oceans compared to SST observations. The resulting globally averaged quantity is called global mean surface temperature (GMST) and it differs from the GSAT which is solely based on surface air temperature. In order to correct this we multiply by 1.06 the GMST from observation to estimate the observed GSAT, as estimated by Richardson et al. (2018).

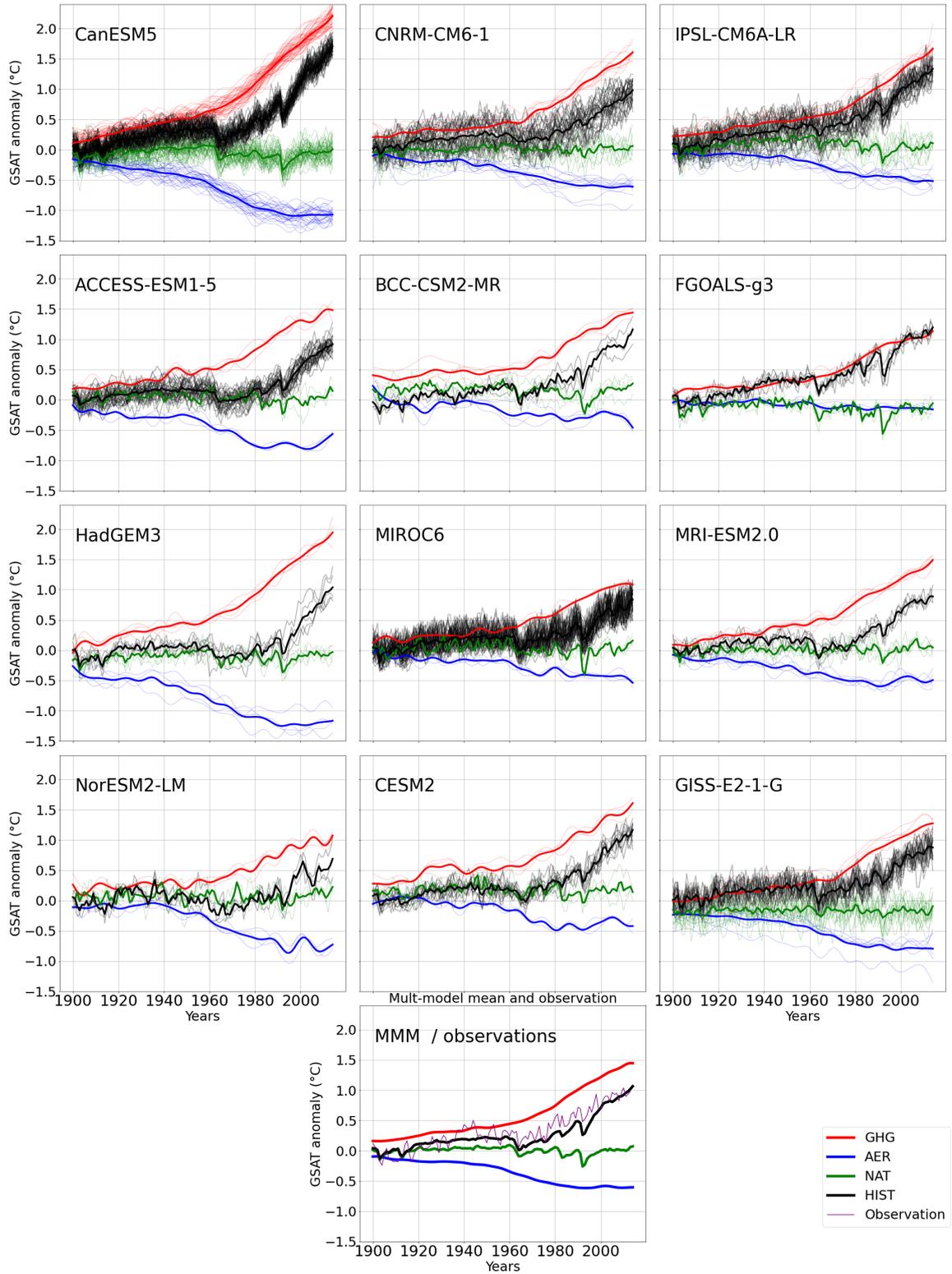
### 176 **2.3 Pre-processing**

177 All monthly climate model data are aggregated to an annual mean and spatially  
178 averaged from 90°S to 90°N to provide the GSAT. We then estimate the temperature anoma-  
179 lies compared to the pre-industrial period.

180 We remove the time mean GSAT of PI from the GHG, AER, and NAT simulations.  
181 For observations and HIST, we compute the average temperature during the 1850-1900  
182 period and remove it from the GSAT. Hereafter we only use the data from 1900-2014  
183 period (115 years).

184 The simulated and observed GSAT can be separated into a forced component and  
185 an internally-generated climate variability component. To reduce the effects of internal  
186 climate variability we apply a low-pass filter to the GSAT of the GHG and AER sim-  
187 ulations. We use a Lanczos low-pass filter (Burger & Burge, 2009), with a window size  
188 of 21 years, and a cutoff period of 10 years. The endpoints are estimated by extending  
189 the time series by replicating the mean value of the first and last ten years of each sim-  
190 ulation. This should not alter the estimated effect of greenhouse gases or aerosols on the  
191 GSAT as both forcings only show multi-decadal and longer fluctuations in terms of ef-  
192 fective radiative forcing (Gulev et al., 2021). We do not apply this procedure to NAT  
193 and HIST because the emission of aerosol from volcanic eruptions induces an intense cool-  
194 ing for the next 2 to 5 years, and such smoothing would degrade the forced anomalies.  
195 This smoothing procedure only lead to minor improvements regarding the estimated un-  
196 certainties (not shown).

197 We illustrate in Fig. 1 the processed data for all climate models, observations and  
198 the multi-model mean (MMM) for each forcing. To compute the MMM we first compute  
199 the ensemble mean (i.e averaging all ensemble member) for each climate model and then  
200 we average the 12 ensemble means. In all models, GHG shows a monotonic warming with  
201 an increasing slope since the 1960's, as expected from the greenhouse gases emissions.  
202 In AER, the aerosol induces a cooling with a pronounced slope from the 1940's to 1980's,  
203 and a plateau from 1980 to 2014. NAT shows small cooling from 0.1 to 0.4°C only oc-  
204 ccurring after the major eruptive volcanic eruptions of Agung (1963), El Chichon (1982)  
205 and Pinatubo (1991). HIST shows a monotonic warming less pronounced than GHG with  
206 also a cooling a few years after the major volcanic eruptions. In all simulations, the in-  
207 ternal variability is important, as illustrated by the fluctuations visible in each members  
208 (thin lines) and is reduced in the ensemble mean (thick lines).



**Figure 1.** GSAT anomaly simulated by each model and (lower panel only) multi-model mean (MMM) and observed GSAT. Black lines show the HIST members. Red lines show the GHG members. Green lines show the NAT members. Blue lines show the AER members. The purple line shows the observations in the lower panel. Bold lines of the same colors show the ensemble mean.

209

## 2.4 Synthetic data

210

211

212

213

214

215

216

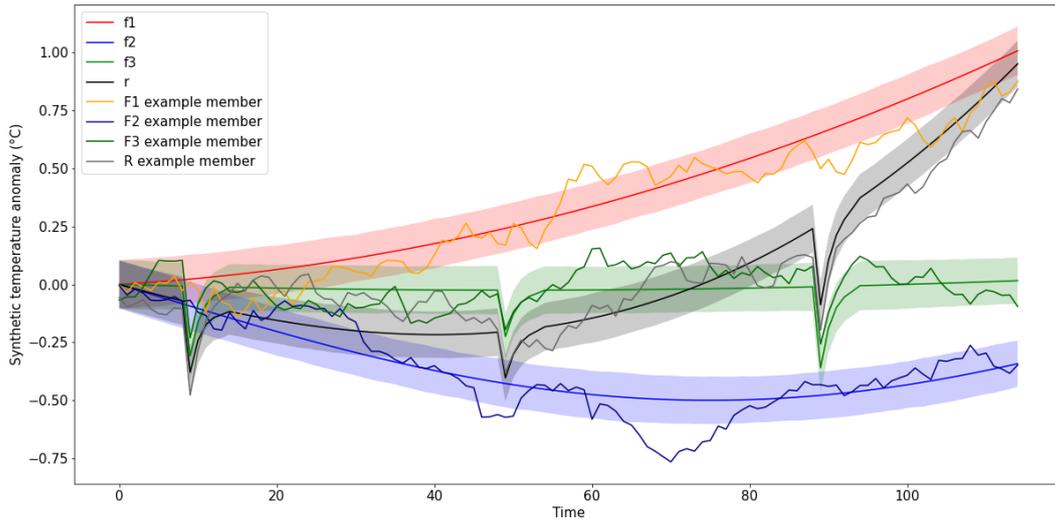
217

218

219

To investigate the performance of the attribution methods when considering external forcings with non-additive influences, a synthetic data set is generated. We generate three time series of size 115 denoted  $f_1$ ,  $f_2$  and  $f_3$ , that represents the forced effects of three synthetic forcings. These time series are constructed to have similarities with the expected influence of the greenhouse gases, aerosol and natural forcing for  $f_1$ ,  $f_2$  and  $f_3$ , respectively (see Fig. red, green and blue lines in 2). However, the expressions of  $f_1$ ,  $f_2$  and  $f_3$  remain arbitrary and are not meant to represent simulated or observed climate. We detail in Text S1 the analytic expressions used to build the time series. We construct the total effect of the three forcings combined, noted  $r$ , using two additional term compared to the additive case :

$$r = f_1 + 0.3f_1^2 + f_2 + f_3 + 0.1f_1f_2 \quad (1)$$



**Figure 2.** Synthetic time series  $f_1$  (red),  $f_2$  (blue),  $f_3$  (green), and  $r$  (black). A randomly chosen time series after adding the variability is illustrated for  $F_1$  (orange),  $F_2$  (dark blue),  $F_3$  (dark green) and  $R$  (grey). Colors shades indicate one standard deviation across the 100 surrogate time series obtained for each pseudo-forcings and their response.

220

221

222

Using an analogy with climate, anomalies are considered to result from the addition of a forced and an internally-generated variability component (see Fig. 1). We add an additional variability to  $f_1$ ,  $f_2$  and  $f_3$  and  $r$  that only represent the forced compo-

223 ment. To generate this variability, we fit a first order autoregressive (AR1) model using  
 224 the time series obtained from the concatenated PI simulations from all models. This AR1  
 225 model is then used to generate 410 surrogate time series that are added to  $f_1$ ,  $f_2$ ,  $f_3$  and  
 226  $r$ . This provides the 100 time series for each forcings denoted  $F_1$ ,  $F_2$ ,  $F_3$  and 110 time  
 227 series  $R$  resulting from the combined forcings (see Fig. 2).

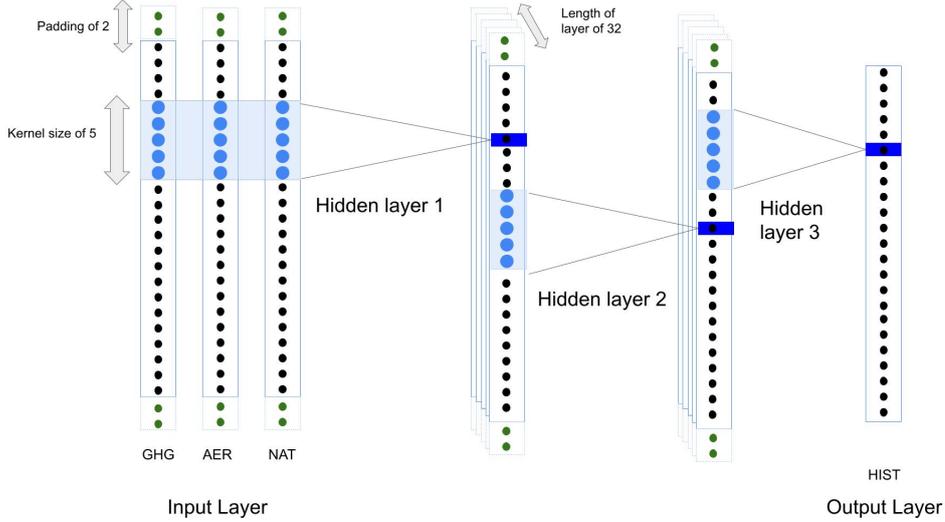
## 228 3 Methods

### 229 3.1 Backward optimization of a neural network

#### 230 3.1.1 Neural network

231 In this section we describe the neural network used. We determine the relationship  
 232 linking the GSAT from HIST to that of GHG, AER, and NAT using a CNN. In the train-  
 233 ing procedure, we use the GSAT from AER, GHG, and NAT as inputs and the GSAT  
 234 from HIST as the target. Our goal is to construct a predictor that captures the role of  
 235 all forcings combined. We assume that stratospheric ozone and land use do not affect  
 236 this relationship.

237 A schematic of the CNN used is shown in Fig 3. CNNs can be used to construct  
 238 relatively simple neural networks as the number of weights and biases is directly decided  
 239 by the size and number of the filters used. We assume that this architecture is suitable  
 240 in the present case the size of the data set is relatively small compared to other neural  
 241 network applications. This might limit the overfitting which occurs when a neural net-  
 242 work model performs significantly better for training data than it does for new data. In  
 243 our case, a one-dimensional kernel is applied to the temporal dimension. To fix the val-  
 244 ues of the weights and biases of the convolutional layers, a neural network needs a learn-  
 245 ing data-set composed of input-output pairs. The outputs are the GSAT of one HIST  
 246 member while the inputs are built with one member for each single-forcing simulations.  
 247 We build this data set by going through all combinations of GHG, AER, NAT and HIST  
 248 members of the same climate model. In order to test the backward optimization (see sec-  
 249 tion 3.3), we removed one HIST member from each climate model and 10 for the IPSL-  
 250 CM6-LR model from these combinations to serve as test data-set. This provides for the  
 251 training of the neural network  $N_d = (n_{HIST} - 1) n_{GHG} n_{AER} n_{NAT}$  4-tuples for each  
 252 climate model except for IPSL-CM6-LR with  $(n_{HIST} - 10) n_{GHG} n_{AER} n_{NAT}$  4-tuples.  
 253 We note  $N_d$  the total number of the 4-tuples obtained for all models. The training data-



**Figure 3.** Diagram of the CNN used. Each white-filled blue rectangles represents a time series of 115 years. The input layer is shown on the left, the hidden layers on the middle and the output layer on the right. Light blue-filled rectangles represent the kernels of the different hidden layers. Dark blue-filled rectangles represent the output of the kernel. Zero-padding is shown in green with dotted lines.

254 set is thus of size  $N_d$  which is of the order of  $10^5$  while an individual input is of size (3,115)  
 255 and its corresponding output of size (1,115). The usual practice is to go through this database  
 256 a number of times to train the CNN. However, we have altered the procedure to provide  
 257 a similar weight to all models during the training.

258 Three steps are applied. First, a climate model is randomly selected. Secondly, we  
 259 randomly select one 4-tuple from the chosen climate model. Then the CNN is trained  
 260 using the three GSAT time series dedicated to (GHG, AER, NAT) as input and the GSAT  
 261 dedicated to HIST as the target. We iterate this process by repeating it  $5 \cdot 10^6$  times. A  
 262 lower number of iterations was found to degrade the backward optimization results (not  
 263 shown), but the results are otherwise similar when increasing the number of iterations.

264 A neural network uses hyperparameters which are the variables that determine the  
 265 network structure and those which determine how the network is trained. The hyper-  
 266 parameters are chosen using a cross-validation, as detailed in Text S2 and Fig. S1. The  
 267 chosen architecture has three convolutional hidden layers, a kernel size of 5 for all lay-  
 268 ers and 32 filters for each layer.

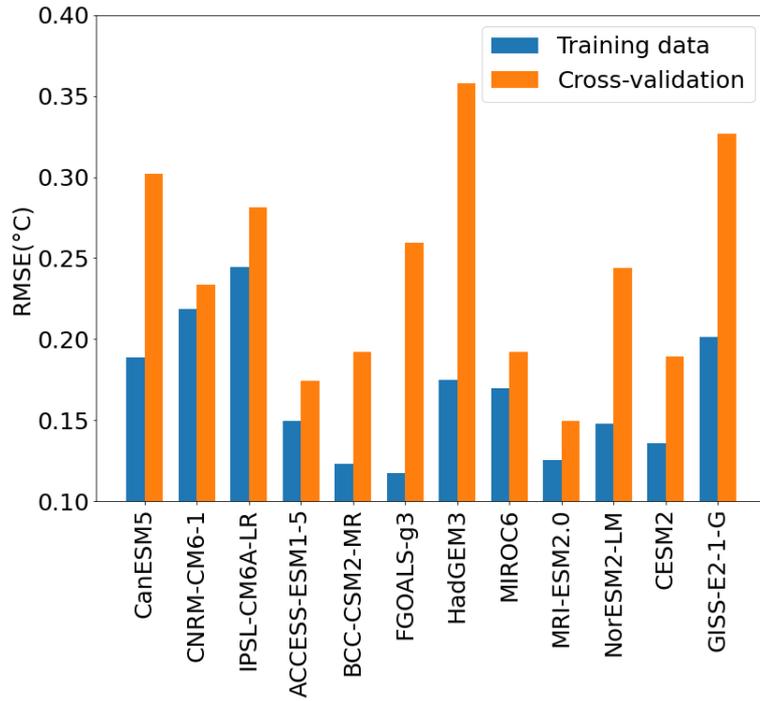
## 269 **3.2 Performance of the CNN**

270 Before presenting the neural network dedicated to the attribution method in the  
 271 next section, we investigate the performance of the CNN in estimating the total effect  
 272 of forcing from the effect of each forcing separately. First, we train the CNN using the  
 273 data from all models and estimate the mean training RMSE made in predicting the data  
 274 for each model separately. Second, we successively train the CNN leaving out the data  
 275 from one model and estimate the mean cross-validation RMSE in predicting the left-out  
 276 model data. Because internal variability is included in the training data, we expect the  
 277 RMSE to exceed the internal variability in all climate models. The training RMSE is  
 278 within  $0.10^{\circ}\text{C}$  and  $0.25^{\circ}\text{C}$  for the different climate models. Indeed, the models with large  
 279 training RSME (Fig. 4 blue bars) corresponds to those simulating a large internal vari-  
 280 ability, as estimated by the standard deviation of the GSAT of the PI simulation (Tab.  
 281 1), where the forced signal is absent.

282 The CNN also should produce an estimated GSAT similar to the mean output from  
 283 the training data, which is expected to be similar to the MMM from HIST. The train-  
 284 ing RMSE may also reflect a forced signal in the HIST simulations distinct from the other  
 285 models. The amplitude of the RMSE increases to  $0.15^{\circ}\text{C}$ - $0.35^{\circ}\text{C}$  when using cross-validation.  
 286 This suggest that the CNN does not overfit. HadGEM3 and, to a lesser extent, FGOALS-  
 287 g3 and GISS-E2-1-G, show differences much larger than the training RMSE when the  
 288 data from these models is used for the validation. This might reflect important singu-  
 289 larities for these three models, which is probably linked to their singular response to forc-  
 290 ings. This might be linked to the equilibrium climate sensitivity which quantifies the abil-  
 291 ity of a model to warm up when greenhouse gases increase. It depends on the feedbacks  
 292 acting in the climate system, and remains poorly constrained by observations (Sherwood  
 293 et al., 2020). GISS-E2-1-G simulate one of the lowest equilibrium climate sensitivity, while  
 294 HadGEM2 has one of the highest sensitivity. In addition, FGOALS-g3 simulate almost  
 295 no response to anthropogenic aerosols (see Fig. 1)

### 296 **3.2.1 Backward optimization**

297 In this section we describe how we use the CNN to perform climate change attri-  
 298 bution. The backward optimization is a method that infers the most likely input of the  
 299 CNN from a given output. To attribute climate change from the CNN, we calculate such



**Figure 4.** RMSE between the CNN output and the GSAT of HIST, in °C, when using (blue bar) the training data and (cross validation, orange bar) when using the data of a model left out in the training.

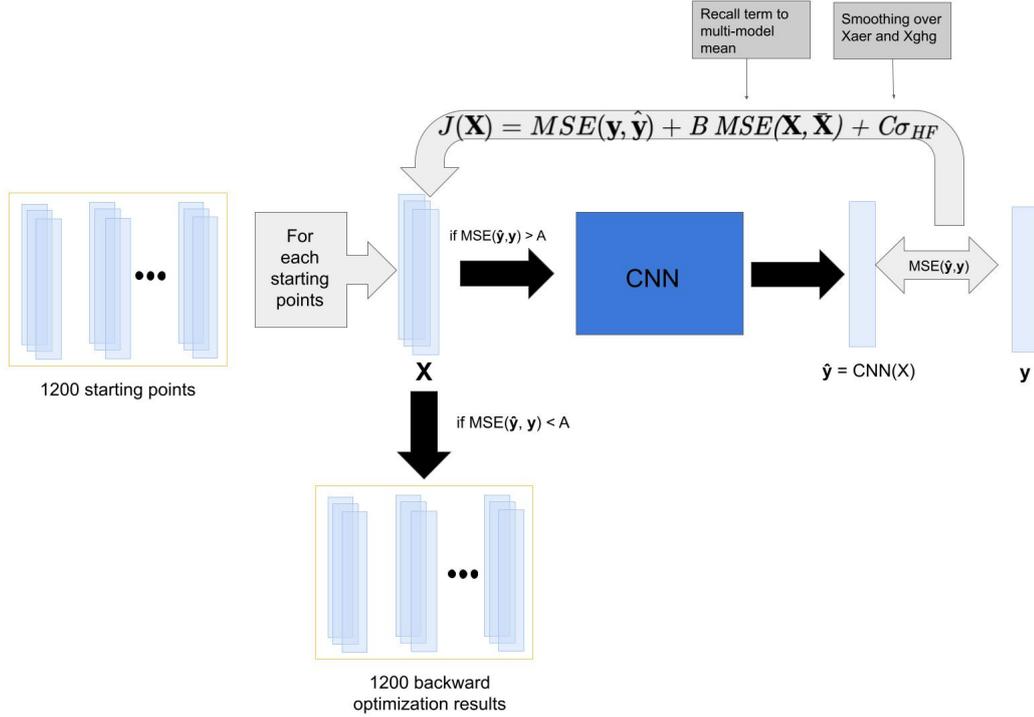
input, which provides the GSAT attributed to the three forcings from the total GSAT anomaly observed or simulated. This is a neural network interpretation method (Toms et al., 2020; Gagne II et al., 2019; McGovern et al., 2019) also known as variational inversion when applied to a geophysical model (Brajard et al., 2012). A scheme of the procedure is given in Fig. 5. This optimal input is determined by minimizing a dedicated cost function and using the backpropagation. The cost function, called  $J$ , is:

$$J(\mathbf{X}) = \text{MSE}(\mathbf{y}, \text{CNN}(\mathbf{X})) + B \text{MSE}(\mathbf{X}, \bar{\mathbf{X}}) + C \sigma_{HF} \quad (2)$$

where  $\mathbf{X} = (x_{GHG}, x_{AER}, x_{NAT})$  is the optimal input to be determined, i.e. a triple of 115-yr time series corresponding to the GSAT induced by greenhouse gases, anthropogenic aerosols and natural forcing.  $\bar{\mathbf{X}}$  is the three time series obtained with the MMM of the simulations GHG, AER and NAT (see Fig. 1, lower panel). MSE denotes the mean squared error.  $\mathbf{y}$  is the desired output of the neural network.  $\sigma_{HF}$  is the sum of the time standard deviation of the high-pass filtered time series obtained from  $x_{GHG}$  and  $x_{AER}$  using a Lanczos high-pass filter with a window of size 21, a cutoff period of ten years.  $B$  and  $C$  are two adjustable real parameters.

The first term on the right hand side of equation (2) measures the the mean square error between the desired output and the CNN output. The second term, also known as a background term, is applied so that the results are similar to a first guess, taken from the MMM in order to avoid absurd and nonphysical solutions. Although this term is not standard for the backward optimization of a neural network, it is however used for the variational inversion procedure used in data assimilation (Brajard et al., 2012; Fablet et al., 2021). The last term is used to build smooth GSAT time series for the forcings associated with greenhouse gases and anthropogenic aerosols. Again, this term is not used for the natural forcings, so that the effects from volcanic aerosols remains unsmoothed, with cooling peaks lasting two to five years, as expected.

When estimating the optimal input, the initial input is iteratively updated using a back-propagation to minimize  $J(\mathbf{X})$  until it is smaller than a fixed value, called  $A$ . To reduce the computational cost, the minimization process is stopped after 500 iterations if  $J(\mathbf{X})$  does not converge. The backward optimization of a neural network has multiple solutions and the method is sensitive to the initial value used for  $\mathbf{X}$ . Therefore, for each of the twelve climate models, we randomly select with repetition 100 (10 during the perfect model approach) triples of the GSAT time series among the members of GHG,



**Figure 5.** Schematic of the backward optimization attribution process with one entry denoted  $\mathbf{y}$  at the right. The 1200 backward optimization results are at the bottom. The learned CNN is in the middle in dark blue.  $J(\mathbf{X})$ , the cost function of the backward optimization is on the top.  $\mathbf{X}$  denote the optimized input and  $\hat{\mathbf{y}}$  denotes its image by the CNN. The 1200 starting points are on the left.

331 AER, and NAT as first guess for the initial states. These initial states are chosen as they  
 332 represent physically coherent inputs. This provides 1200 initial physically coherent val-  
 333 ues for  $\mathbf{X}$  which sample the internal climate variability and the spread among the dif-  
 334 ferent models. This generates 1200 backward optimizations. This estimation is empir-  
 335 ical and does not account for the internal variability of the target of the backward op-  
 336 timization. For each year, the 90% confidence intervals of the optimal input is then es-  
 337 timated using  $\pm 1.64$  standard deviations among the backward optimization results as-  
 338 suming a Gaussian distribution.

339 The choice of  $A$  (iteration stop treshold),  $B$  (background term) and  $C$  (smooth-  
 340 ing term) was fixed empirically as the other hyperparameters of the neural network. We  
 341 found that these parameters do not significantly modify the results of the backward op-  
 342 timisation (see Text S3, Tab. S1 and Tab. S2). We select  $A = 0.05$ ,  $B = 0.01$  and  $C =$   
 343  $0.1$ .

### 3.3 Regularized optimal fingerprints

We evaluate the performance of the neural network based method for detection and attribution by comparing its results to those obtained with the regularized optimal fingerprinting (ROF, Ribes et al. (2013)). This last method is widely used and has already been applied to the air surface temperature using CMIP6 data by Gillett et al. (2021).

The ROF method is based on a multivariate linear regression and on the assumption that the observed change can be obtained with the sum of the forced anomalies for each forcing (the so-called fingerprints) plus internal variability.

The observed GSAT denoted  $\mathbf{y}$ , is given by:

$$\mathbf{y} = \beta \mathbf{X} + \epsilon \quad (3)$$

with  $\beta = (\beta_{GHG}, \beta_{AER}, \beta_{NAT})$  the scaling factors and  $\mathbf{X} = (X_{GHG}, X_{AER}, X_{NAT})$  the effects of all the forcings on the GSAT.  $\epsilon$  represents the effect of internal variability, assumed to be a Gaussian white noise.

We use greenhouse gases, anthropogenic aerosols and natural forcings as three individual forcings and neglect the other forcings.  $\mathbf{X}$  is estimated in this case by using the MMM of GHG, AER and NAT simulations.

To perform such a regression, a common method is to reduce the dimension of data using the leading empirical orthogonal functions calculated in PI. This reduces the number of spatial dimensions and allows an accurate estimation of the internal variability covariance matrix. But such a method involves an arbitrary choice of the number of EOFs used to truncate the data. The ROF method (Ribes et al., 2013) avoids this arbitrary choice using a regularized estimation of the covariance matrix to estimate the scaling factors.

The response of climate to the  $i$ -th forcing is detected if  $\beta_i$  is significantly different from zero. If the confidence interval of  $\beta_i$  includes one, this shows consistency between observations and simulated climate model responses. We use the total least square (TLS) method (Allen & Stott, 2003) to perform the regression and estimate the scaling factors, which accounts for the residual internal variability in the MMM. The internal variability is assumed to be the same in GHG, AER and NAT members, which prevents the use of different smoothing to the GSAT simulated in GHG and AER, as done for the backward optimization, or in NAT. As the internal variability is largely reduced by the ensemble averaging in the MMM, we estimate the attributable warming in GSAT by  $\beta_i X_i$

375 for the  $i$ -th forcing. This should lead to an attributable warming similar to  $\beta_i \hat{X}_i$  using  
 376 the estimated  $X_i$  by the TLS instead of  $X_i$ . Estimates of attributable warming in GSAT  
 377 for each year can then be obtained by  $\sum \beta_i X_i$ . Following Gillett et al. (2021), the in-  
 378 ternal variability is sampled by concatenating all available simulations after subtraction  
 379 of the mean of the corresponding model ensemble. To account for the subtraction of the  
 380 ensemble mean, we multiply for each model, the anomalies by  $\sqrt{\frac{n}{n-1}}$ , where  $n$  is the en-  
 381 semble size. For each simulation, the equivalent size corresponding to the MMM is es-  
 382 timated using:

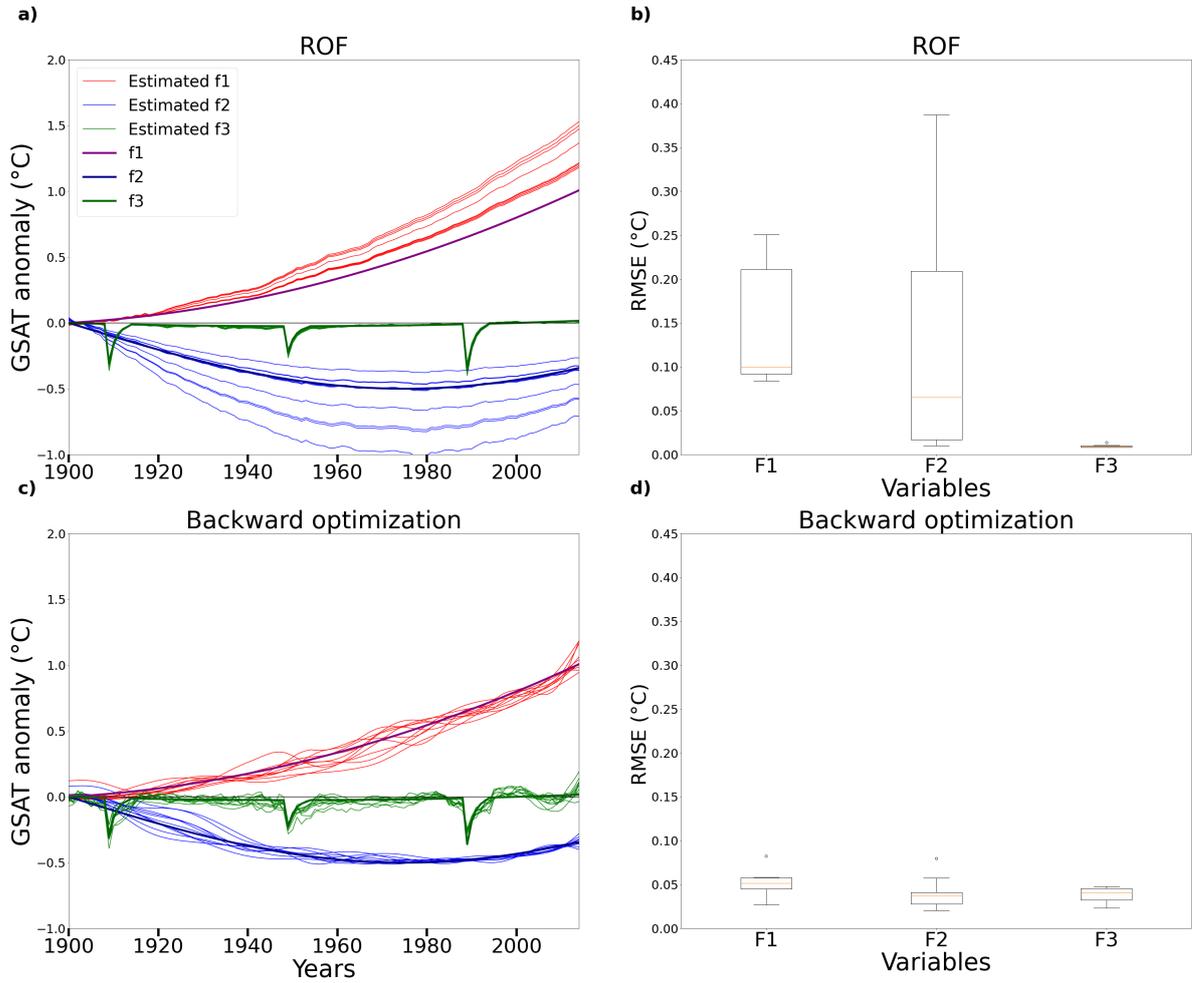
$$N = \frac{M^2}{\sum_{i=1}^M \frac{1}{n_i}} \quad (4)$$

383 with  $M$  the number of different climate models used (in our case 12) and  $n_i$  the num-  
 384 ber of members available for the  $i$ -th climate model. To estimate the uncertainty in the  
 385 GSAT effect attributable to the  $i$ -th forcing, it is necessary to take into account the un-  
 386 certainty of  $\beta_i$  and the internal variability contained in  $X_i$ . For each year and forcing,  
 387 the uncertainty in the attributable GSAT is calculated using 1000 random draws assum-  
 388 ing a gaussian distribution for both  $\beta_i$  and  $X_i$ . The mean and standard errors of  $\beta_i$  are  
 389 estimated as in Allen and Stott (2003). The mean and standard deviation of  $X_i$  are es-  
 390 timated from the size  $N$  of the MMM and the standard deviation of the GSAT obtained  
 391 from the PI runs. We first calculate the standard deviation for each model (as given in  
 392 Tab. 1), average the values obtained across models, and then divide by the square root  
 393 of  $N$ . This procedure is valid under the conditions that the uncertainties of  $\beta_i$  and  $X_i$   
 394 are Gaussian, uncorrelated and small compared to their respective means. The latter hy-  
 395 pothesis is not verified for GSAT anomalies close to zero for  $X_i$ , such as those obtained  
 396 in the first decades of our time series (see Fig. 1), or for the GSAT of NAT. Thus the  
 397 uncertainties for the attributable GSAT are to be taken with caution.

## 398 4 Attribution performances

### 399 4.1 Performance on synthetic data

400 To investigate the performance of the backward optimization and ROF in the case  
 401 of non-additive data, we applied the two attributions methods to the synthetic data pre-  
 402 sented in section 2.4. Fig. 6ac shows the time series of the estimated effect of the three  
 403 synthetic forcings and  $f_1$ ,  $f_2$  and  $f_3$  the ground truth time series. We use the 100 sur-



**Figure 6.** Estimated  $f_1$ ,  $f_2$  and  $f_3$  given by (a) ROF and (c) backward optimization. The original  $f_1$ ,  $f_2$  and  $f_3$  ground truth lines are shown in bold. Histograms shows the distribution of the RMSE of the results of b) ROF and d) backward optimization compared to the ground truth.

404 rogate time series generated for each forcings and their response denoted  $F_1$ ,  $F_2$ ,  $F_3$  and  
 405  $R$ , instead of the simulated GSAT from GHG, AER, NAT and HIST, respectively. The  
 406 10 R time series remaining are used as pseudo-observation, noted  $\mathbf{y}$  previously. For the  
 407 backward optimisation the estimated forced effect  $f_1$  (Fig. 6c, red lines) show some vari-  
 408 ability but is centred around the true  $f_1$  (purple line). For ROF (Fig. 6a, red lines), the  
 409 estimated  $f_1$  are systematically larger than the true  $f_1$  at the end of the time series. Sim-  
 410 ilarly,  $f_2$  (Fig. 6c, blue and dark blue lines) is well estimated by the backward optimiza-  
 411 tion, while ROF (Fig. 6a) produce an estimated  $f_2$  with an important variability and  
 412 an overestimation in most of the cases. The  $f_3$  forcing is well estimated by both meth-  
 413 ods, but with more variability for backward optimization.

414 The RMSE between the effect estimated by the different attribution methods and  
 415 the ground truth are shown in 6bd in the form of boxplot. For ROF the mean RMSE  
 416 value is  $0.14^\circ\text{C}$  for  $f_1$ ,  $0.12^\circ\text{C}$  for  $f_2$  and  $0.01^\circ\text{C}$  for  $f_3$ . These values are for backward op-  
 417 timization of  $0.05^\circ\text{C}$  for  $f_1$ ,  $0.04^\circ\text{C}$  for  $f_2$  and  $0.04^\circ\text{C}$  for  $f_3$ . Backward optimization there-  
 418 fore provides errors smaller than ROF in case of the non-additive forcing generated, while  
 419 the use of ROF lead to important errors.

## 420 **4.2 Evaluation of the performances in attributing climate changes : per-** 421 **fect model approach**

422 To evaluate the performance of the backward optimization and ROF we use a per-  
 423 fect model approach that relies on climate model data only. This approach consists of  
 424 using the data from all but one of the climate models to perform our two attribution meth-  
 425 ods. In the case of backward optimization, this implies that we do not use the data from  
 426 a climate model during the CNN training phase, in the starting points, or in the MMM  
 427 calculation. For ROF, the data of a model are not used to construct the climate noise  
 428 estimate or included in the MMM. We use a HIST member of the test dataset (see Sec-  
 429 tion 3.3) from each climate model as the target for the attribution methods. The attributable  
 430 anomalies associated with each forcing are then compared with the ensemble mean of  
 431 the GHG, AER and NAT simulations of the removed climate model, even if it includes  
 432 some residual internal variability, especially when the number of members is small. We  
 433 use the paradigm that “climate models are statistically indistinguishable from the truth”  
 434 (Ribes et al., 2017; Hargreaves, 2010; van Oldenborgh et al., 2013), where the difference  
 435 between observations and models is assumed to be distributed as the difference between

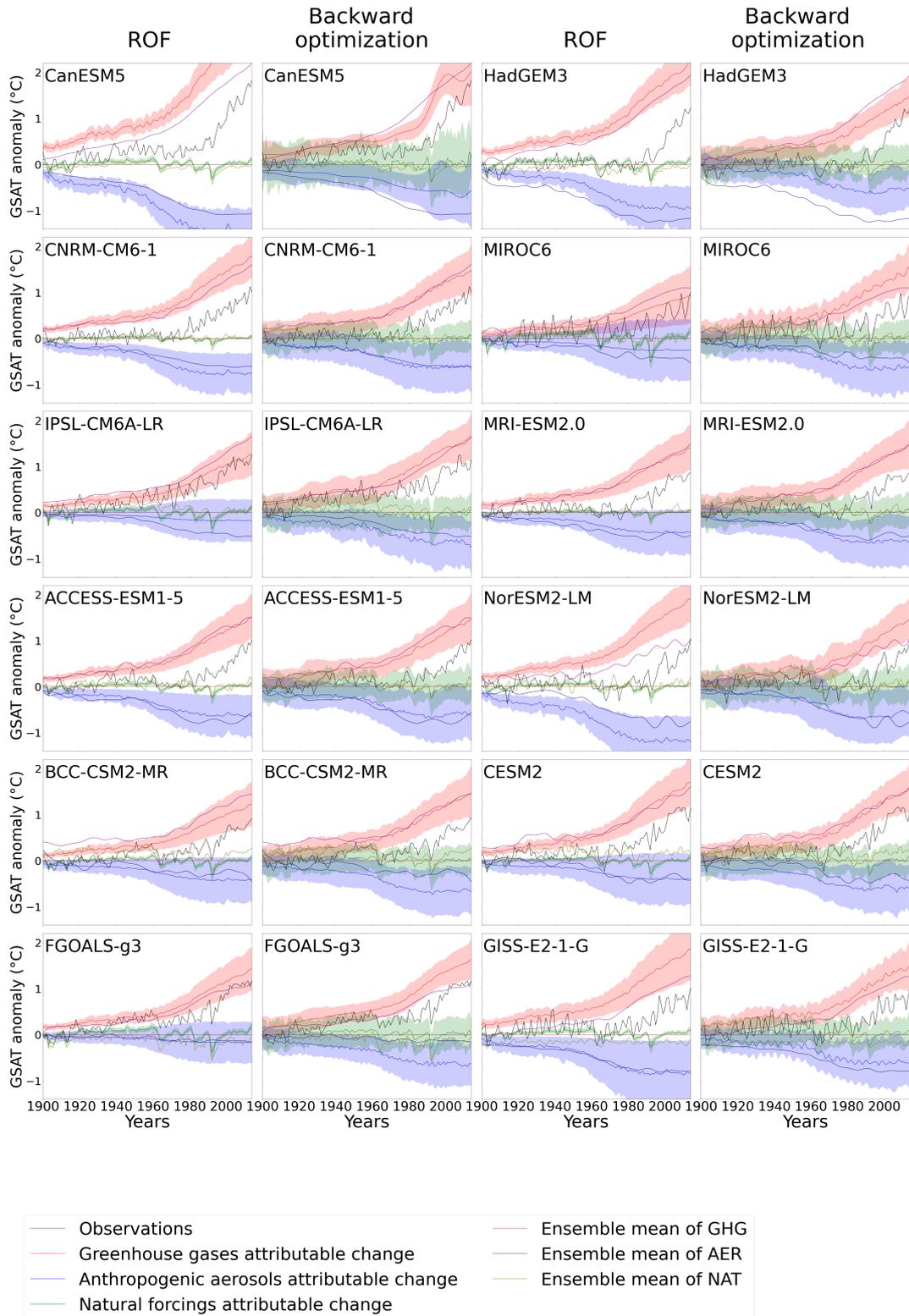
any pairs of climate models. We therefore assess the capability of the attribution methods when using observations by investigating only climate models. This approach is called a perfect model approach by analogy with the methods developed for seasonal (Doblas-Reyes et al., 2013) or decadal (Hawkins et al., 2011) climate forecast.

Figure 7 illustrates the attributable anomalies calculated from an HIST member for each climate model. The ensemble means of GHG, AER and NAT simulations for that climate model are shown for comparison. The differences between the attributable anomalies and the ensemble means of GHG, AER and NAT are also quantified in Fig. 8ab and 8ef with the RMSE and the time mean difference between the two time series. Lastly, the widths of the 90% confidence intervals in 2000-2014 are compared in Fig. 8cd.

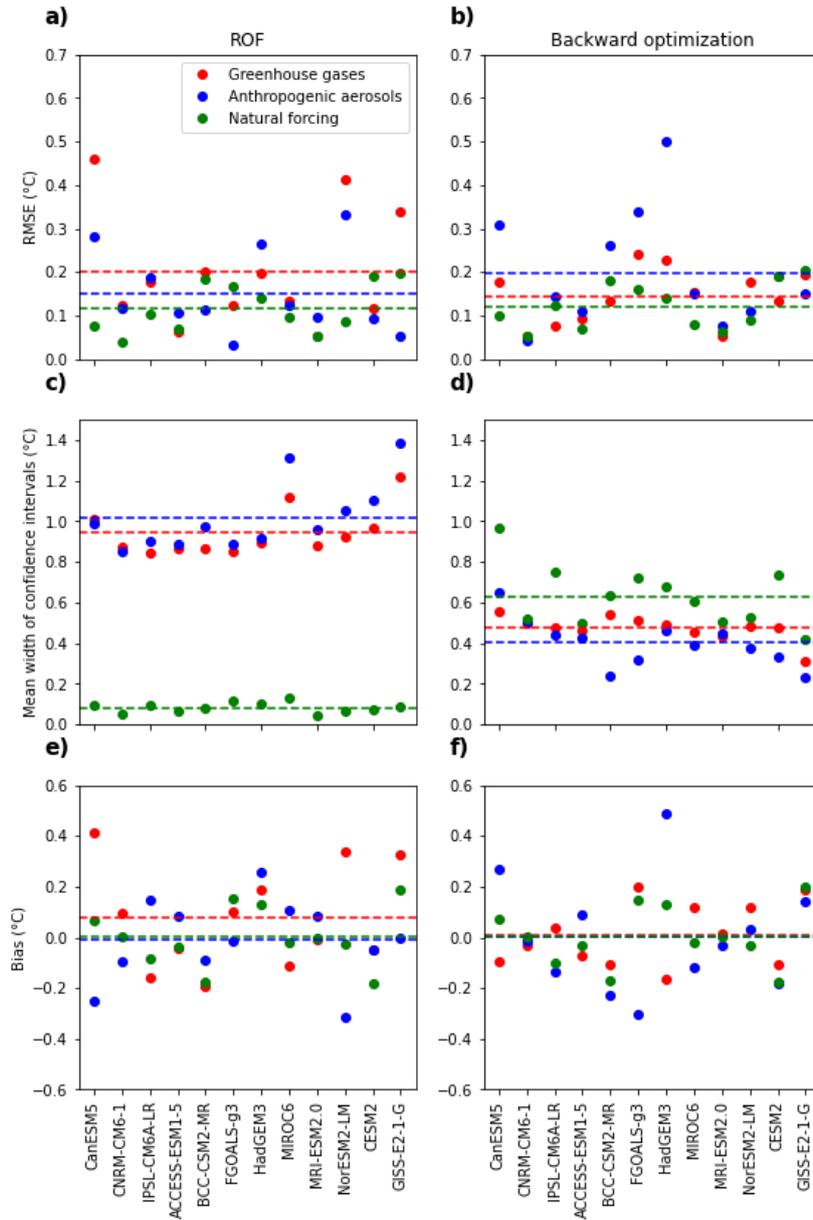
The two methodologies show a monotonic warming induced by the greenhouse gases that intensified in the 1970's for all climate models. The cooling effect of anthropogenic aerosols is also consistent for both methods, with an intensified cooling in the 1970's, also known as global dimming (Wild, 2009), followed by a stabilization in the 2000's. Lastly, the changes attributable to natural forcings are small in both methods, except for the cooling following the major volcanic eruptions.

For the backward optimization, the RMSE is  $0.14^{\circ}\text{C}$ ,  $0.20^{\circ}\text{C}$  and  $0.12^{\circ}\text{C}$  when averaged across the 12 models for the effects of greenhouse gases, anthropogenic aerosols and natural forcing, respectively (see dashed line in Fig. 8a). ROF provides an average RMSE of  $0.20^{\circ}\text{C}$ ,  $0.15^{\circ}\text{C}$  and  $0.12^{\circ}\text{C}$  for these forcings (dashed lines in Fig. 8b), so the errors are similar in both methods. Moreover ROF shows an average positive bias of  $0.09^{\circ}\text{C}$  for greenhouse gases. All other biases for ROF and for backward optimization are almost zero. ROF, therefore seems to over-estimate the effect of greenhouse gases which is not the case of the backward optimization.

However, RMSE and biases are affected by the residual internal variability included in ensemble means especially when only a few members are available. The RMSE and biases are therefore weak indicators for models with few members. The width of the confidence intervals for greenhouse gases and anthropogenic aerosols obtained with the backward optimisation are smaller than those obtained with ROF from the 1970's, while they are larger from 1900 to 1940. Although the uncertainty provided by the confidence intervals of ROF was verified using a perfect model approach in Gillett et al. (2021), some



**Figure 7.** Attributable GSAT in °C calculated for ROF and backward optimization from (black line) a HIST member. The GSAT is decomposed into the attributable changes due to (red line) greenhouse gases; (blue line) anthropogenic aerosols and (green line) natural forcings. For comparison, the ensemble mean of (purple line) GHG, (dark blue line) AER and (beige line) NAT is indicated. Color shades show the 90% confidence intervals of the attributed GSAT.



**Figure 8.** Performances of attribution methods using a perfect model approach. a) RMSE when using ROF for the attributable GSAT anomaly of (red) greenhouse gases, (blue) anthropogenic aerosols, (green) natural forcing. b) Same as a) for the backward optimization. c) Width of the 90% percent confidence intervals in 2000-2014 when using ROF. d) same as c) but for backward optimization e) Time mean difference between the estimated and ensemble mean GSAT attributable to the forcings when using ROF. f) Same as e) for backward optimization. Dashed lines shows average values across the 12 climate models.

468 authors suggested that ROF underestimates such uncertainty because of insufficient con-  
469 sideration in the internal variability (Li et al., 2021; DelSole et al., 2019). This suggests  
470 that the confidence intervals given by the backward optimisation are also underestimated,  
471 and that further improvements would be needed to evaluate them in more details.

472 The width of the confidence intervals for the effect of natural forcing (Fig. 8cd, green  
473 points) is in ROF much lower than this obtained with the backward optimization. This  
474 might be explained by the calculation of the confidence intervals of ROF which is not  
475 adapted to small anomalies (see section 3.4) as obtained for natural forcings. Moreover  
476 we evaluate the uncertainty for the backward optimization by sampling both the inter-  
477 model and internal variability contained in the starting points, so that the confidence  
478 intervals are rather homogeneous in time and for the three forcings. We suggest that both  
479 estimations need to be refined using larger ensembles of simulations. This would allow  
480 a more systematic assessment of the uncertainties using the perfect model approach.

481 Figures 7 and 8 also show that the cooling from anthropogenic aerosols is overes-  
482 timated in FGOALS-g3 in backward optimization results compared to the ensemble mean,  
483 and underestimated in CanESM5 and HadGEM3. It is likely that effect of external forc-  
484 ings in these three models is very different from the other models. For instance, FGOALS-  
485 g3 simulates a negligible effect for the aerosols in AER (see Fig. 1). CanESM5 and HadGEM3  
486 simulate a warming induced by greenhouse gases (see GHG simulation) larger than the  
487 other models, probably associated with the important equilibrium climate sensitivity of  
488 these models. The backward optimization fails to reproduce these singular behaviors,  
489 being mostly governed by the multi-model consensus. The CNN-based method, i.e. the  
490 backward optimization, shows results less variable between models than ROF. The back-  
491 ward optimization attributable changes are more consistent with the multi-model con-  
492 sensus, which is hardly affected by removing the data from one climate model. In con-  
493 trast, in ROF the MMM time series is rescaled with the scaling factors (see section 3.3).  
494 This leads to important errors when the data used as pseudo-observation is taken from  
495 a model with a large sensitivity (see for instance CanESM5).

496 Figure 7 is only based on the use of a single historical simulation for each model.  
497 Therefore, we also investigate if the attributable changes are affected by a modification  
498 of the historical member. The attributable GSAT is estimated with the two methods from  
499 the ten HIST IPSL-CM6-LR member from the test data (see section 3.1.1). The RM-

500 SEs, the biases and the width of confidence intervals are obtained with respect to the  
 501 ensemble mean of the single-forcing simulations of the IPSL-CM6-LR model (Fig. S2).  
 502 Backward optimization presents much less variable results between members than ROF  
 503 in terms of RMSE or bias, except for natural forcing. The amplitude of the confidence  
 504 intervals is slightly increases for the backward optimization compared to ROF. It results  
 505 that backward optimization is less affected by internal variability than ROF.

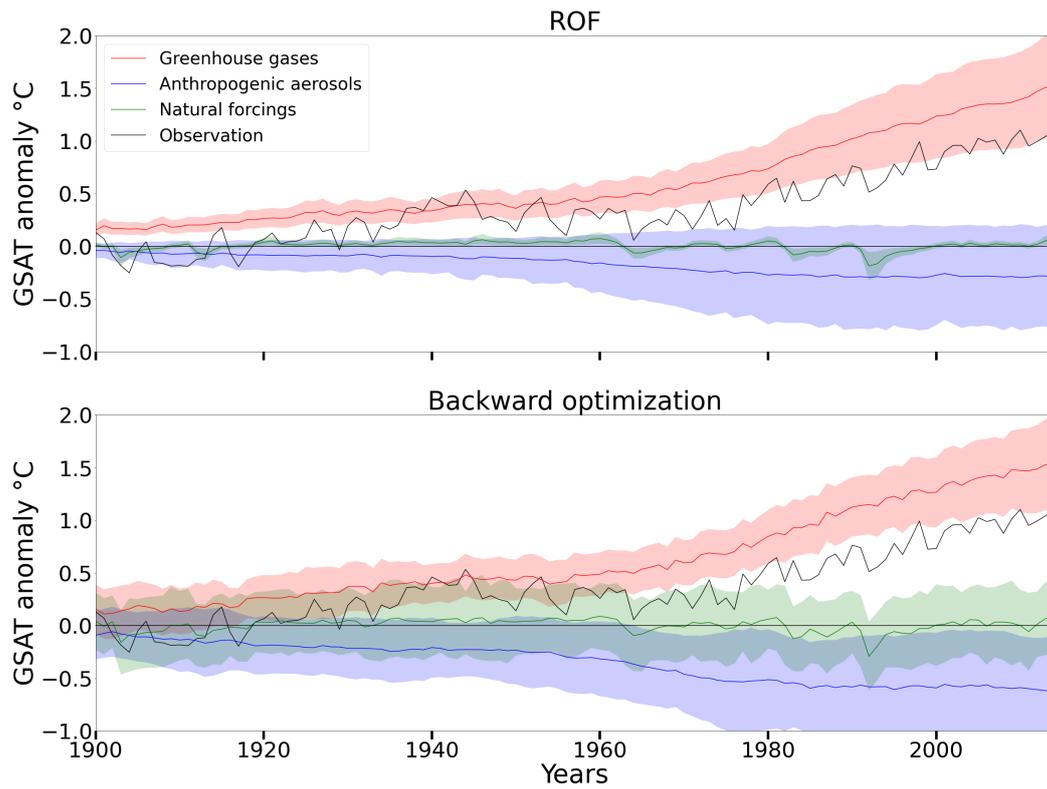
### 506 **4.3 Attribution of the observed GSAT**

507 After studying the performance of ROF and backward optimization for synthetic  
 508 data and in a perfect model approach, we apply both methods to the observed GSAT  
 509 anomalies.

510 The attributable GSAT changes are similar for ROF and backward optimization  
 511 (see Fig. 9). For example, in 2000-2014, ROF provides a GSAT attributable to green-  
 512 house gases of  $1.28^{\circ}\text{C}$  ( 90% confidence interval of  $[0.85^{\circ}\text{C},1.71^{\circ}\text{C}]$ ), while it is  $-0.33^{\circ}\text{C}$   
 513 ( $[-0.80^{\circ}\text{C},0.12^{\circ}\text{C}]$ ) for anthropogenic aerosols and  $0.01^{\circ}\text{C}$  ( $[0.0^{\circ}\text{C},0.02^{\circ}\text{C}]$ ) for natural forc-  
 514 ing. In comparison, backward optimization finds attributable changes of  $1.42^{\circ}\text{C}$  ( $[1.03^{\circ}\text{C},1.80^{\circ}\text{C}]$ ),  
 515  $-0.61^{\circ}\text{C}$  ( $[-1.16^{\circ}\text{C},-0.06^{\circ}\text{C}]$ ) and  $0.02^{\circ}\text{C}$  ( $[-0.33^{\circ}\text{C},0.38^{\circ}\text{C}]$ ), respectively, for these three forc-  
 516 ings. Nevertheless, backward optimization provides more noisy time series and more cool-  
 517 ing during volcanic eruptions. The similarity of the results between ROF and backward  
 518 optimization suggests that the GSAT changes are largely additive as found in Marvel  
 519 et al. (2015) or Shiogama et al. (2013).

520 The attributable changes of the GSAT given by ROF are much comparable to that  
 521 of Gillett et al. (2021) who studied the effect of other forcings (land use and ozone) to-  
 522 gether with the greenhouse gases. Their results for the 2010-2019 decade provide a 5%-  
 523 95% range of the attributable warming of  $[1.2^{\circ}\text{C},1.9^{\circ}\text{C}]$  for greenhouse gases and other  
 524 forcings,  $[-0.7^{\circ}\text{C},-0.1^{\circ}\text{C}]$  for anthropogenic aerosols and  $[0.01^{\circ}\text{C},0.06^{\circ}\text{C}]$  for natural forc-  
 525 ing. We verified that the ROF results shown in Fig. 9 remain similar when we take into  
 526 account other forcings together with the greenhouse gases influence (see Fig. S3).

527 Backward optimization shows a slightly smaller uncertainty for greenhouse gases  
 528 and anthropogenic aerosols than ROF toward the end of the time series, but a larger un-  
 529 certainty range for natural forcings, as found and discussed in section 4.2. We can note  
 530 that the reconstruction of the observations by the backward optimization is by construc-



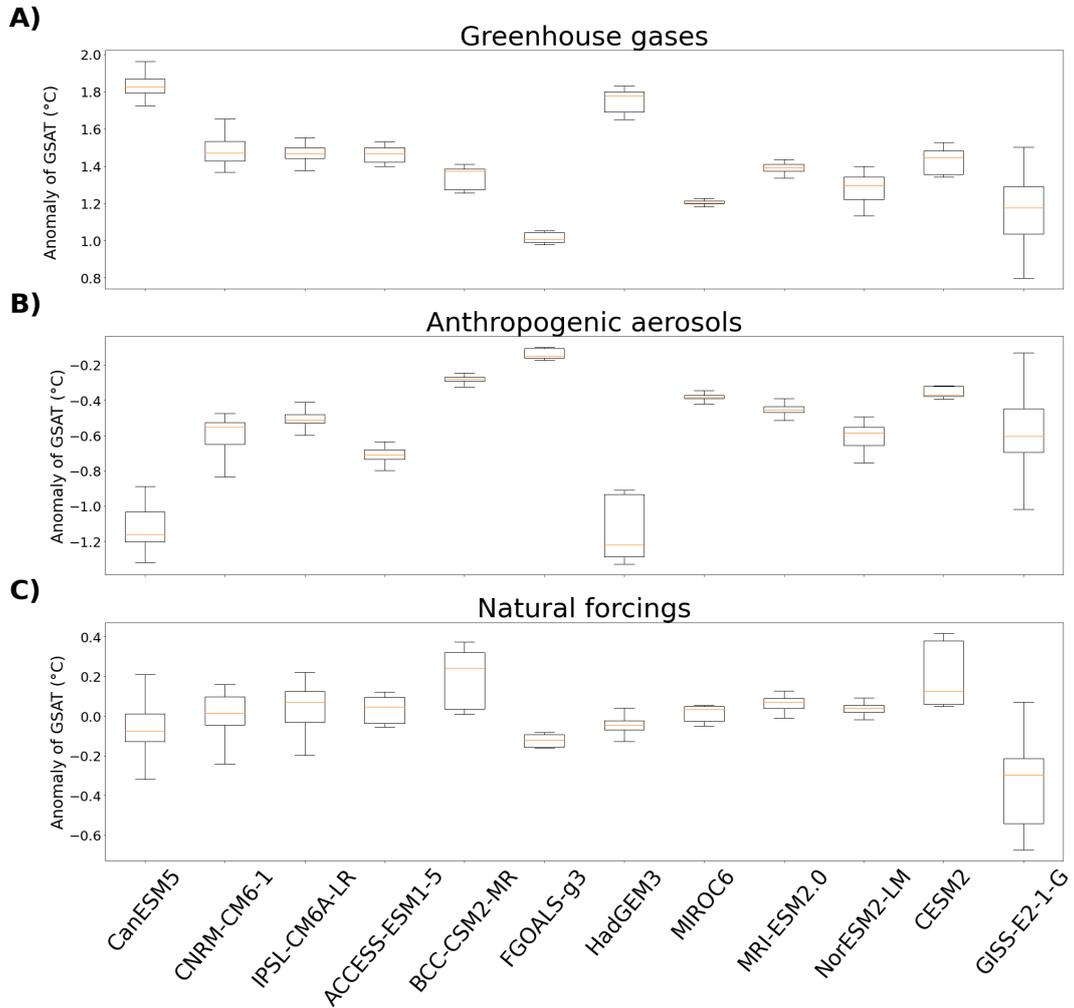
**Figure 9.** (Top) Attributable GSAT anomaly, in °C, as given by ROF for the effect of the (red) greenhouse gases, (green) natural forcings and (blue) anthropogenic aerosols. The black line shows the observed GSAT. The color shade shows the 90% confidence interval. (Bottom) : same as top, but for backward optimization.

531 tion very close to the observations (see Fig. S4) and captures most of the internal vari-  
532 ability contained within the observations.

#### 533 4.4 Focus on the main backward optimization results

534 The backward optimization uncertainties are computed sampling various initial in-  
535 puts. Backward optimization is often used with an all-zeros starting point (Toms et al.,  
536 2020) even if McGovern et al. (2019) have optimised the initial inputs by using coher-  
537 ent starting points as done in the present study. Figure 10 shows the boxplots of the at-  
538 tributable changes in 2000-2014 when using the observations and backward optimiza-  
539 tion, as previously discussed in section 4.3, classified according to the climate models used  
540 for the initial input.

541 The attributable changes produced by the backward optimization are influenced  
542 by the climate model used to generate the initial input. For example, CanESM5 sim-  
543 ulates large warming in response to greenhouse gases (see Fig. 1), probably linked to its  
544 large equilibrium climate sensitivity. When using the outputs of CanESM5 as initial in-  
545 put of the backward optimization, large attributable changes are obtained for both the  
546 greenhouse gases and the anthropogenic aerosols. On the other hand, when using an ini-  
547 tial input from FGOALS-g3 the changes due to the greenhouse gases and the anthro-  
548 pogenic aerosols are small. For each forcing, we analyse the dispersion of the GSAT anoma-  
549 lies over the years 2000-2014 by estimating the mean GSAT attributable to the use of  
550 all starting points for each of the 12 climate models. The variability explained by the  
551 model is calculated by is the standard deviation across these 12 attributable GSAT. The  
552 residual variability which accounts for the internal variability of the starting points is  
553 estimated by the standard deviation of the 1200 attributable GSAT after subtracting for  
554 each time series the average response obtained with their respective climate model. The  
555 standard deviation explained by the model of the starting point is  $0.22^{\circ}\text{C}$  for greenhouse  
556 gases,  $0.29^{\circ}\text{C}$  for anthropogenic aerosols and  $0.14^{\circ}\text{C}$  for natural forcings. The residual  
557 standard deviation is of  $0.06^{\circ}\text{C}$  for the greenhouse gases,  $0.09^{\circ}\text{C}$  for the anthropogenic  
558 aerosols and  $0.1^{\circ}\text{C}$  for the natural forcings. The residual variance therefore is smaller than  
559 this associated with the climate model for each forcing, especially for the greenhouse gases.  
560 The range of attribution results is about  $1^{\circ}\text{C}$  for all forcings, with some particular mod-  
561 els providing attributable anomalies at the head or the tail of the inter-models distribu-  
562 tion when used as starting point. Removing or modifying these outliers to improve the



**Figure 10.** Boxplots of the attributable changes in 2000-2014 when using observation and backward optimization, classified according to the climate model used as initial inputs for (A) the greenhouse gases (B) the anthropogenic aerosols and (C) the natural forcings.

563 backward optimization results have been considered. However, selecting these initial in-  
564 puts may imply a selection of climate models which needs to be associated with a care-  
565 ful investigation of the physical mechanisms (Coquard et al., 2004).

## 566 **5 Discussion and conclusion**

567 We present a method for detection and attribution of climate data based on a back-  
568 ward optimization of a convolutional neural network (CNN). We trained the CNN on  
569 the simulated GSATs obtained from outputs of twelve CMIP6 climate models. We then  
570 performed a backward optimization to estimate the attributable changes. This method-  
571 ology does not assume that the effects of the external forcings are additive. Such addi-  
572 tivity implies that the total changes simulated by the forcings can be obtained by the  
573 sum of the changes due to the individual forcings. The additivity assumption is an im-  
574 portant limitation when focusing on precipitation (Marvel et al., 2015) or at regional scale  
575 (Pope et al., 2020; Deng et al., 2020). We evaluated the effect of internal variability and  
576 model dispersion by using different starting points sampling the simulated distributions.  
577 We compared the results of the CNN backward optimization with those obtained using  
578 the regularized optimal fingerprinting (ROF) (Allen & Stott, 2003; Ribes et al., 2013).  
579 In order to assess the ability of backward optimization to deal with non-additivities in  
580 forcing compared to ROF we used synthetic data, which, unlike GSAT, have a strong  
581 non-additive behavior. In that case, the backward optimization results are more simi-  
582 lar to the true forced effect of the forcings than when using ROF which assumes addi-  
583 tivity. To see if this results can be generalised additional investigations need to be con-  
584 ducted using either different synthetic data or real non-additive climate data, as for in-  
585 stance the precipitation field.

586 We also designed a perfect model approach to evaluate the skill of the two meth-  
587 ods. We successively removed the data of each climate model and used an historical mem-  
588 ber of the removed climate model as pseudo-observation. The attributable changes of  
589 each forcing are then compared to their actual effect simulated in the corresponding en-  
590 semble mean of single-forcing simulations. Backward optimization is found to provide  
591 performances similar to that obtained with ROFs in terms of RMSEs or bias. The con-  
592 fidence intervals of the backward optimization are smaller for greenhouse gases and an-  
593 thropogenic aerosols in the last years of the studied period and much larger for natu-  
594 ral forcings than those obtained by ROF. As the calculation of the uncertainty applied

595 in ROF has been previously shown to be also underestimated (DelSole et al., 2019), this  
596 suggests that backward optimization leads to an even larger underestimation. This might  
597 be linked to the internal variability of the target time series, which is not accounted for  
598 in the neural network-based method. A solution to solve this issue would be to gener-  
599 ate surrogate time series for the backward optimization and repeat the backward opti-  
600 mization. Larger ensemble of single forcing simulations, such as those proposed in the  
601 Large Ensemble Single Forcing Model Intercomparison Project (D. M. Smith et al., 2022),  
602 would also be required to refine of the estimated errors. In addition, the changes attributable  
603 to natural forcings in the backward optimization have a larger uncertainty than the one  
604 of ROF. This is suggested to be an artefact of the estimated uncertainty used, which may  
605 be flawed for small changes. Many aspects of the backward optimization can be improved  
606 in future works. The backward optimization process can also be improved by giving weights  
607 based on the realistic simulation of the interannual to decadal variability. Indeed, the  
608 procedure presented here is designed to produce a close agreement between the recon-  
609 structed time series and the observations (or pseudo-observations). As shown in Fig. S4,  
610 the reconstructed time series, i. e. the image of the CNN using the backward optimiza-  
611 tion results, closely follow the observations. The CNN might instead be designed to only  
612 reproduce the forced component of the anomalies excluding the internal variability un-  
613 related to climate forcings. A better treatment of the initial state could be also inves-  
614 tigated, excluding or penalizing the time series used as initial input when inconsistent  
615 with observations. In addition, giving different weights to each climate models accord-  
616 ing to their performance in reproducing observed features could be considered, such as  
617 the observed GSAT evolution in Ribes et al. (2021).

618 Overall, the attributable changes obtained with the backward optimization are con-  
619 sistent with recent attribution results, as reviewed in Eyring et al. (2020a). This con-  
620 firms the previous detection and attribution results on the GSAT. This study also shows  
621 that neural networks can be used to explore the CMIP databases through the backward  
622 optimization presented here. Such a method could be deployed on other physical vari-  
623 ables, such as precipitation. It could also easily be applied to spatial average instead of  
624 global mean where the non-additivities could be an obstacle. Lastly, a similar method  
625 applied on gridded data could also be considered without major modifications given that  
626 CNNs can easily process images.

## 6 Open Research

### Data Availability Statement

The CMIP6 data is available through the Earth System Grid Federation (Cinquini et al., 2014) and can be accessed through different international nodes. For example,:

<https://esgf-node.ipsl.upmc.fr/projects/esgf-ipsl/>

Codes used in this article for the backward optimization and the figures are from Bône (2023) software available freely at <https://doi.org/10.5281/zenodo.7248662>. The ROF results have been obtained using the Eyring et al. (2020b) software (version 2.9.0) that can be freely found at <https://github.com/ESMValGroup/ESMValTool/releases/tag/v2.9.0>.

### Acknowledgments

We thank three anonymous reviewers for their careful reading and their insightful comments and suggestions. We acknowledge the support of the SCAI doctoral program managed by the ANR with the reference ANR-20-THIA-0003, the support of the EUR IPSL Climate Graduate School project managed by the ANR under the "Investissements d'avenir" programme with the reference ANR-11-IDEX-0004-17-EURE-0006. This work was performed using HPC resources from GENCI-TGCC A0090107403 and A0110107403, and GENCI-IDRIS AD011013295. Guillaume Gastineau was funded by the JPI climate/JPI ocean ROADMAP project (grant number ANR-19-JPOC-003).

### References

- Allen, M. R., & Stott, P. A. (2003). Estimating signal amplitudes in optimal fingerprinting, Part I : Theory. *Climate Dynamics*, *21*(5), 477–491.
- Allen, M. R., & Tett, S. F. (1999). Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, *15*(6), 419–434.
- Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019). Viewing forced climate patterns through an AI lens. *Geophysical Research Letters*, *46*(22), 13389–13398.
- Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., ... others (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010.

- 658 Brajard, J., Santer, R., Crépon, M., & Thiria, S. (2012). Atmospheric correction  
659 of MERIS data for case-2 waters using a neuro-variational inversion. *Remote*  
660 *Sensing of Environment*, 126, 51–61.
- 661 Burger, W., & Burge, M. J. (2009). *Principles of digital image processing: core algo-*  
662 *rithms*. Springer London.
- 663 Bône, C. (2023). *Codes for "Detection and attribution of climate change" [Software]*.  
664 Retrieved from <https://doi.org/10.5281/zenodo.7248662>
- 665 Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.-S., &  
666 Caltabiano, N. (2018). Decadal climate variability and predictability: Chal-  
667 lenges and opportunities. *Bulletin of the American Meteorological Society*,  
668 99(3), 479–490.
- 669 Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., ... others  
670 (2022). Recent advances and applications of deep learning methods in materi-  
671 als science. *npj Computational Materials*, 8(1), 1–26.
- 672 Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., ...  
673 others (2014). The Earth System Grid Federation: An open infrastructure for  
674 access to distributed geospatial data [Dataset]. *Future Generation Computer*  
675 *Systems*, 36, 400–417.
- 676 Coquard, J., Duffy, P., Taylor, K., & Iorio, J. (2004). Present and future surface  
677 climate in the western USA as simulated by 15 global climate models. *Climate*  
678 *Dynamics*, 23(5), 455–472.
- 679 Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D., DuVivier, A., Ed-  
680 wards, J., ... others (2020). The community earth system model ver-  
681 sion 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2),  
682 e2019MS001916.
- 683 DelSole, T., Trenary, L., Yan, X., & Tippett, M. K. (2019). Confidence intervals in  
684 optimal fingerprinting. *Climate Dynamics*, 52(7), 4111–4126.
- 685 Deng, J., Dai, A., & Xu, H. (2020). Nonlinear climate responses to increasing co2  
686 and anthropogenic aerosols simulated by cesm1. *Journal of Climate*, 33(1),  
687 281–301.
- 688 Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues,  
689 L. R. (2013). Seasonal climate predictability and forecasting: status and  
690 prospects. *Wiley Interdisciplinary Reviews: Climate Change*, 4(4), 245–268.

- 691 Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., . . . others  
692 (2020a). Earth System Model Evaluation Tool (ESMValTool) v2.0—an ex-  
693 tended set of large-scale diagnostics for quasi-operational and comprehensive  
694 evaluation of Earth system models in CMIP. *Geoscientific Model Development*,  
695 *13*(7), 3383–3438.
- 696 Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., . . . Zimmer-  
697 mann, K. (2020b). *Earth System Model Evaluation Tool (ESMValTool) v2.0 –*  
698 *an extended set of large-scale diagnostics for quasi-operational and comprehen-*  
699 *sive evaluation of Earth system models in CMIP [Software]* (Vol. 13) (No. 7).  
700 Retrieved from [https://github.com/ESMValGroup/ESMValTool/releases/](https://github.com/ESMValGroup/ESMValTool/releases/tag/v2.9.0)  
701 [tag/v2.9.0](https://github.com/ESMValGroup/ESMValTool/releases/tag/v2.9.0)
- 702 Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., &  
703 Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project  
704 Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model*  
705 *Development*, *9*(5), 1937–1958.
- 706 Eyring, V., Gillett, N., Rao, K. A., Barimalala, R., Parrillo, M. B., Bellouin, N., . . .  
707 Zhu, B. (2021). Human Influence on the Climate System. In *Climate Change*  
708 *2021: The Physical Science Basis. Contribution of Working Group I to the*  
709 *Sixth Assessment Report of the Intergovernmental Panel on Climate Change.*  
710 *Cambridge University Pres.*
- 711 Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., & Rousseau, F.  
712 (2021). Learning variational data assimilation models and solvers. *Journal of*  
713 *Advances in Modeling Earth Systems*, *13*(10), e2021MS002572.
- 714 Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Inter-  
715 pretable deep learning for spatial analysis of severe hailstorms. *Monthly*  
716 *Weather Review*, *147*(8), 2827–2845.
- 717 Gillett, N. P., Kirchmeier-Young, M., Ribes, A., Shiogama, H., Hegerl, G. C.,  
718 Knutti, R., . . . others (2021). Constraining human contributions to ob-  
719 served warming since the pre-industrial period. *Nature Climate Change*, *11*(3),  
720 207–212.
- 721 Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., . . .  
722 Tebaldi, C. (2016). The detection and attribution model intercomparison  
723 project (DAMIP v1.0) contribution to CMIP6. *Geoscientific Model Develop-*

- 724 *ment*, 9(10), 3685–3697.
- 725 Good, P., Lowe, J. A., Andrews, T., Wiltshire, A., Chadwick, R., Ridley, J. K., ...  
726 others (2015). Nonlinear regional warming with increasing CO2 concentrations.  
727 *Nature Climate Change*, 5(2), 138–142.
- 728 Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- 729 Gulev, S., Thorne, P., Ahn, J., Dentener, F., Domingues, C., Gerland, S., & Vose,  
730 R. (2021). Changing state of the climate system. In climate change 2021: The  
731 physical science basis. Contribution of working group I to the sixth assessment  
732 report of the intergovernmental panel on climate change.
- 733 Gupta, A. S., Jourdain, N. C., Brown, J. N., & Monselesan, D. (2013). Climate Drift  
734 in the CMIP5 Models. *Journal of Climate*, 26(21), 8597 - 8615. doi: 10.1175/  
735 JCLI-D-12-00521.1
- 736 Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO  
737 forecasts. *Nature*, 573(7775), 568–572.
- 738 Hargreaves, J. C. (2010). Skill and uncertainty in climate models. *WIREs Climate*  
739 *Change*, 1(4), 556-564. Retrieved from [https://wires.onlinelibrary.wiley](https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.58)  
740 [.com/doi/abs/10.1002/wcc.58](https://doi.org/10.1002/wcc.58) doi: <https://doi.org/10.1002/wcc.58>
- 741 Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent cli-  
742 mate change. *Journal of Climate*, 6(10), 1957–1971.
- 743 Hawkins, E., Robson, J., Sutton, R., Smith, D., & Keenlyside, N. (2011). Evaluating  
744 the potential for statistical decadal predictions of sea surface temperatures  
745 with a perfect model approach. *Climate dynamics*, 37(11), 2495–2509.
- 746 Hobbs, W., Palmer, M. D., & Monselesan, D. (2016). An energy conservation anal-  
747 ysis of ocean drift in the CMIP5 global coupled models. *Journal of Climate*,  
748 29(5), 1639–1653.
- 749 Irving, D., Hobbs, W., Church, J., & Zika, J. (2021). A Mass and Energy Conser-  
750 vation Analysis of Drift in the CMIP6 Ensemble. *Journal of Climate*, 34(8),  
751 3157 - 3170. doi: 10.1175/JCLI-D-20-0281.1
- 752 Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell,  
753 G. L., ... others (2020). GISS-E2. 1: Configurations and climatology. *Journal*  
754 *of Advances in Modeling Earth Systems*, 12(8), e2019MS002025.
- 755 Kennedy, J. J., Rayner, N. A., Atkinson, C. P., & Killick, R. E. (2019). An  
756 Ensemble Data Set of Sea Surface Temperature Change From 1850: The

- 757 Met Office Hadley Centre HadSST.4.0.0.0 Data Set. *Journal of Geophysical*  
758 *Research: Atmospheres*, 124(14), 7719-7763. Retrieved from [https://](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JD029867)  
759 [agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JD029867](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JD029867) doi:  
760 <https://doi.org/10.1029/2018JD029867>
- 761 Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J.  
762 (2021). 1D convolutional neural networks and applications: A survey. *Me-*  
763 *chanical Systems and Signal Processing*, 151, 107398. doi: [https://doi.org/](https://doi.org/10.1016/j.ymssp.2020.107398)  
764 [10.1016/j.ymssp.2020.107398](https://doi.org/10.1016/j.ymssp.2020.107398)
- 765 Labe, Z. M., & Barnes, E. A. (2021). Detecting climate signals using explainable  
766 AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth*  
767 *Systems*, 13(6), e2021MS002464.
- 768 Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M.,  
769 Pritzel, A., ... others (2022). GraphCast: Learning skillful medium-range  
770 global weather forecasting. *arXiv preprint arXiv:2212.12794*.
- 771 Li, Yu, Y., Tang, Y., Lin, P., Xie, J., Song, M., ... Wang, L. (2020). The flexible  
772 global ocean-atmosphere-land system model grid-point version 3 (FGOALS-  
773 g3): description and evaluation. *Journal of Advances in Modeling Earth*  
774 *Systems*, 12(9), e2019MS002012.
- 775 Li, Zwiers, F., Zhang, X., Li, G., Sun, Y., & Wehner, M. (2021). Changes in Annual  
776 Extremes of Daily Temperature and Precipitation in CMIP6 Models. *Journal*  
777 *of Climate*, 34(9), 3441 - 3460. Retrieved from [https://journals.ametsoc](https://journals.ametsoc.org/view/journals/clim/34/9/JCLI-D-19-1013.1.xml)  
778 [.org/view/journals/clim/34/9/JCLI-D-19-1013.1.xml](https://journals.ametsoc.org/view/journals/clim/34/9/JCLI-D-19-1013.1.xml) doi: [https://doi](https://doi.org/10.1175/JCLI-D-19-1013.1)  
779 [.org/10.1175/JCLI-D-19-1013.1](https://doi.org/10.1175/JCLI-D-19-1013.1)
- 780 Li, Z., Zhang, W., Jin, F.-F., Stuecker, M. F., Sun, C., Levine, A. F., ... Liu, C.  
781 (2020). A robust relationship between multidecadal global warming rate vari-  
782 ations and the Atlantic Multidecadal Variability. *Climate Dynamics*, 55(7),  
783 1945-1959.
- 784 Marvel, K., Schmidt, G. A., Shindell, D., Bonfils, C., LeGrande, A. N., Nazarenko,  
785 L., & Tsigaridis, K. (2015). Do responses to different anthropogenic forcings  
786 add linearly in climate models? *Environ. Res. Lett.*, 10(10), 104010. doi:  
787 [10.1088/1748-9326/10/10/104010](https://doi.org/10.1088/1748-9326/10/10/104010)
- 788 Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., ...  
789 Zhou, B. (2021). 2021: Changing State of the Climate System. In *Climate*

- 790 Change 2021: The Physical Science Basis. Contribution of Working Group I  
791 to the Sixth Assessment Report of the Intergovernmental Panel on Climate  
792 Change. *Cambridge University Press*, 287-422.
- 793 McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Home-  
794 yer, C. R., & Smith, T. (2019). Making the black box more transparent:  
795 Understanding the physical implications of machine learning. *Bulletin of the*  
796 *American Meteorological Society*, 100(11), 2175–2199.
- 797 Meehl, G. A., Hu, A., Santer, B. D., & Xie, S.-P. (2016). Contribution of the In-  
798 terdecadal Pacific Oscillation to twentieth-century global surface temperature  
799 trends. *Nature Climate Change*, 6(11), 1005–1008.
- 800 Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E.,  
801 ... Simpson, I. R. (2021). An Updated Assessment of Near-Surface Temper-  
802 ature Change From 1850: The HadCRUT5 Data Set. *Journal of Geophysical*  
803 *Research: Atmospheres*, 126(3), e2019JD032361. Retrieved from [https://](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032361)  
804 [agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032361](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032361)  
805 (e2019JD032361 2019JD032361) doi: <https://doi.org/10.1029/2019JD032361>
- 806 Neelin, J. D., Battisti, D. S., Hirst, A. C., Jin, F.-F., Wakata, Y., Yamagata, T., &  
807 Zebiak, S. E. (1998). ENSO theory. *Journal of Geophysical Research: Oceans*,  
808 103(C7), 14261–14290.
- 809 Osborn, T. J., Jones, P. D., Lister, D. H., Morice, C. P., Simpson, I. R., Winn, J. P.,  
810 ... Harris, I. C. (2021). Land Surface Air Temperature Variations Across the  
811 Globe Updated to 2019: The CRUTEM5 Data Set. *Journal of Geophysical*  
812 *Research: Atmospheres*, 126(2), e2019JD032352. Retrieved from [https://](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032352)  
813 [agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032352](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032352)  
814 (e2019JD032352 2019JD032352) doi: <https://doi.org/10.1029/2019JD032352>
- 815 O’Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks.  
816 *arXiv preprint arXiv:1511.08458*.
- 817 Pope, J. O., Orr, A., Marshall, G. J., & Abraham, N. L. (2020). Non-additive re-  
818 sponse of the high-latitude Southern Hemisphere climate to aerosol forcing in a  
819 climate model with interactive chemistry. *Atmospheric Science Letters*, 21(12),  
820 e1004. doi: <https://doi.org/10.1002/asl.1004>
- 821 Ribes, A., Planton, S., & Terray, L. (2013). Application of regularised optimal fin-  
822 gerprinting to attribution. Part I: method, properties and idealised analysis.

- 823 *Climate dynamics*, 41(11), 2817–2836.
- 824 Ribes, A., Qasmi, S., & Gillett, N. P. (2021). Making climate projections conditional  
825 on historical observations. *Science Advances*, 7(4), eabc0671.
- 826 Ribes, A., Zwiers, F. W., Azais, J.-M., & Naveau, P. (2017). A new statistical ap-  
827 proach to climate change detection and attribution. *Climate Dynamics*, 48(1),  
828 367–386.
- 829 Richardson, M., Cowtan, K., & Millar, R. J. (2018). Global temperature definition  
830 affects achievement of long-term climate goals. *Environmental Research Let-  
831 ters*, 13(5), 054004.
- 832 Roberts, M. J., Baker, A., Blockley, E. W., Calvert, D., Coward, A., Hewitt, H. T.,  
833 ... others (2019). Description of the resolution hierarchy of the global cou-  
834 pled HadGEM3-GC3. 1 model as used in CMIP6 HighResMIP experiments.  
835 *Geoscientific Model Development*, 12(12), 4999–5028.
- 836 Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., ...  
837 others (2020). Overview of the Norwegian Earth System Model (NorESM2)  
838 and key climate response of CMIP6 DECK, historical, and scenario simula-  
839 tions. *Geoscientific Model Development*, 13(12), 6165–6200.
- 840 Sherwood, S., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Har-  
841 greaves, J. C., ... others (2020). An assessment of Earth’s climate sen-  
842 sitivity using multiple lines of evidence. *Reviews of Geophysics*, 58(4),  
843 e2019RG000678.
- 844 Shiogama, H., Stone, D. A., Nagashima, T., Nozawa, T., & Emori, S. (2013). On the  
845 linear additivity of climate forcing-response relationships at global and conti-  
846 nental scales. *International Journal of Climatology*, 33(11), 2542–2550. Re-  
847 trieved from [https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/](https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3607)  
848 [joc.3607](https://doi.org/10.1002/joc.3607) doi: <https://doi.org/10.1002/joc.3607>
- 849 Smith, Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., ... Michou,  
850 M. (2020). Effective radiative forcing and adjustments in CMIP6 models.  
851 *Atmospheric Chemistry and Physics*, 20(16), 9591–9618.
- 852 Smith, D. M., Gillett, N. P., Simpson, I. R., Athanasiadis, P. J., Baehr, J., Bethke,  
853 I., ... Ziehn, T. (2022). Attribution of multi-annual to decadal changes in the  
854 climate system: The Large Ensemble Single Forcing Model Intercomparison  
855 Project (LESFMIP). *Front. Clim.*, 4, 955414. doi: 10.3389/fclim.2022.955414

- 856 Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P.,  
 857 ... others (2019). The Canadian earth system model version 5 (CanESM5.  
 858 0.3). *Geoscientific Model Development*, *12*(11), 4823–4873.
- 859 Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., ... others  
 860 (2019). Description and basic evaluation of simulated mean state, internal vari-  
 861 ability, and climate sensitivity in MIROC6. *Geoscientific Model Development*,  
 862 *12*(7), 2727–2765.
- 863 Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neu-  
 864 ral networks for the geosciences: Applications to earth system variability. *Jour-  
 865 nal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002.
- 866 van Oldenborgh, G. J., Reyes, F. J. D., Drijfhout, S. S., & Hawkins, E. (2013, mar).  
 867 Reliability of regional climate model trends. *Environmental Research Letters*,  
 868 *8*(1), 014055. Retrieved from [https://dx.doi.org/10.1088/1748-9326/8/1/  
 869 014055](https://dx.doi.org/10.1088/1748-9326/8/1/014055) doi: 10.1088/1748-9326/8/1/014055
- 870 Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M.,  
 871 ... others (2019). Evaluation of CMIP6 deck experiments with CNRM-CM6-1.  
 872 *Journal of Advances in Modeling Earth Systems*, *11*(7), 2177–2213.
- 873 Wild, M. (2009). Global dimming and brightening: A review. *Journal of Geo-  
 874 physical Research: Atmospheres*, *114*(D10). doi: [https://doi.org/10.1029/  
 875 2008JD011470](https://doi.org/10.1029/2008JD011470)
- 876 Wu, T., Lu, Y., Fang, Y., Xin, X., Li, L., Li, W., ... others (2019). The Beijing  
 877 Climate Center climate system model (BCC-CSM): the main progress from  
 878 CMIP5 to CMIP6. *Geoscientific Model Development*, *12*(4), 1573–1600.
- 879 Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neu-  
 880 ral networks: an overview and application in radiology. *Insights into imaging*,  
 881 *9*(4), 611–629.
- 882 Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., ...  
 883 others (2019). The Meteorological Research Institute Earth System Model  
 884 version 2.0, MRI-ESM2. 0: Description and basic evaluation of the physical  
 885 component. *Journal of the Meteorological Society of Japan. Ser. II*.
- 886 Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M.,  
 887 ... Srbnovsky, J. (2020). The Australian earth system model: ACCESS-  
 888 ESM1. 5. *Journal of Southern Hemisphere Earth Systems Science*, *70*(1),

