Détection et attribution du changement climatique à l'aide de réseaux de neurones

Constantin Bone

SORBONNE-UNIVERSITÉ

THÈSE DE DOCTORAT

Détection et attribution du changement climatique à l'aide de réseaux de neurones

---

Auteur:

Constantin BÔNE

Président:

Francis Codron

Rapporteurs:

Bernadette Dorizzi

Ronan Fablet

Examinateur:

Aurélien Ribes

Superviseurs:

Guillaume GASTINEAU

Patrick GALLINARI

Sylvie THIRIA

Thèse soumise pour l'obtention du titre de

docteur de Sorbonne-Université

LOCEAN ISIR

Sorbonne-Université, IRD, CNRS, MNHN

22 septembre 2023

"Rejoice, for I bring you glorious news. God walks among us."

Lectitio Divinitatus

Remerciements

Je tiens à remercier mes directeurs de thèse, Guillaume Gastineau, Sylvie Thiria et Patrick Gallinari pour avoir été une source inépuisable de conversation, à la fois scientifique mais aussi bien plus informelle et détendue.
Je tiens également à remercier mes camarades doctorants Clovis, Luther, Léonard et j'en passe tant... Vous avez rendu ces trois années de thèse plaisantes et m'avez convaincu de me rendre tout les jours au labo (ce qui n'était pas gagné...).
Je remercie également ma famille entière, et en particulier mes parents, mes frères et ma soeur qui m'ont soutenu et ce malgré mes discours incompréhensible durant ces trois dernières années.
Enfin je tiens particulièrement à remercier mes plus proches amis, Benjamin, Baltheus, Charles, Gauvain et surtout Pax, vous avez été tout ce temps une grande source de joie et de réconfort.

Sommaire

Remerciements **v**

1 Introduction **1**

2 Concepts et méthodes **7**

2.1 La variabilité interne et forcée 7

2.1.1 Observations de la température de surface 7

2.1.2 Mécanismes du changement climatique 9

Bilan radiatif 9

Forçage radiatif 11

Rétroactions climatiques 11

2.1.3 Modes de variabilité interne 15

La variabilité multidécennale Atlantique 15

Oscillation australe 17

2.2 Modélisation climatique 18

2.2.1 Principe de la modélisation climatique 18

2.2.2 Simulations issues des exercices d'inter-comparaison . 21

Simulations utilisées 23

Simulations historical 23

Simulation Pi-Control 25

Simulations DAMIP 25

2.2.3 Variabilité interne et forcé dans les modèles climatiques 26

2.3 Méthodes pour séparer la variabilité interne et forcée 28

2.3.1 Contexte général 28

2.3.2 Méthodes de séparation de la variabilité interne et forcé 29
Tendances 29
Décomposition multidimensionnelle d'ensemble empirique 29
Analyse des composants basses fréquences 30
Méthodes d'échelonnage 30
2.3.3 Étude la variabilité interne et forcé 31
Tendances du réchauffement global 31
Signal temporel de la variabilité multidécennal de l'Atlantique 31
2.4 Méthodes de détection et attribution 34
2.4.1 Contexte général 34
2.4.2 Empreintes optimisées 34
2.4.3 Méthode de contraintes par les observations 39
2.4.4 Autres méthodes 40
2.4.5 Attribution à l'échelle globale 41
2.5 Réseau de neurones 44
2.5.1 Définition 44
2.5.2 Entraînement d'un réseau de neurones 45
2.5.3 Réseaux perceptron multicouches 48
2.5.4 Réseaux convolutionnels 49
2.5.5 Intelligence artificielle explicable 51

3 Séparation de la variabilité interne et forcé 55
3.1 Introduction 55

4 Attribution de la température globale de l'air en surface 103
4.1 Introduction 103
4.2 Article de *Environmental Data Sciences* 103
4.3 Article de *Journal of Advances in Modelling Earth Systems* 117

5 Conclusion et perspectives **167**

5.1 Conclusion 167

5.2 Perspectives 170

Bibliographie **175**

Chapitre 1

Introduction

Le changement climatique est une des problématiques incontournables du 20ème et 21ème siècle. Ses conséquences se manifestent par une augmentation de la température (Hansen et al., 2006), une augmentation de la fréquence et l'intensité d'événements extrêmes comme les vagues de chaleur (Luber and McGeehin, 2008), les cyclones tropicaux (Walsh et al., 2016) ou les inondations (Hirabayashi et al., 2013), une accélération du cycle de l'eau global (Allan et al., 2020), des phénomènes de mousson (Zhisheng et al., 2015) ou des sécheresses (Mukherjee et al., 2018). Ces différents phénomènes peuvent entraîner des changements dans les écosystèmes (Walther, 2010), un effondrement de la biodiversité (Bellard et al., 2012) ou une montée du niveau de la mer (Mimura, 2013).

Le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), un organisme de l'ONU chargé de faire des revues telles que les *Assessment reports* (AR) de l'état du consensus scientifique des causes, fonctionnement et conséquences du changement climatique, a publié des rapports détaillés à ce sujet. Le premier (AR1) en 1990 et le dernier (AR6) en 2021. Dans ce dernier, il a identifié 127 risques pour les sociétés humaines dus aux conséquences du changement climatique. Parmi ces risques, on retrouve l'accès à des sources d'eau potable, la baisse de la production agricole, le développement de maladies infectieuses, les dommages humains et matériels durant

les évènements climatiques extrêmes, etc. Une compréhension du changement climatique est donc nécessaire pour trouver et mettre en place des politiques d'adaptation et de mitigation efficaces.

On définit par climat la distribution statistique des différentes variables physiques (température, pression atmosphérique, vents, humidité, précipitations, etc.) décrivant les différentes composantes du système climatique. Ces composantes sont les surfaces terrestres, l'atmosphère, l'océan et la cryosphère.

La dynamique propre de ces composantes ou leurs interactions (Cassou et al., 2018) affecte le système climatique. Cela donne une première source de variabilité au climat dite « interne » à celui-ci. Les effets de cette variabilité peuvent être importants à toutes les échelles de temps. Ils peuvent également être à impact local ou global. Des exemples de cette variabilité interne sont les phénomènes "El Niño" et "La Niña" de l'oscillation Sud El Niño (ENSO pour *El Niño – Southern oscillation*) ou bien la variabilité décennale du Pacifique et de l'Atlantique. Ces phénomènes peuvent affecter la température globale de quelques dixièmes de degrés (Meehl et al., 2016) et provoquent des modifications des précipitations, de la circulation atmosphérique, ou de la fréquence des évènements extrêmes dans de nombreuses régions du monde.

Un changement climatique peut être causé par des éléments dits « externes » aux différentes composantes du système climatique. Ces conditions limites externes au système climatique se nomment « forçages » et ils peuvent être d'origine naturelle ou humaine. La variabilité du climat due aux forçages est dite « forcée ». Les principaux forçages dus à l'activité humaine sont les émissions de gaz à effet de serre, le rejet d'aérosols, l'ozone stratosphérique et le changement dans l'utilisation des terres. Ils sont dus à différentes activités humaines comme l'industrie, les transports, l'agriculture, la déforestation, etc. Les principaux forçages naturels sont le rejet d'aérosols naturels lors d'éruptions volcaniques et la variation de rayonnement solaire reçu par la Terre, par exemple, lors d'éruptions solaires ou de changement dans les

paramètres astronomiques.

La détection et l'attribution du changement climatique, est abordé dans tous les rapports du GIEC. La détection du changement climatique vise à démontrer l'existence d'une variabilité forcée et donc que les changements climatiques excèdent ceux provenant de la variabilité interne. L'attribution du changement climatique consiste à isoler le rôle de ces forçages dans le changement climatique observé. L'étude de la détection et attribution du changement climatique est fondamentale, car elle permet d'établir la responsabilité de l'Homme dans celui-ci. L'étude de ce problème permet de comprendre les causes du changement climatique passé. Cette compréhension est importante afin de pouvoir prévoir l'évolution du changement climatique (Huggel et al., 2015; Marjanac et al., 2017; Frame et al., 2020) en évaluant les conséquences possibles des activités humaines présentes et futures. Ainsi la détection et attribution permettent donc de créer des politiques d'adaptation des impacts du changement climatique ou de mitigation de celui-ci, mais également de les évaluer (Nauels et al., 2019; Banerjee et al., 2020). Un sous-problème de l'attribution du changement climatique et également étudié depuis les premiers rapports du GIEC est celui de la séparation de la variabilité interne et forcée dans les observations (Harzallah and Sadourny, 1995; Hawkins and Sutton, 2009; Ting et al., 2009; Solomon et al., 2011; Deser et al., 2014; Frankcombe et al., 2015). Ce problème peut être vu comme une version plus complexe de celui de la détection du changement climatique, car il permet de caractériser entièrement la variabilité forcée et non plus seulement son existence. Son étude permet une meilleure compréhension de la variabilité interne ou forcée. Il peut servir d'étape préalable à d'autres études de détection et d'attribution par un prétraitement des données utilisées ou par une diminution de la quantité de données utilisées venant des modèles climatiques.

L'apprentissage automatique (*machine learning* en anglais) est une technique de traitement de données, sous-branche de l'intelligence artificiel, qui vise à permettre à un ordinateur d'apprendre à partir de données à réaliser des tâches sans être explicitement programmé pour cela. Les réseaux de neurones sont un des types d'algorithmes les plus emblématiques du *machine learning*. Ils sont utilisés pour traiter les données et apprendre des relations à partir d'elles et permettent de traiter de grandes quantités de données de manière efficace en termes de temps et de puissance de calcul. Leur polyvalence en a fait un des outils les plus utilisés en statistique depuis plusieurs décennies (Choudhary et al., 2022) dans de nombreux domaines scientifiques. L'analyse d'images est l'un des domaines où les réseaux de neurones ont connu le plus de succès (Egmont-Petersen et al., 2002). Par exemple, l'une des applications les plus connues du traitement des images par les réseaux neuronaux est le débruitage des images (Ilesanmi and Ilesanmi, 2021; Tian et al., 2020) qui consiste à restaurer des images bruitées.

Les réseaux de neurones sont des outils relativement nouveaux en géosciences, bien que leur utilisation soit en plein développement, et ce, dans les sciences marines (Malde et al., 2020), les sciences de la terre (Bergen et al., 2019) ou en météorologie (Barnes et al., 2019; Boukabara et al., 2019; Reichstein et al., 2019). Ils ont été utilisés en climatologie, pour de nombreux problèmes (Ham et al., 2019; Gagne II et al., 2019; Labe and Barnes, 2021). Cependant, à notre connaissance, ils n'ont jamais été utilisés pour les problèmes de la détection et attribution du changement climatique. L'objectif de cette thèse fut donc d'utiliser ces réseaux de neurones et des techniques d'interprétation de ceux-ci pour la détection et l'attribution du changement climatique.

Dans un premier temps, nous présenterons les mécanismes de la variabilité interne et forcée, les différentes méthodes existantes pour les problèmes de la détection et attribution du changement climatique et de la séparation entre la variabilité interne et forcée et leurs principaux résultats. Nous y

présenterons également le fonctionnement général des réseaux de neurones. Dans un second temps, nous étudierons la séparation de la variabilité interne et forcée avec l'utilisation de réseaux de neurones. Dans un troisième temps, nous aborderons la détection et attribution du changement climatique à l'échelle globale avec l'utilisation de réseaux de neurones et de méthodes d'interprétation. Finalement, nous ferons une synthèse du travail accompli et nous discuterons de ses limites et de ses perspectives.

Chapitre 2

Concepts et méthodes

2.1 La variabilité interne et forcée

2.1.1 Observations de la température de surface

Afin de décrire l'évolution passée du climat, nous nous intéressons à l'évolution de la température globale de la surface de l'air. Nous nous sommes concentrés sur la température de l'air en surface, car il s'agit de la variable physique la plus observée et illustrative du changement climatique.

L'observation de la température commence dès l'invention du thermomètre au 17ème siècle. Cependant, l'établissement des premiers réseaux de stations de surface à grande échelle commence au 19ème siècle (Maury, 1849, 1855, 1860). L'évolution des différents moyens techniques d'observations permet d'avoir des mesures de plus en plus précises. Ces mesures se font grâce à des milliers de stations météorologiques sur les terres. Pour la mer, les mesures d'observations se font majoritairement et historiquement sur la température de surface (SST pour *sea surface temperature*) par bateaux et bouées.

Les données SST et de température de l'air au-dessus des surfaces terrestres sont traitées et mises sous la forme de données grillées par différentes agences météorologiques telles que la *National Aeronautics and Space Administration* (NASA), la *National Oceanic and Atmospheric Administration* (NOAA) ou *Berkeley Earth* pour ne citer qu'eux. Chaque groupe agrège les mesures

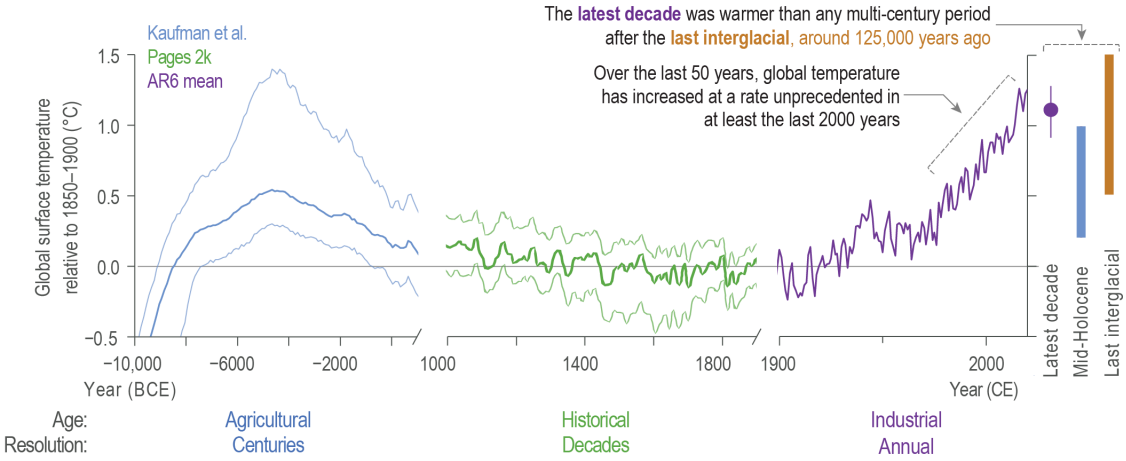

FIGURE 2.1: Anomalies de GMST en °C par rapport à la période 1850-1900 pour différentes périodes temporelles (tiré de Gulev et al., 2021)

et applique des ajustements (par exemple pour traiter les îlots de chaleur urbains), et traite les inégalités d'échantillonnage spatiales et temporelles. Le traitement des zones sans mesure est particulièrement incertain. Des algorithmes de krieggeage ont par exemple été proposés (Cowtan and Way, 2014; Kadow et al., 2020). Les bases de données sont ainsi grillées spatialement sur toute la surface de la Terre. Pour obtenir des valeurs globales ou régionales il faut moyenner les points de grille concernés en leur donnant un poids proportionnel à l'aire de la surface qu'ils décrivent. Ces groupes obtiennent ainsi des estimations de la température globale de surface (GMST pour *global mean surface temperature*). La GMST diffère de la température de surface globale de surface de l'air (GSAT pour *global surface air temperature*) car elle utilise les valeurs de SST au contraire de la GSAT qui utilise la température de surface de l'air pour tous les points de grille. La GSAT n'est pas disponible pour les observations à cause de la nature même des observations disponibles de l'océan. Ces deux variables physiques ont des différences, mais nous ne nous n'y sommes pas intéressés dans cette thèse. Nous avons utilisé la GSAT, car les modèles climatiques, décrits plus tard, ne fournissent de manière directe que la GSAT. Afin d'obtenir de la GMST des

observations la GSAT correspondante, nous multiplions la première par une valeur numérique comme fait dans Gillett et al., 2021. Cette valeur est le ratio moyen de l'anomalie de GSAT dans la période 2010-2019 (avec 1850-1900 comme période de référence) simulé dans différents modèles climatiques de référence comparée à la GMST des observations, partiellement masquées, de la base de données HadCRUT4.

Le GIEC a compilé les données disponibles sur la température de la surface du globe et elles montrent toutes que la GMST a augmenté de 0,85 °C de 1880 à 2012. Il a également constaté que chacune des trois décennies suivant 1980 a été successivement plus chaude à la surface de la Terre que toute décennie précédente depuis 1850 (Gulev et al., 2021). On illustre dans la Fig. 2.1 l'évolution observée et reconstruite de la GMST pour différentes époques historiques. Il y a une tendance au réchauffement du globe depuis le début du vingtième siècle et qui s'est accéléré depuis les années soixante. Il est de l'ordre de 1,07 °C pour la période 2010-2019 par rapport à la période 1850-1900.

2.1.2 Mécanismes du changement climatique

Bilan radiatif

Le climat est un système interconnecté dirigé par l'énergie solaire. Les travaux de Fourier, 1822, ont formulé le principe d'équilibre radiatif. Ce principe est fondamental pour déterminer la température de surface de la Terre et montre l'importance de l'atmosphère pour garder la chaleur terrestre. En effet, il dicte qu'il doit y avoir un équilibre entre l'énergie entrant dans le système climatique et sortant de celui-ci si le système est à l'équilibre. La répartition des flux énergétiques dans le système climatique pour atteindre cet équilibre est appelée bilan radiatif de la Terre. Les forçages peuvent cependant provoquer

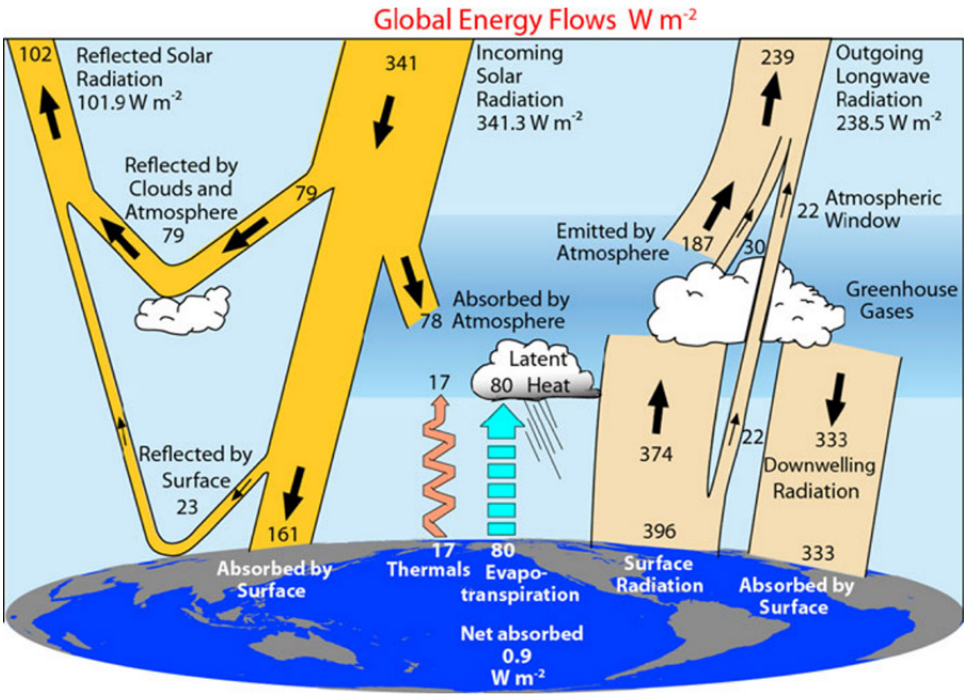

FIGURE 2.2: Moyenne annuelle globale du bilan énergétique de la Terre pour la période allant de mars 2000 à mai 2004 (W/m^2). Les flèches larges indiquent schématiquement les flux d'énergie en fonction de leur importance. À partir de Trenberth et al., 2009.

un déséquilibre de ce budget énergétique (Trenberth et al., 2014; Von Schuckmann et al., 2016). Ce déséquilibre représente une métrique importante du changement climatique global (Hansen et al., 2005; Von Schuckmann et al., 2020).

La figure 2.2 illustre un tel budget pour la période mars 2000 à mai 2004. L'équilibre n'a pas été atteint pour cette période, la Terre a donc réalisé un gain énergétique de l'ordre de 0.9 $W.m^{-2}$ sur cette période, principalement dû à l'effet de l'augmentation de la concentration des gaz à effet de serre. Cet effet de serre diminue le rayonnement tellurique émis par la Terre vers l'espace. La température doit donc augmenter afin de combler ce déficit de rayonnement émis. Cela, par un phénomène nommé rétroaction de Planck et détaillé plus tard, conduit à une augmentation du rayonnement terrestre qui conduit à un refroidissement relatif de la terre qui ne compense cependant pas le réchauffement dû à l'effet de serre.

Forçage radiatif

Le forçage radiatif effectif quantifie le gain énergétique net du système climatique que provoque la modification d'une variable atmosphérique en maintenant les autres inchangées. Il s'agit d'un calcul théorique qui permet de quantifier l'impact énergétique d'un forçage.
Par exemple, les gaz à effet de serre modifient le budget énergétique par effet de serre : le rayonnement émis par la terre est capté en partie par ces gaz et est réémis vers la Terre, ce qui contribue à la réchauffer.
Les aérosols ont un impact direct sur le budget énergétique de la Terre en changeant la diffusion et l'absorption de la radiation solaire entrante. Ils ont également un effet indirect en affectant la microphysique des nuages, et donc leurs propriétés en servant de noyaux de condensation ou de glaciation pour la formation de gouttelettes d'eau nuageuse. Leurs effets produisent globalement un refroidissement de la Terre pour la plupart des types d'aérosols.
L'utilisation des terres modifie l'albédo de la surface de la Terre et les flux turbulents. L'ozone stratosphérique absorbe quant à lui une partie du rayonnement solaire dans la stratosphère.
Le forçage radiatif effectif ne permet cependant pas d'évaluer l'impact climatique des forçages. En effet, un forçage ne change jamais seul : son existence provoque des changements du système physique qui peuvent à leur tour affecter le climat. Ce phénomène est appelé une rétroaction climatique. Nous allons dans la suite présenter les principales rétroactions affectant le système climatique.

Rétroactions climatiques

Les rétroactions climatiques sont des interactions dans laquelle une perturbation climatique provoque une seconde perturbation qui influe à son tour sur la perturbation initiale. Une rétroaction positive accentue la perturbation

initiale, une rétroaction négative l'atténue. La perturbation initiale provient typiquement d'un forçage radiatif.

La présence de ces boucles de rétroactions complexifie considérablement les problèmes climatiques, car elle fait de la Terre un système physique complexe où tout phénomène peut affecter tous les autres (Roe and O'Neal, 2009). Nous allons présenter ici les principales rétroactions du système climatique lié à l'augmentation de la température.

La rétroaction de la vapeur d'eau est une rétroaction positive se provoquant quand la Terre se réchauffe. L'atmosphère se réchauffant, la pression de vapeur saturante augmente également, or la vapeur d'eau étant un gaz à effet de serre, cela conduira à une nouvelle augmentation de la température. Cette rétroaction a une grande importance et est capable de doubler l'augmentation de température qui aurait lieu sans cette rétroaction.

La rétroaction des nuages est la plus incertaine. En effet, les nuages ont un double effet de réchauffement et de refroidissement du climat. Le réchauffement est dû au fait que les nuages renvoient une partie du rayonnement terrestre vers la surface. L'effet de refroidissement est dû au fait que les nuages réfléchissent également la lumière du soleil en émettant un rayonnement infrarouge vers l'espace. Le bilan net de température dépend en grande partie de l'attitude du nuage. Un nuage élevé retiendra plus de chaleur et favorisera donc le réchauffement au contraire d'un nuage bas qui favorisera le refroidissement. Une augmentation de température peut influer la distribution et l'attitude des nuages, provoquant des rétroactions positives ou négatives. Les nuages étant mal observés avant l'avènement des observations satellites, leur effet est très incertain dans les modèles climatiques, particulièrement au niveau de la microphysique des nuages.

La rétroaction de l'albédo est une rétroaction positive qui se produit quand la glace fond dû à l'augmentation de la température. La glace, en effet, réfléchit

une plus grande partie du rayonnement solaire incident que les surfaces continentales ou l'eau liquide. Cela conduit donc à un nouveau réchauffement.
La rétroaction du gradient de température est une rétroaction négative se déroulant dans la troposphère et influant sur l'effet de serre. En effet, l'importance de l'effet de serre dépend du gradient de température de l'atmosphère avec l'altitude. Un réchauffement de la Terre conduira à un affaiblissement de ce gradient vertical de température, ce qui affaiblira mécaniquement l'effet de serre.
La rétroaction de Planck est une rétroaction négative qui concerne le rayonnement thermique de la Terre. La loi physique de Stefan-Boltzmann établit en effet que tout corps émet un rayonnement proportionnel à la puissance quatre de sa température. Ainsi donc, la Terre émet plus de rayonnement quand elle se réchauffe, provoquant ainsi un refroidissement de celle-ci.
Ces rétroactions sont plus ou moins intenses selon les différentes régions du globe. Par exemple, la rétroaction de Planck est relativement plus faible aux pôles alors que les rétroactions de l'albédo et du gradient vertical y sont au contraire plus fortes. Un exemple de l'importance de rétroactions est celui de l'amplification polaire (Holland and Bitz, 2003; Pithan and Mauritsen, 2014) qui est un phénomène où les changements de températures de surface à hautes latitudes excèdent le réchauffement global dû à l'ensemble de ces rétroactions (Pithan and Mauritsen, 2014; Goosse et al., 2018; Dai and Bloecker, 2019; Feldl et al., 2020). La Figure 2.3 montre la tendance de réchauffement observé par décade dans la période 1900-1980 et 1981-2020. Nous pouvons y voir que les tendances sont effectivement bien plus marquées dans la période récente aux hautes latitudes où elles atteignent 0,8 °C par décade en 1981-2020 contre 0,15 °C sur le reste du Globe.

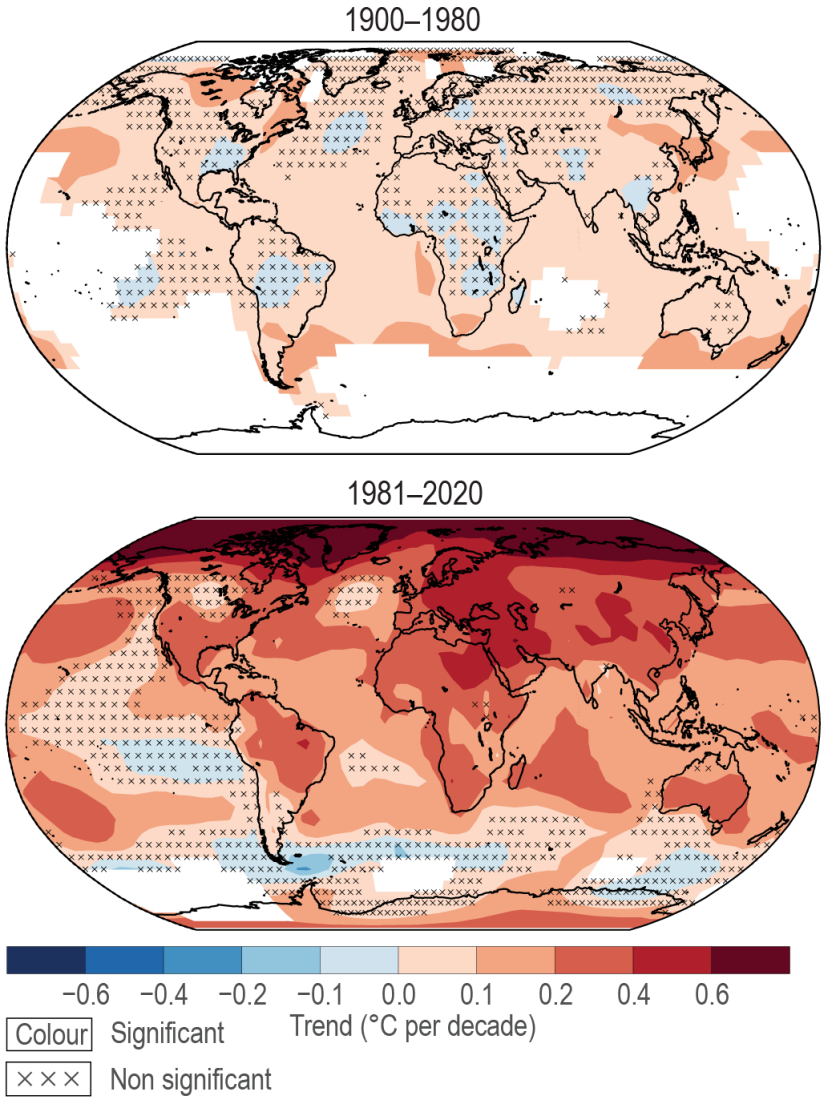

FIGURE 2.3: Tendances des températures dans les observations HadCRUTv5 par décade pour la période 1900-1980 et 1981-2020. Les croix montrent les régions dont les tendances ne sont pas significatives. Tiré de Gulev et al., 2021

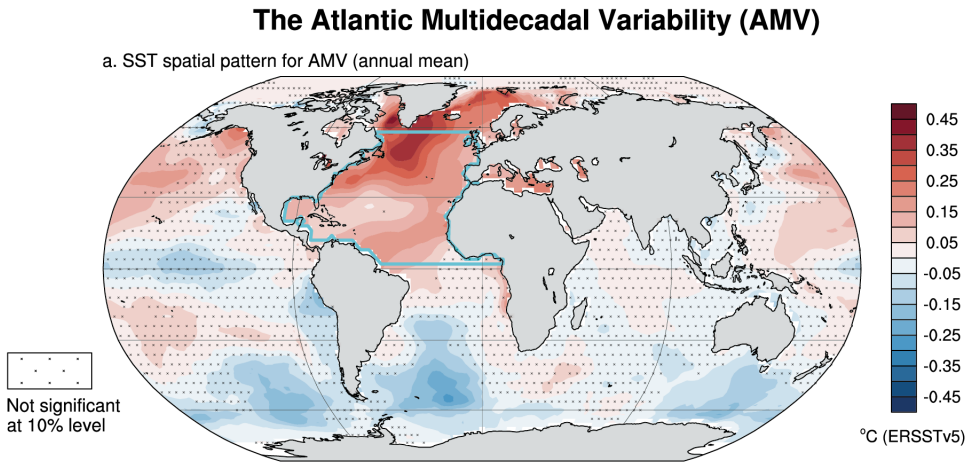

FIGURE 2.4: Anomalies de SST dû à l'AMV après un filtrage passe-bas de 10 ans. Les pointillés représentent les points de grille ne montrant pas de changement significatifs avec un intervalle de confiance de 90%. Tiré de Masson-Delmotte et al., 2021a

2.1.3 Modes de variabilité interne

Les changements du climat sont également dus à des fluctuations internes du système climatique. La variabilité interne présente des structures spatio-temporelles préférentielles. On appelle « modes » ces structures de variabilité du système climatique possédant une structure spatiale, une échelle de temps et parfois une saisonnalité.

Nous présentons ici deux modes : la variabilité multidécennale Atlantique (AMV pour *Atlantic multi-decadal variability*) et l'oscillation australe (noté ENSO pour *El Niño – Southern oscillation*).

La variabilité multidécennale Atlantique

La variabilité multidécennale de l'Atlantique (AMV) décrit les fluctuations lentes et à grande échelle observées d'une décennie à l'autre dans une variété d'enregistrements instrumentaux et de reconstructions par proxy sur l'ensemble de l'océan Atlantique Nord et des continents environnants.

L'AMV se caractérise par des variations des anomalies de la SST à l'échelle du bassin, avec une période de l'ordre de 70 ans. L'AMV est liée aux fluctuations décennales à multicentennales de la circulation méridienne de retournement de l'Atlantique (AMOC pour *Atlantic meridional overturning circulation*) ainsi qu'au transport méridien océanique de chaleur/salinité associé (Zhang, 2017).

La phase positive de l'AMV se caractérise par un réchauffement anormal sur l'ensemble de l'Atlantique Nord, dont l'amplitude est la plus forte dans le gyre subpolaire et le long des zones de marge de la glace de mer dans la mer du Labrador et la mer de Groenland/Barents (environ +0,5 °C) et, dans une moindre mesure, dans le bassin subtropical de l'Atlantique Nord. Les anomalies de température spatiales durant la phase positive de l'AMV sont illustrées dans la Figure 2.10. Ces anomalies sont obtenues après utilisation d'un filtre passe-bas avec une période de coupure à 10 ans. On peut y voir que les anomalies de températures sont particulièrement marqués dans la mer du Labrador.

L'AMV est un facteur clé des anomalies de température et de précipitations le long des continents entourant l'Atlantique Nord (Sutton and Hodson, 2005), mais également à distance par le biais de téléconnexions atmosphériques mondiales (par exemple, les moussons (Monerie et al., 2019)).

La prise en compte de l'influence régionale induite par l'AMV est cruciale, car elle agit comme un modulateur des impacts dus aux forçages anthropiques ou naturels. Par exemple, l'AMV peut avoir une empreinte prononcée dans les processus intégrés dans le temps, tels que les flux fluviaux, et pourrait expliquer la plus grande partie de la variance observée dans certaines zones locales spécifiques depuis 1900 (Bonnet et al., 2020). Il s'agit donc d'un phénomène clé pour l'attribution à l'échelle régionale ou les processus des variations climatiques observées dans le passé.

Oscillation australe

ENSO se caractérise par une alternance entre un refroidissement et un réchauffement anormal de la température de surface océanique au niveau du Pacifique central et est. Ces changements coïncident avec des changements dans les vents et la précipitation de la même région (Philander, 1990; Neelin et al., 1998; Wang, 2018). ENSO est le premier mode de variabilité dans à des échelles de temps interannuelles (3 à 8 ans en moyenne) et est considéré comme un phénomène interne du couplage océans-atmosphère. Bien que ce mode soit en grande partie dû à des phénomènes ayant lieu au niveau du Pacifique tropical, il a des impacts au-delà des tropiques et dans de nombreuses autres parties du monde. En effet, ENSO est le principal modulateur de la température globale de surface à des échelles de temps interannuelles (Pan and Oort, 1983; Trenberth et al., 2002). Il est aussi une des sources de prévisibilité du climat à des échelles de temps saisonnières ou plus faiblement à des échelles interannuelles (Philander, 1990; Smith et al., 2012). Les évènements dits El Niño de cette oscillation se caractérisent par des températures de surface de la mer plus chaudes que la normale pour le centre et l'est du Pacifique équatorial, une différence négative de l'anomalie de la pression de surface entre les parties est et ouest de l'océan Pacifique tropical, des vents de surface anormaux vers l'ouest, une augmentation de la couverture nuageuse et des précipitations sur le centre et l'est du Pacifique équatorial central et oriental et les zones terrestres adjacentes. En revanche, les événements La Niña se caractérisent généralement par des températures de surface de la mer plus froides que la normale dans le centre et l'est du Pacifique équatorial, une différence d'anomalie de pression de surface positive entre les parties est et ouest de l'océan Pacifique tropical et des vents de surface anormaux d'est, reflétant une intensification du gradient thermique climatologique est-ouest dans l'océan Pacifique équatorial.

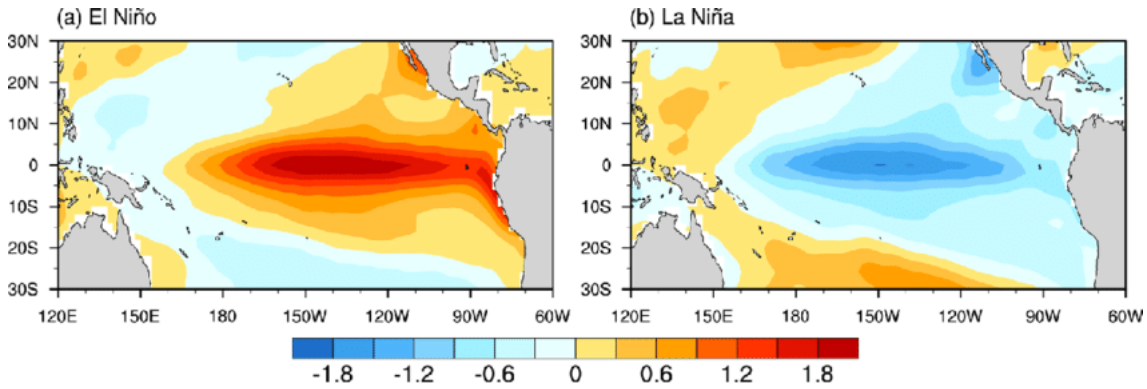

FIGURE 2.5: Anomalies composites de la SST sur l'océan Pacifique tropical pour certains événements (a) El Niño et (b) La Niña durant la période 1949-2015. Tiré de Li et al., 2018

La Figure 2.5 montre le motif spatial moyen des anomalies de SST d'El Niño et La Niña sur la période 1949 - 2015 obtenues par régression. Nous pouvons effectivement y voir sur une zone spatiale quasi identique un réchauffement marqué de 1,8 °C lors des épisodes El Niño et un refroidissement équivalent durant les évènements La Niña.

La complexité et l'interconnexion des différents composantes et effets, tels que les modes de variabilité internes, du système climatique a rendu nécessaire le développement de simulations numériques de celui-ci.

2.2 Modélisation climatique

2.2.1 Principe de la modélisation climatique

La compréhension du système climatique mondial exige autant une compréhension théorique que des mesures des principaux facteurs et forces qui régissent le transport de l'énergie et de la masse (air, eau et vapeur d'eau) autour du globe, des propriétés physiques de l'atmosphère, de l'océan, de la cryosphère et des surfaces terrestres, ainsi que les nombreuses rétroactions

(positives et négatives) entre ces processus. La modélisation permet aux scientifiques de combiner un large éventail de connaissances théoriques et empiriques issues en particulier de la mécanique des fluides et de produire des estimations du climat sous forme de simulation, du passé, du présent ou du futur (Nebeker, 1995; Edwards, 2011). Ils sont des outils indispensables pour étudier le climat à différentes échelles spatiales et temporelles.

Un modèle climatique est un outil informatique fonctionnant sur des ordinateurs de haute performance (André et al., 2014; Balaji et al., 2017) qui utilise des équations de la mécanique des fluides pour simuler le climat de la Terre en simulant une ou plusieurs des différentes composantes du climat (atmosphère, océan, cryosphère) et leurs interactions. Une de ces équations fondamentales est celle de Navier-Stokes qui décrit le mouvement d'un fluide. Elle est généralement écrite sous la forme d'un système d'équations différentielles aux dérivés partielles. On peut également citer la loi de conservation de la masse et de l'énergie ou bien les équations de Clausius-Clapeyron décrivant les changements d'état d'un corps en fonction de la température.

Leur fonctionnement repose sur une discrétisation spatio-temporelle de ces équations avec une grille spatiale en trois dimensions (Staniforth and Thuburn, 2012). La taille du maillage spatial et temporel utilisé est appelée la résolution du modèle climatique. Plus ce maillage est fin, plus les niveaux de détails du modèle climatique sont élevés, mais plus la capacité de calcul nécessaire est élevée. La résolution de la grille détermine en effet quels processus physiques peuvent être résolus ou lesquels doivent être paramétrés, car se passant à une échelle de temps ou spatial trop fine pour la grille du modèle. La paramétrisation implique de choisir des valeurs numériques appelés paramètres qui régissent ces phénomènes non résolus. Nous illustrons dans la Fig 2.6 le schéma de fonctionnement d'un modèle climatique. On peut voir que, à chaque point de grille, différents phénomènes physiques sont modélisés et leur résultat influent sur les points de grilles voisins.

Les effets des forçages sont aussi transmis aux modèles climatiques comme conditions limites. Il existe plusieurs types de modèles climatiques. Nous avons, lors de cette thèse, utilisé les modèles de circulation générale atmosphère-océans (AOGCM) qui modélisent un grand nombre des composantes climatiques sur tout le globe et leurs couplages. Les composantes modélisées d'un AOGCM sont l'océan, la glace de mer, les surfaces terrestres et l'atmosphère. Les modèles utilisés sont sans cycle du carbone interactif. Le CO2 et le CH4 sont en effet prescrits et paramétré dans l'atmosphère.

La création d'un modèle climatique nécessite de faire un grand nombre de choix dans beaucoup de domaines. Outre la formulation formelle du modèle et la résolution spatiale et temporelle, les phénomènes non résolus doivent être paramétrés. Parmi les phénomènes paramétrés, on peut citer le transfert radiatif, la formation et la vitesse de chute des gouttes de pluie, la convection atmosphérique ou la turbulence dans la couche limite.

Ces paramètres sont choisis pour que les résultats obtenus par le modèle climatique soient consistants avec le climat observé (Hourdin et al., 2017). Tous les modèles utilisent pour cela le climat moyen correspondant à la période récente de 1979-2015, où l'on dispose d'observations en quantité suffisante, notamment depuis qu'il existe des satellites dédiés aux observations météorologiques. Certains modèles utilisent également les tendances historiques.

Ces cibles d'ajustement sont propres pour chaque groupe de modélisation et chaque modèle climatique. Les modèles variant également entre eux pour leur formulation, la résolution des processus physiques et la grille utilisée, cela conduit à une grande diversité de résultats obtenus avec les modèles climatiques. Une illustration de leur différence est la variabilité de la sensibilité climatique d'équilibre qui quantifie le réchauffement du modèle dû à un doublement de la concentration de dioxyde de carbone par rapport à la période préindustrielle une fois que le système climatique modélisé aura

trouvé un nouvel état d'équilibre.

Le fait de fixer la paramétrisation des modèles climatique pour correspondre aux observations et le fait qu'ils partagent un grand nombre de composantes et de paramétrisations (Knutti et al., 2013) a toujours été une des critiques adressées aux modèles climatiques. En effet, la majorité de ces modèles ont été paramétrés pour reproduire le climat passé, (Hourdin et al., 2017) ce qui peut jeter des doutes sur leurs capacités physiques à reproduire des variations non observées. Une question importante est donc de déterminer si les modèles sont adéquats pour diverses questions scientifiques d'importance comme les causes du changement climatique ou l'évolution future du climat (Parker, 2009; Notz, 2015; Winsberg, 2018). Ce problème pousse aussi les scientifiques abordant ces sujets à utiliser non pas un modèle unique, mais un ensemble de modèles climatiques dont leurs paramétrisations respective, modèle physique et grille varient entre eux.

2.2.2 Simulations issues des exercices d'inter-comparaison

Pour permettre l'utilisation d'un ensemble de modèles, des expériences ont été effectuées avec les mêmes conditions limites. Les projets d'inter-comparaison de modèles couplés (*Coupled Model Intercomparison Project*, CMIP, (Meehl et al., 2000; Meehl et al., 2007; Taylor et al., 2012; Eyring et al., 2016) ont été mis en place avec des protocoles expérimentaux précis, des périodes fixes et des variables de sortie prédéfinies. Les modèles adhérant aux standards des différents CMIP forment ainsi des générations de modèles climatiques.
Ces ensembles multimodèles permettent d'échantillonner et de quantifier les incertitudes entre et à l'intérieur des différentes générations de modèles climatiques. Leur usage permet de mieux échantillonner les incertitudes des modèles climatiques (Knutti et al., 2010) même si les modèles présentent toujours des parties communes qui peuvent induire des erreurs systématiques

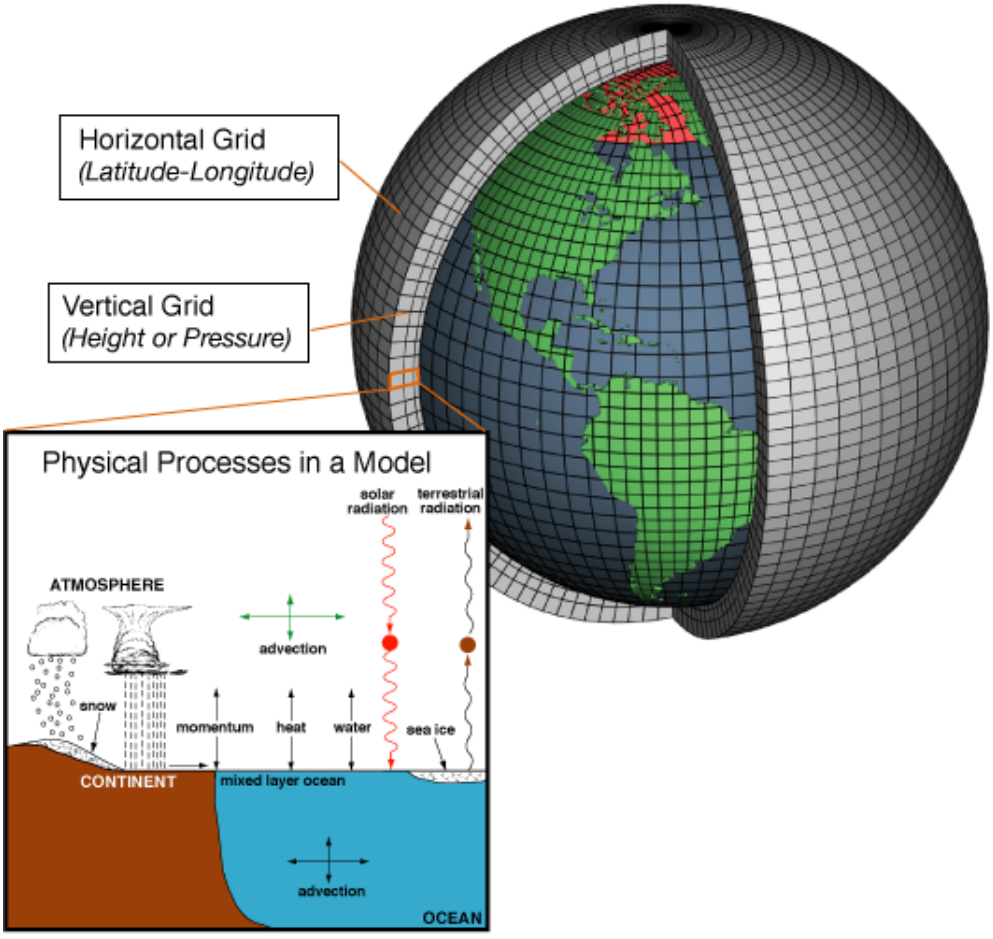

FIGURE 2.6: Schéma d'un modèle climatique pourvu d'une grille horizontale et verticale. (Tiré de NOAA).

(Boé, 2018; Abramowitz et al., 2019).

Nous avons lors de cette thèse utilisé des AOGCM issues de CMIP5 (Taylor et al., 2012) et de CMIP6 (Eyring et al., 2016).

Simulations utilisées

Plusieurs simulations différentes ont été menées à travers CMIP ou à travers la communauté SMILE (*Single model initial condition large ensemble*, Maher et al., 2021) que nous décrirons plus tard. Ces simulations sont caractérisées par les valeurs spatio-temporelles de forçage utilisées comme conditions limites. Ces conditions limites ont été estimées à l'aide de reconstructions historiques estimés des forçages. La période historique commence en 1850, où les premières mesures fiables de température du globe sont réalisés et se poursuit jusqu'à nos jours.
Aussi, des estimations des conditions limites futures de ces forçages à l'échelle de plusieurs siècles ont été menés selon différents scénarios socio-économiques (Rounsevell and Metzger, 2010; O'Neill et al., 2014). Un scénario est une description de la manière dont l'avenir peut évoluer, sur la base d'un ensemble cohérent et homogène d'hypothèses sur la démographie, l'économie, l'innovation technologique, la gouvernance et les modes de vie.

Simulations historical

Les simulations les plus utilisées lors de nos travaux sont les simulations dites *"historical"* issues de CMIP5 et de CMIP6. Leur période va de 1850 à 2005 pour CMIP5 et 2016 pour CMIP6 et utilisent comme conditions limites les concentrations de gaz à effet de serre, les aérosols anthropiques ou naturels, l'ozone ou l'utilisation des terres. D'autres simulations utilisent les conditions limites provenant des scénarios de chemins socio-économiques partagés (SSP pour *Shared Socio-economics Pathways*) allant de 2015 à 2100

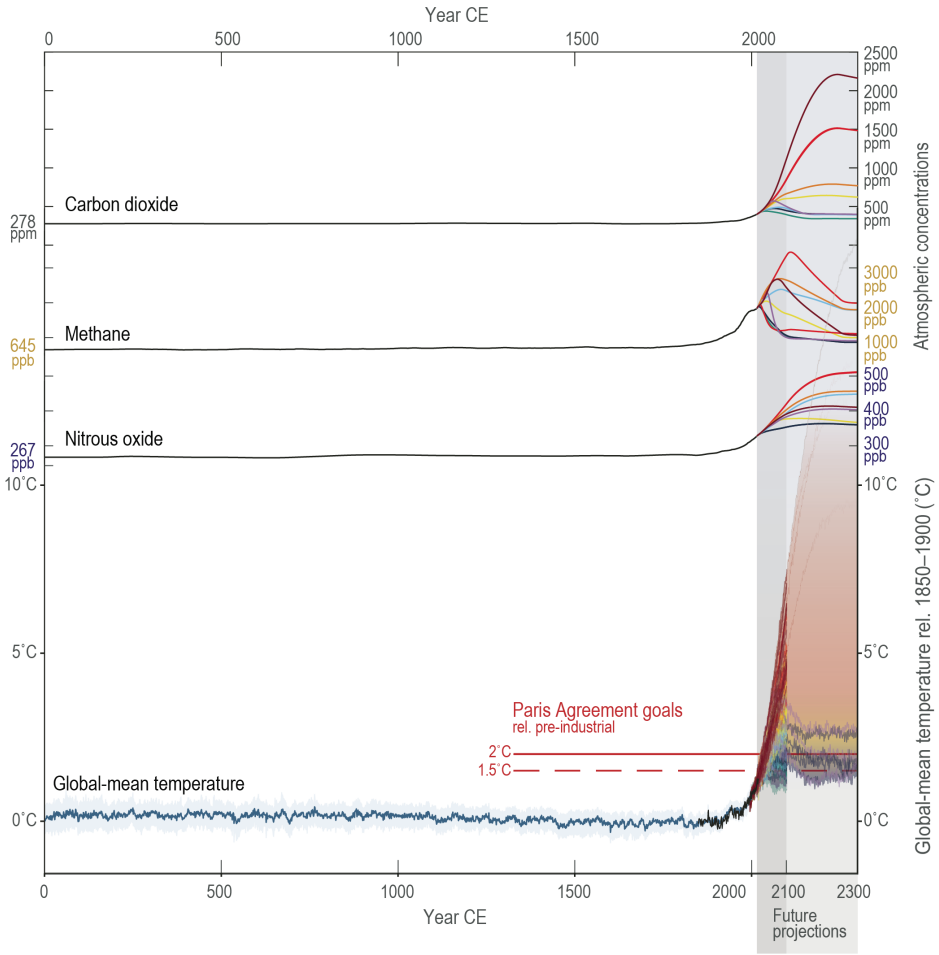

FIGURE 2.7: Concentrations historiques et projetées de dioxyde de carbone, de méthane et d'oxyde nitreux utilisés dans CMIP6 et températures moyennes mondiales (GMST) pour différentes projections par rapport à la période 1850 - 1900. Tiré de Chen and Tréguier, 2021

pour CMIP6 et aux RCP (*Representative concentration Pathway*) de CMIP5 allant de 2006 à 2100. Nous illustrons dans la figure 2.7 l'évolution de différents gaz à effet de serre prescrits dans les simulations historiques et scénarios de CMIP6 ainsi que la température globale moyenne de l'ensemble des modèles CMIP6. Comme on peut le voir, l'impact des scénarios est prédominant sur la GMST simulée.

Simulation Pi-Control

Les simulations Pi-Control ne considèrent aucune évolution des valeurs de forçage qui prennent leurs valeurs de 1850, période considérée comme préindustrielle. Elles représentent ainsi un climat de contrôle sans variation des forçages externes, et donc évaluent uniquement la variabilité interne océans-atmosphère. Les résultats d'une simulation Pi-Control servent généralement de conditions initiales aux simulations *historical*.

Simulations DAMIP

Les simulations de DAMIP (*Detection and attribution model intercomparaison project*; (Gillett et al., 2016)) sont des simulations d'un volet de CMIP6 utilisées en particulier pour la détection et attribution du changement climatique. Ce sont des simulations dites "mono-forçage" car elles ne considèrent l'évolution que d'un seul forçage. Les autres forçages sont gardés à leurs valeurs pré-industriel. Nous avons utilisé trois types de simulation DAMIP : les simulations *historical-GHG* qui considèrent uniquement l'évolution des émissions de gaz à effet de serre, les *historical-aer* pour les simulations considérant uniquement l'évolution des aérosols anthropique et les simulations *historical-nat* qui considèrent uniquement l'effet des forçages naturels comme les éruptions solaires ou volcaniques. Ces phénomènes constituent également des forçages, car ils affectent le bilan radiatif de la Terre bien qu'ils ne soient pas liés à l'activité humaine. On peut voir dans la Fig 2.8 l'anomalie de GSAT annuelle avec la période 1850-1900 comme référence, produite par les simulations DAMIP et *historical* du modèle IPSL-CM6-LR en affichant la moyenne (ligne) de leurs membres et la dispersion entre les membres (donné par un écart-type, zone coloré). Les résultats obtenus montrent un réchauffement prononcé pour les simulations *historical-GHG*, un refroidissement dans les simulations *historical-aer* et une stabilité pour les simulations *historical-nat*

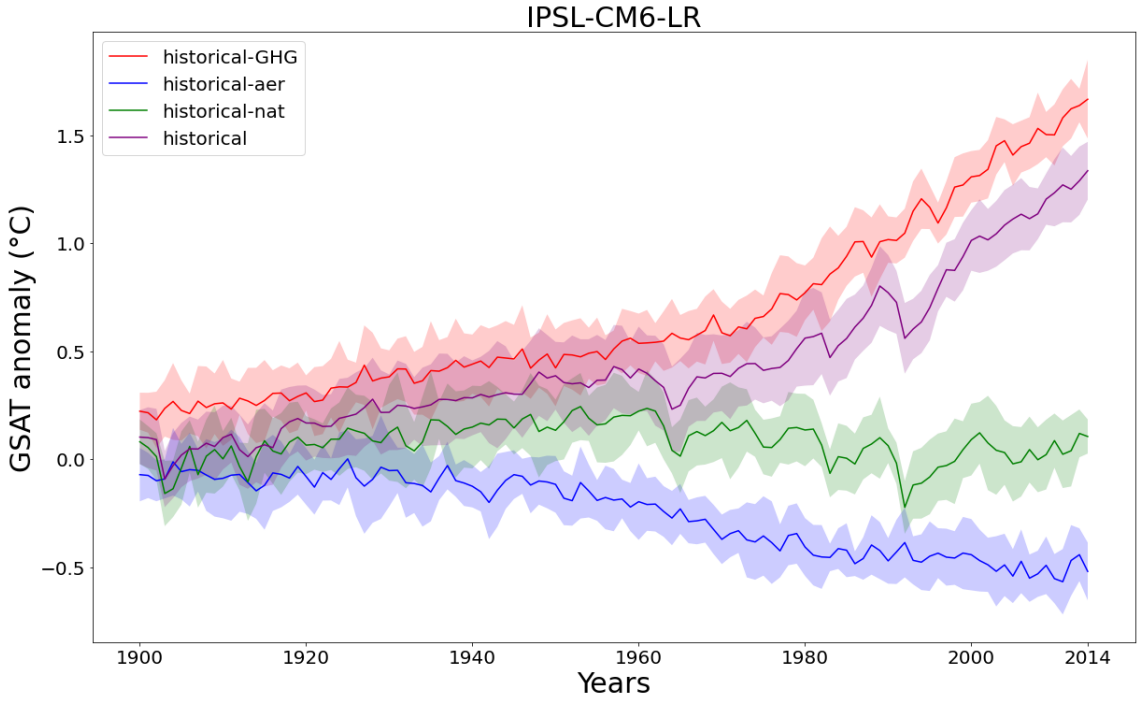

FIGURE 2.8: Moyenne d'ensemble des anomalies de température (en °C) par rapport à la période 1850-1900 des simulations (en rouge) *historical-GHG*, (bleu) *historical-aer*, (vert) *historical-nat* ou (violet) *historical* pour le modèle IPSL-CM6-LR. Les ombres de même couleur montrent un écart-type de la variabilité inter-membre.

excepté durant les épisodes d'éruptions volcaniques comme Agung (1963), El Chichón (1982) ou Pinatubo (1991) qui rejettent des aérosols naturels dans l'atmosphère. En effet, les gaz à effet de serre contribuent à un réchauffement du climat par effet de serre alors que les aérosols contribuent à un refroidissement du climat.

2.2.3 Variabilité interne et forcé dans les modèles climatiques

Pour chaque modèle climatique et simulation de celui-ci, une unique réalisation n'est souvent pas suffisante compte tenu de l'effet de la variabilité interne. En effet, chaque simulation nécessite de fixer des conditions initiales. Ces conditions initiales ont un grand impact sur les résultats des simulations à cause du chaos inhérent au système climatique. Une simulation a donc souvent plusieurs réalisations appelées membres qui échantillonnent la

variabilité interne du modèle et de la simulation. La variabilité interne se décorrèle entre les membres après plusieurs décades. Les membres deviennent alors indépendants entre eux. Disposer de beaucoup de membres pour un même modèle climatique et simulation permet donc d'étudier l'impact de la variabilité interne dans le modèle climatique. C'est ce qui a motivé la création des larges ensembles tels que SMILE (Maher et al., 2021). Ces ensembles contiennent un grand nombre de membres (de l'ordre de 30 membres ou plus). Malgré leur existence relativement récente, ils ont été grandement utilisés pour étudier la variabilité interne du climat (Frankignoul et al., 2017) ou les évènements extrêmes (Kirchmeier-Young et al., 2017) mais également pour tester et valider de nouvelles approches (Wills et al., 2020) ou communiquer avec les instances politiques par exemple, sur la question de savoir si des réductions d'émission de gaz à effet de serre auront un impact visible à quelques années ou seront masqués par la variabilité interne (Lehner et al., 2020).

Une méthode simple pour estimer la variabilité forcée du modèle est alors de faire la moyenne de ses membres. Cela permet de réduire la variabilité interne par un facteur de $\sqrt{n}$ avec n le nombre de membres utilisés. Les travaux de Deser et al., 2012; Deser et al., 2014 ont montré qu'un ensemble de 10 à 40 membres peuvent être nécessaires pour identifier les changements de la température de surface du climat en Amérique du Nord dans les simulations historiques. Ce chiffre peut être plus important sur des variables climatiques ayant un rapport du signal de variabilité forcé sur signal de variabilité interne plus faible comme la précipitation ou la pression au niveau de la mer. Ces études ont conduit à de nouvelles méthodes cherchant à estimer le signal forcé à partir d'une simulation d'ensemble (Wills et al., 2020), réduisant ainsi le nombre de membres nécessaires pour bien estimer la réponse forcée d'un modèle climatique.

2.3 Méthodes pour séparer la variabilité interne et forcée

2.3.1 Contexte général

Des méthodes ont été développées pour séparer la variabilité interne et forcée dans les observations. Une telle séparation est essentielle pour mener des études de détection et d'attribution, permettant une estimation et une simulation précises de la réaction du climat aux altérations du forçage radiatif. En outre, cette différenciation permet de reconnaître et de comprendre la variabilité interne du climat. Néanmoins, la disponibilité des observations instrumentales est limitée à la période depuis 1850, et la durée relativement brève de ces observations présente des défis pour caractériser efficacement et avec confiance la variabilité interne. De plus, une partie importante de cette période (environ 1950 jusqu'à aujourd'hui) est fortement affecté par l'évolution des forçages externes.

Ces méthodes se catégorisent par les données qu'elles utilisent. Certaines méthodes utilisent uniquement les observations (Schneider and Held, 2001; Smoliak et al., 2015; Wallace et al., 2012; Deser et al., 2016; Frankignoul et al., 2017; Wills et al., 2018; Sippel et al., 2019) afin de ne pas être affectés par les erreurs potentielles des modèles climatiques. Les modèles peuvent en effet ne pas réussir à bien capter certaines tendances régionales (Shin and Sardeshmukh, 2011).

D'autres méthodes utilisent les résultats de modèles climatiques (Frankcombe et al., 2015; Ting et al., 2009; DelSole et al., 2011) avec parfois une approche multimodèles. En effet, il a été argumenté qu'une approche multimodèle devrait grandement réduire l'impact d'un modèle climatique individuel.

Nous allons dans cette section décrire plusieurs méthodologies utilisées pour séparer la variabilité interne et forcée dans les observations et utiliser

certaines d'entre elles pour caractériser la variabilité interne et forcé contenu dans les observations sur la variable de la température de surface.

2.3.2 Méthodes de séparation de la variabilité interne et forcé

Tendances

Le signal forcé est estimé à chaque point de grille par une tendance linéaire ou quadratique estimée par la méthode des moindres carrés. Cette méthode utilise uniquement l'information temporelle pour chaque point de grille, et ainsi ignore les informations de covariances spatiales qui peuvent être utiles (Schneider and Held, 2001; Wallace et al., 2012; Smoliak et al., 2015; Deser et al., 2016; Frankignoul et al., 2017; Wills et al., 2018; Sippel et al., 2019).

Décomposition multidimensionnelle d'ensemble empirique

Des méthodes plus complexes que de simples tendances furent développées pour caractériser la variabilité forcée. La méthode de la décomposition multidimensionnelle d'ensemble empirique (Wu et al., 2009; Wu et al., 2011) se fonde sur le calcul, pour une série temporelle, des enveloppes caractérisées par leurs minimums et maximums locaux. Si la moyenne de ces enveloppes est suffisamment différente de zéro alors elle est retranchée à la série temporelle, donnant ainsi un résidu. Ce processus est itéré, utilisant le résidu comme nouvelle série temporelle, jusqu'à ce que le résidu obtenu soit une fonction monotone ou une fonction ayant au plus un seul extremum interne. Cette méthodologie étant sensible au bruit, des estimations plus robustes peuvent être obtenues en ajoutant différentes réalisations de bruits blancs à la série temporelle d'origine. La méthodologie est alors appliquée pour chaque réalisation du bruit et utilisant la moyenne des résidus finaux comme résultat final.

Analyse des composants basses fréquences

Introduite dans Wills et al., 2018, cette méthode est fondée sur l'analyse discriminante linéaire, une méthode statistique largement utilisée dans la reconnaissance des formes et l'apprentissage automatique et qui utilise l'information contenue dans les covariances spatiales des données. La méthode consiste à tout d'abord utiliser une décomposition en composantes principales pour trouver un petit nombre de degrés de liberté spatiales expliquant une grande partie de la variabilité des données. On projette alors les données, filtrées à l'aide d'un filtre passe-bas, sur les différentes composantes trouvées et cherchons les valeurs propres des combinaisons linéaires des composantes qui maximisent le ratio de variabilité des basses fréquences sur la variabilité totale quand les données sont projetées dessus. On projette alors les données non filtrées sur la combinaison linéaire précédemment obtenue et obtenons les composants basses fréquences. Ces composants peuvent alors être utilisés pour caractériser la variabilité forcée et interne dans les observations.

Méthodes d'échelonnage

Afin de mitiger l'impact individuel des modèles climatiques Steinman et al., 2015 ont introduit une autre méthode très semblable à celle utilisée en détection et attribution du changement climatique (Allen and Tett, 1999; Allen and Stott, 2003). Dans cette méthode, la moyenne multimodèles n'est pas retranchée simplement aux différents membres *historical*. Elle est utilisée comme variable explicative d'une régression linéaire ayant comme cible les observations. Le reste non expliqué par cette méthode représente donc la variabilité interne. Cette méthode peut être complexifiée avec une régression multivariée et les simulations DAMIP. Par exemple, une régression à deux variables explicatives peut être faite avec les moyennes multimodèles des simulations *historical-GHG* et *historical-nat*. Ou bien une régression à trois

variables explicatives avec les simulations *historical*, *historical-nat* et *historical-GHG*. Cette méthode sera plus détaillée dans la section "empreintes optimisées".

2.3.3 Étude la variabilité interne et forcé

Tendances du réchauffement global

L'utilisation de ces différentes méthodes permet d'obtenir des estimations de la localisation et intensité de la variabilité forcée. La figure 2.9 montre le réchauffement estimé par tendances dans les observations et les changements simulés dans la moyenne multimodèles de CMIP6 pour 1 °C de changement global de température par rapport à la période 1850-1900. Nous pouvons voir que le réchauffement climatique touche toutes les régions du globe, mais pas au même niveau. Les surfaces terrestres sont bien plus touchées par le réchauffement que les océans. Les hautes latitudes sont sensiblement plus touchées par le réchauffement climatique dû au phénomène d'amplification polaire. A contrario l'Atlantique Nord montre un réchauffement très limité, voire quasiment nul dans les observations. Cela est dû au phénomène de trou du réchauffement climatique dans l'Atlantique Nord (Keil et al., 2020) dû à un ralentissement de la circulation méridienne de retournement de l'Atlantique sous l'effet du changement climatique. La variabilité forcée retrouvée dans la moyenne multimodèles de CMIP6 et dans les observations est similaire pour la majorité des régions, même si le réchauffement est surestimé dans les simulations pour l'Atlantique nord et pour l'Antarctique.

Signal temporel de la variabilité multidécennal de l'Atlantique

Séparer la variabilité interne et forcé permet également de mieux caractériser la variabilité interne. En effet, certains modes de variabilité interne ont un effet se confondent avec la variabilité forcée, rendant leur étude plus complexe.

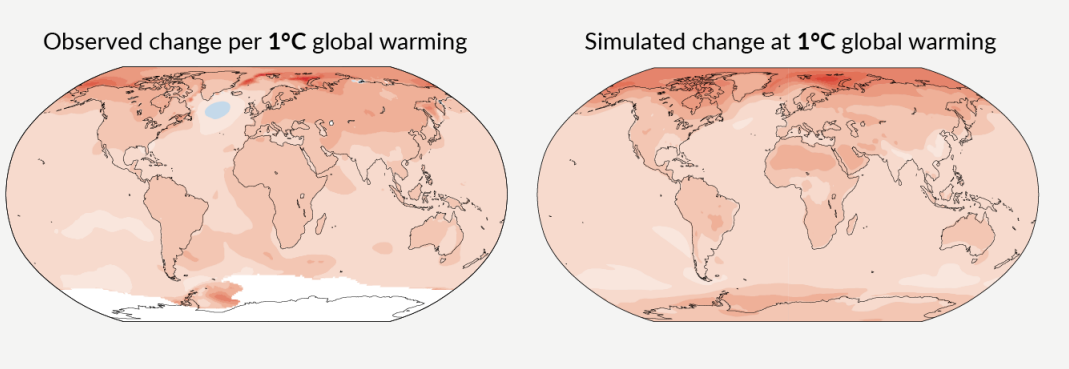

FIGURE 2.9: Comparaison entre les tendances dans la température de surface de l'air dans la période 1850-2020 par °C de réchauffement global dans (carte de gauche) les observations de *Berkely Earth* et (carte de droite) pour la température multi-modèle de CMIP6. Tiré de Masson-Delmotte et al., 2021b

C'est le cas pour l'AMV dont le signal basse fréquence peut se confondre avec la variabilité forcée. Il est donc de coutume de retirer la variabilité forcée dans le signal de l'AMV et d'étudier uniquement le résidu régional dans les SST. Il n'existe pas de méthode unique et privilégiée pour éliminer l'impact des forçages externe dans les observations de l'AMV. Retirer la tendance est utilisée dans la définition traditionnelle de l'indice AMV qui correspond à la moyenne sur l'ensemble du bassin de l'Atlantique Nord des anomalies annuelles de la SST où l'on a enlevé la tendance et auxquelles on applique une moyenne glissante sur 10 ans, comme le proposent Enfield et al., 2001. Cette série temporelle montre quelques phases distinctes sur les quelque 120 ans d'enregistrement instrumental, à savoir des périodes chaudes de 1930 à 1965 et depuis 1995, et des périodes froides de 1900 à 1925 et de 1965 à 1995 (voir 2.10). Pour éliminer le signal non linéaire résiduel, différentes approches ont été proposées, soit basés sur la base d'observations en appliquant diverses méthodes statistiques (Trenberth and Shea, 2006; Frajka-Williams et al., 2017; Frankignoul et al., 2017; Sutton et al., 2018; Yan et al., 2019), soit sur la base d'estimations des modèles climatique du signal forcé par les forçages naturelles (solaire et volcanique) et anthropique (gaz à effet de serre

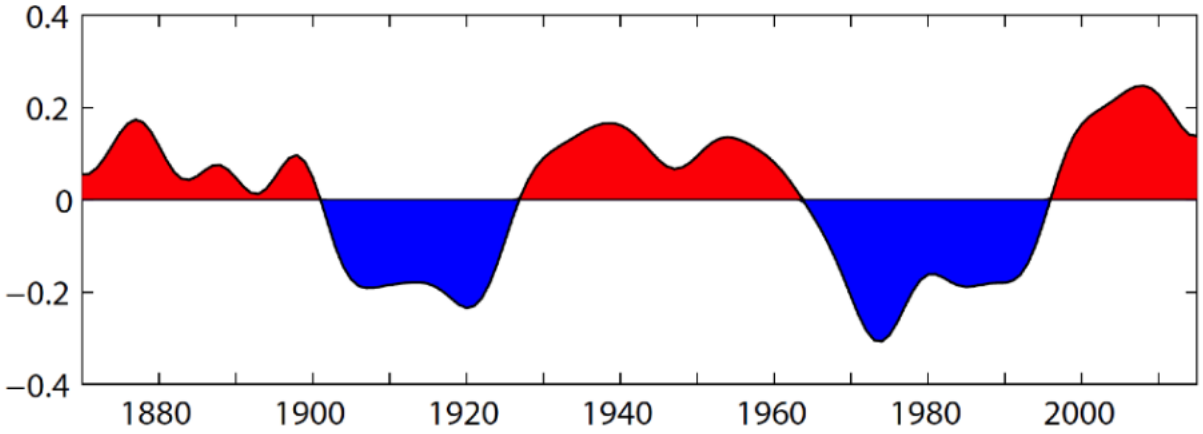

FIGURE 2.10: Indice AMV observé, défini comme la moyenne annuelle des anomalies de la SST, filtrée passe-bas sur 10 ans, sur le bassin de l'Atlantique Nord (0N-65N, 80W-0E), avec le jeu de données HadISST (Rayner et al. 2003) pour la période 1870-2015. (Créé par Dennis Shea et Dr Rhong Zhang pour le *Climate Data Guide*).

et aérosols) évaluée à partir de simulations historiques (Ting et al., 2009; Terray, 2012; Steinman et al., 2015; Tandon and Kushner, 2015). Plus précisément, dans Trenberth and Shea, 2006, la série temporelle d'anomalies de la SST moyenne mondiale observée annuellement est soustraite de la série temporelle annuelle observée de l'Atlantique Nord pour obtenir l'indice AMV brut non filtré. Dans Ting et al., 2009, une analyse EOF maximisant le rapport signal/bruit est appliquée aux moyennes annuelles globales des SST dérivées de l'ensemble multimodèle CMIP5 pour extraire une estimation de la composante forcée, qui est a priori supprimée avant le calcul de l'indice AMV, puis définie comme un résidu à basse fréquence. Cette dernière technique a été retenue dans les expériences de sensibilité CMIP6 DCPP-C (Boer et al., 2016) qui visent à mieux comprendre les téléconnexions associées à l'AMV.

2.4 Méthodes de détection et attribution

2.4.1 Contexte général

Différentes méthodes de résolution du problème de la détection et attribution du changement climatique existent. De nouvelles méthodes sont créées afin de pouvoir mieux quantifier le rôle de la variabilité interne et de résoudre des problèmes d'attribution à des échelles plus locales, sur d'autres variables que la température moyenne de surface, ou pour des évènements extrêmes (Trenberth et al., 2015).

Le but d'une méthode de détection et d'attribution consiste à donner une évaluation de la variabilité forcée dû à chaque forçage sur la variable étudiée, associé à une mesure de confiance de ces résultats.

Nous allons dans cette partie présenter différentes méthodes de détection et d'attribution utilisées pour la variable de température de surface à l'échelle globale ou à l'échelle régionale ainsi que certains de leurs résultats.

2.4.2 Empreintes optimisées

Les empreintes optimisées ont été développées dans Hasselmann, 1993. Cette méthode se fonde sur une régression linéaire multivariée avec les observations comme cible et les estimations de l'impact de chaque forçage (appelées empreintes) comme variables explicatives. L'hypothèse sous-jacente est que les changements observés consistent en une addition entre les effets des forçages et de la variabilité interne. Soit l'équation :

$$Y = \sum_i \beta_i X_i + \epsilon \tag{2.1}$$

Avec Y les observations de la variable physique étudié, X_i l'effet estimé du forçage i (les fameuses empreintes), β_i le coefficient de régression associé au forçage i et ϵ la matrice de covariance de la variabilité interne.

L'estimation des empreintes de chaque forçage se fait avec les différentes simulations d'un ou plusieurs modèles climatiques suivi de moyennes temporelles et/ou spatiales et de projections sur différents espaces. Par exemple, pour étudier la variable de température globale Ribes et al., 2013 ont utilisé des moyennes temporelles décanales et en projetant la moyenne obtenue sur des harmoniques sphériques spatiales.

Les coefficients de régression pour chaque forçage obtenu sont appelés facteurs d'échelles. Leurs intervalles de confiance peuvent être calculés par des tirages aléatoires considérant la variabilité interne dans les observations et les modèles climatiques (Allen and Stott, 2003). Si les valeurs d'intervalles de confiance obtenus n'incluent pas zéro alors le forçage est détecté dans les observations. Si l'intervalle de confiance du facteur d'échelle comprend la valeur un alors l'estimation du forçage par les modèles climatiques est cohérente avec les observations. Une estimation du changement attribuable peut alors être déterminée grâce aux facteurs d'échelles et les empreintes issues des modèles climatiques. Deux méthodes de régression linéaire peuvent être utilisées pour effectuer cette régression. La première est la méthode simple des moindres carrés. La seconde, est la méthode des moindres carrés totaux et permet de prendre en compte la variabilité interne des X_i (Allen and Stott, 2003) ainsi que l'incertitude structurelle des modèles climatiques (Huntingford et al., 2006). L'équation fondamentale devient alors :

$$Y - v_0 = \sum_i \beta_i (X_i - v_i - \mu_i) \tag{2.2}$$

avec v_0 et v_i le bruit climatique présent dans les observations et les modèles climatiques (une hypothèse est qu'ils ont la même structure) et μ_i l'incertitude dans les projections de modèles du fait d'utiliser de multiples modèles climatiques.

Dans les deux cas, une estimation de la variabilité interne est nécessaire

afin d'améliorer le rapport signal/bruit. Les observations et les réponses simulées par le modèle sont en effet généralement normalisées par une estimation de la variabilité interne dérivée des simulations du modèle climatique. Cette procédure nécessite une estimation de la matrice de covariance inverse de la variabilité interne, et certaines approches ont été proposées pour une estimation plus fiable de celle-ci (Ribes et al., 2009). Un signal peut être détecté de manière erronée en raison d'un bruit trop faible, et la variabilité interne simulée doit donc être évaluée avec soin. La variabilité simulée par le modèle peut être vérifiée en comparant la variance modélisée provenant de simulations non forcées avec la variance résiduelle dans Eq 1 à l'aide d'un test de cohérence résiduelle standard (Allen and Tett, 1999; Ribes et al., 2013). Imbers et al., 2014 ont testé la sensibilité des résultats de détection et d'attribution à différentes représentations de la variabilité interne associées aux processus à mémoire courte et à mémoire longue.

Afin d'obtenir une estimation plus précise de la matrice de covariance de la variabilité interne, une méthode courante est d'utiliser une décomposition en EOF pour réduire la dimension spatiale des données en choisissant une troncature adaptées aux données. Cependant, cette méthode implique un choix arbitraire du nombre de fonctions orthogonales empiriques utilisées pour tronquer les données. La méthode des empreintes optimisées régularisées (Ribes et al., 2013) évite ce choix arbitraire avec une estimation régularisée de la matrice de covariance de la variabilité interne.

On illustre dans la Fig 2.11 les facteurs d'échelle obtenus avec la méthode des empreintes optimisés régularisés et les changements de température de surface de l'air attribuables aux différents forçages sur la période 2010-2019 par rapport à la période 1850-1900 pour différents modèles climatiques issues de CMIP6 ainsi que pour une approche multimodèles et différents groupes de forçages. Les différents groupes de forçages étudiés sont dans un premier temps l'ensemble des forçages anthropique et des forçages naturels.

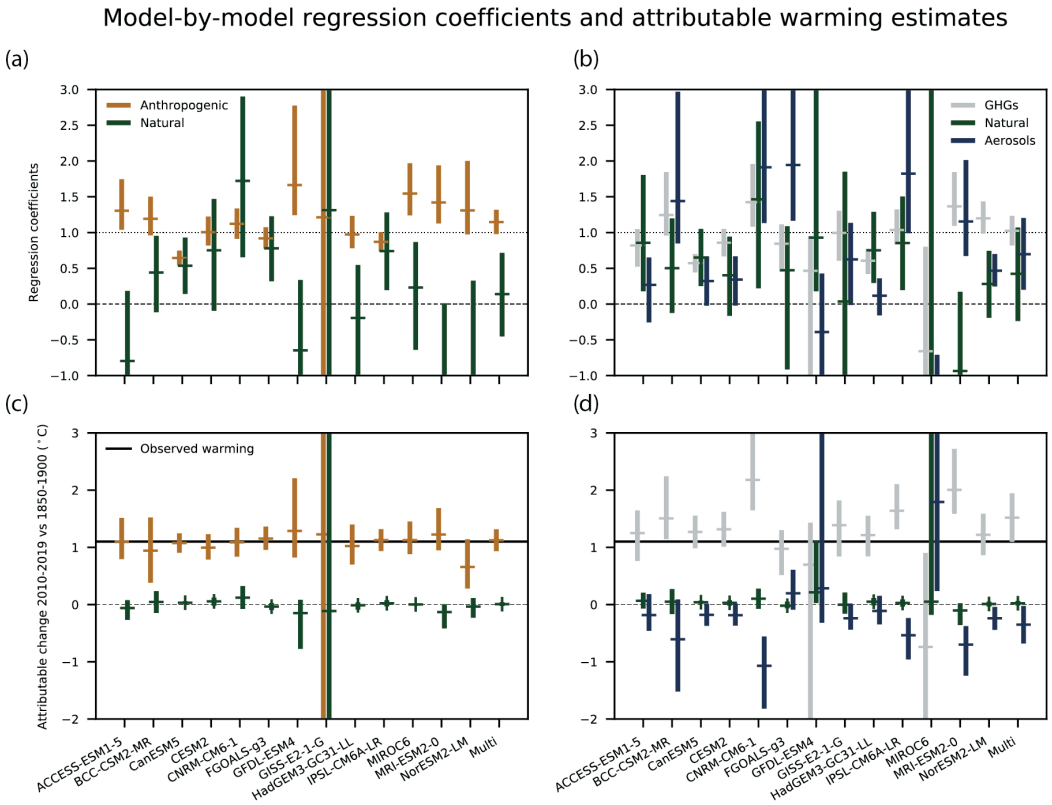

FIGURE 2.11: (a) Facteurs d'échelles obtenues avec deux groupes de forçages, un regroupant tous les forçages anthropique (*Anthropogenic*) et l'autre tous les forçages naturels (*Natural*) et différents modèles climatiques. (b) Comme (a) mais pour trois groupes de forçages : les forçages naturels (*Natural*), les aérosols anthropiques (*Aerosols*) et le reste des forçages anthropique (GHGs). (c) Changement attribuable en 2010-2019 comparé à la période 1850-1900 pour les facteurs d'échelle obtenue dans la figure (a). (d) Identique à (c) mais pour les facteurs d'échelles de (b). Tiré de Gillett et al., 2021

Dans la Fig 2.11a Les facteurs d'échelles obtenues pour les effets anthropique n'incluent pas la valeur 0 pour tous les modèles climatiques et l'approche multimodèle excepté pour le modèle HadGEM3-GC31-LL qui montre un très grand intervalle de confiance. Cela signifie que le forçage anthropique est détecté. À l'inverse, les intervalles de confiance du facteur d'échelle relié aux forçages naturels incluent les valeurs 0 pour une majorité de modèles et pour l'approche multimodèle ce qui indique que les forçages naturels ne sont pas détectés pour les changements de température de surface. Les changements de température attribuables pour ces deux groupes de forçages sont eux relativement similaires pour les différents modèles climatiques et l'approche multimodèles. Il est de l'ordre de 1,1 °C en 2010-2019 comparé à la période 1850-1900 pour les forçages anthropiques et de 0 °C pour les forçages naturels. Dans Fig 2.11cd) nous nous intéressons à différents groupes de forçages : l'ensemble des forçages naturels, les aérosols anthropiques et le reste des forçages anthropique noté "GHGs" car les gaz à effet de serre en représentent une grande partie de ceux-ci. Les facteurs d'échelles montrent pour tous les forçages des non-détection pour certains modèles climatiques. Les changements de température attribuables ont une plus grande variabilité entre les différents modèles climatiques. On remarque en particulier que le refroidissement dû aux aérosols est particulièrement incertain. L'approche multimodèle montre toutefois un refroidissement de l'ordre de 0,3 °C pour les aérosols anthropique, un réchauffement de 1,4 °C pour les autres forçages anthropique et un impact nul des forçages naturels.

D'autres variantes de la méthode des empreintes optimisées existent. Hannart et al., 2014 ont décrit une procédure d'inférence pour les facteurs d'échelle qui évite de faire l'hypothèse que l'erreur de modèle et la variabilité interne ont la même structure de covariance. Une approche intégrée de l'empreinte optimale a également été suggérée dans laquelle toutes les sources d'incertitude

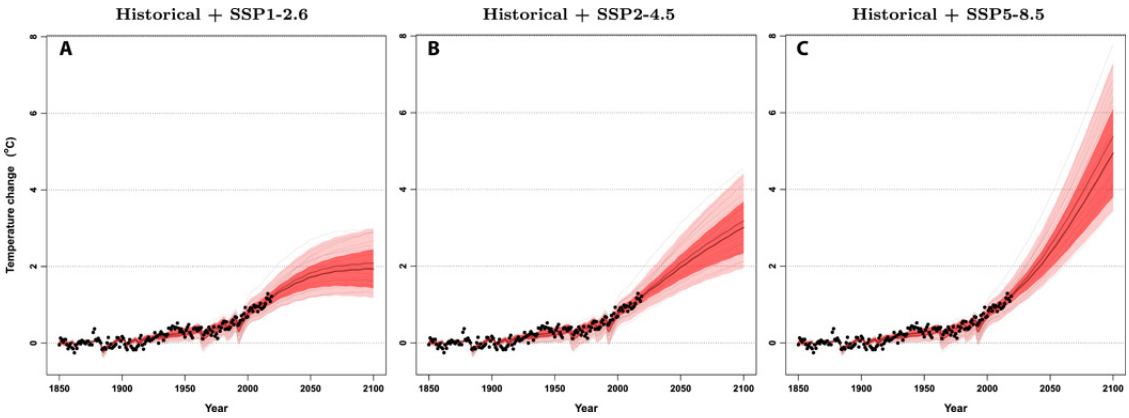

FIGURE 2.12: (A à C) La contrainte observationnelle est appliquée aux simulations concaténées des scénarios historiques et SSP (SSP1-2.6, SSP2-4.5, ou SSP5-8.5). Les valeurs de GSAT observées annuellement (points noirs) sont comparées aux plages de confiance de 5 à 95 % de la réponse forcée non contrainte (rose) et contrainte (rouge), telles qu'estimées à partir de 22 modèles CMIP6. Les séries temporelles des réponses forcées des modèles CMIP6 individuels sont également représentées (lignes grises fines). Tiré de Ribes et al., 2021

(c'est-à-dire l'erreur d'observation, l'erreur des modèles climatiques et la variabilité interne) sont traitées dans un seul modèle statistique sans étape préliminaire de réduction de la dimension (Hannart et al., 2016). Katzfuss et al., 2017 ont introduit une approche intégrée similaire basée sur une moyenne de modèles bayésiens. Aussi, DelSole et al., 2019 ont suggéré une méthode *bootstrap* pour mieux estimer les intervalles de confiance des facteurs d'échelle, même dans un régime de signaux faibles.

2.4.3 Méthode de contraintes par les observations

Afin de prendre en compte les incertitudes dû à la variabilité interne dans les approches basées sur la régression et afin d'utiliser directement toute la série observée sans perdre d'information en réduisant ses dimensions, Ribes et al., 2021 ont introduit un nouveau cadre d'inférence statistique. Cette méthode leur permet d'utiliser l'entièreté des observations pour affiner les estimations de changement climatique futures et passées, et a été appliquée pour réduire la plage d'incertitude du réchauffement anthropique estimé. Cette méthode

vise à contraindre les simulations des modèles climatiques avec les observations. Elle utilise une hypothèse d'additivité des forçages et une méthodologie de maximisation de la vraisemblance. Soit x la réponse du climat à tous les forçages ainsi qu'aux forçages individuels. On note $\pi(x)$ une estimation de x faite avec uniquement les modèles climatiques, soit $\pi(x) = N(\mu, \Sigma_{mod})$ avec μ le vecteur des réponses moyennes des modèles climatiques aux différents forçages (ou tous les forçages simultanément) et Σ_{mod} un vecteur regroupant les différentes matrices de variances-covariances des réponses des modèles climatiques aux différents forçages. Ils font ensuite l'hypothèse que les observations y peuvent être écrites comme une somme des réponses forcées plus la variabilité interne (l'hypothèse d'additivité) et des erreurs de mesure. Soit :

$$y = Hx + \epsilon \tag{2.3}$$

avec y les observations, H la matrice d'erreur d'observation, x la réponse du climat aux différents forçages et ϵ la variabilité interne assumé gaussienne.
Comme $\pi(x)$ et ϵ suivent des lois gaussiennes, il devient possible d'estimer la distribution $p(x, y)$ contrainte par les observations.
On illustre dans la Fig 2.12 la contrainte des différentes simulations historiques de CMIP6 par les observations obtenues pour différents scénarios SSP. On peut voir que cette contrainte continue d'avoir un effet important même après plusieurs décennies.

2.4.4 Autres méthodes

Hannart and Naveau, 2018 ont étendu l'application de la théorie causale standard (Pearl, 2009) au contexte de la détection et de l'attribution en convertissant une série chronologique en un événement, et en calculant la probabilité de causalité, une approche qui maximise la preuve causale associée au forçage. D'autre part, Schurer et al., 2018 ont employé un cadre bayésien

pour considérer explicitement l'incertitude de la modélisation climatique dans la méthode des empreintes optimisées. Des travaux utilisent une analyse discriminante afin d'étudier des échelles de temps plus courtes comme les températures saisonnières (Jia and DelSole, 2012) ou la mousson d'Asie du Sud (Srivastava and DelSole, 2014). La même approche a été appliquée pour séparer les réponses au forçage des aérosols d'autres forçages (Yan et al., 2016) et les résultats obtenus à l'aide des sorties de modèles climatiques ont indiqué que la détectabilité de la réponse des aérosols est maximisée avec un ensemble de données de température et de précipitations. Paeth et al., 2017 ont introduit une méthode de détection et d'attribution applicable à de multiples variables, basée sur une analyse discriminante et une méthode de classification bayésienne. Enfin, une approche systématique a été proposée pour traduire l'analyse quantitative en une description de la confiance dans la détection et l'attribution d'une réponse climatique aux facteurs anthropiques (Hansen and Stone, 2016). Une dernière méthode à mentionner est une méthode purement physique n'utilisant ni les observations ni les modèles climatiques. Il s'agit de construire un modèle physique du bilan énergétique de la Terre vis-à-vis du forçage radiatif. Les résultats obtenus par toutes ces méthodes sont assez proches entre eux à l'échelle globale.

2.4.5 Attribution à l'échelle globale

La grande variété de ces méthodes et la similarité de leurs résultats permettent d'avoir des résultats robustes à l'échelle globale. Ces méthodes montrent cependant certaines dissimilarités de résultats pour la décomposition de l'impact du forçage de l'activité humaine en ces différentes composantes comme les gaz à effet de serre, les aérosols anthropiques, etc.

Les différentes méthodes de détection et attribution du changement climatique ont montré de manière virtuellement certaine l'existence d'un réchauffement climatique depuis le 20ème siècle dans la température annuel globale de surface. Ce changement est en quasi-totalité imputable aux activités humaines, et ce, de manière virtuellement certaine. Ce réchauffement a renversé la tendance de refroidissement existant depuis 5 000 ans dans l'hémisphère nord, causé par un forçage astronomique.

La figure 2.13 compare les changements attribuables dans la GSAT complète à l'échelle mondiale pour la période 2010-2019 par rapport à la période 1850-1900 à partir de trois études de détection et d'attribution utilisant une méthode d'empreintes, dont deux utilisent les moyennes multimodèles de CMIP6 (Gillett et al., 2021; Ribes et al., 2021) et une des modèles de CMIP5. Une quatrième estimation notée, "*Chapter 7*", se base sur un modèle énergétique du climat. Cette figure montre également les changements de GSAT directement simulés en réponse à ces forçages dans treize modèles CMIP6. En dépit de leurs méthodologies et de leurs ensembles de données d'entrée différents, les trois approches d'attribution donnent des résultats très similaires. La plage de réchauffement attribuable à l'homme englobant le réchauffement observé, situé en moyenne à 1,1 °C sur la période 2010-2019 par rapport à la période 1850-1900. Le réchauffement attribuable aux forçages naturels est proche de zéro. Le réchauffement dû à l'augmentation des gaz à effet de serre est compensé en partie par le refroidissement dû à d'autres agents de forçage anthropiques, principalement les aérosols, bien que l'incertitude liée à ces derniers soit importante. Les estimations basées sur le modèle énergétique (*Chapter 7* sur la figure) sont proches des estimations des études d'attribution, bien qu'elles soient le produit d'une approche différente. Cet accord renforce la confiance dans l'ampleur et les causes du réchauffement de la température de surface.

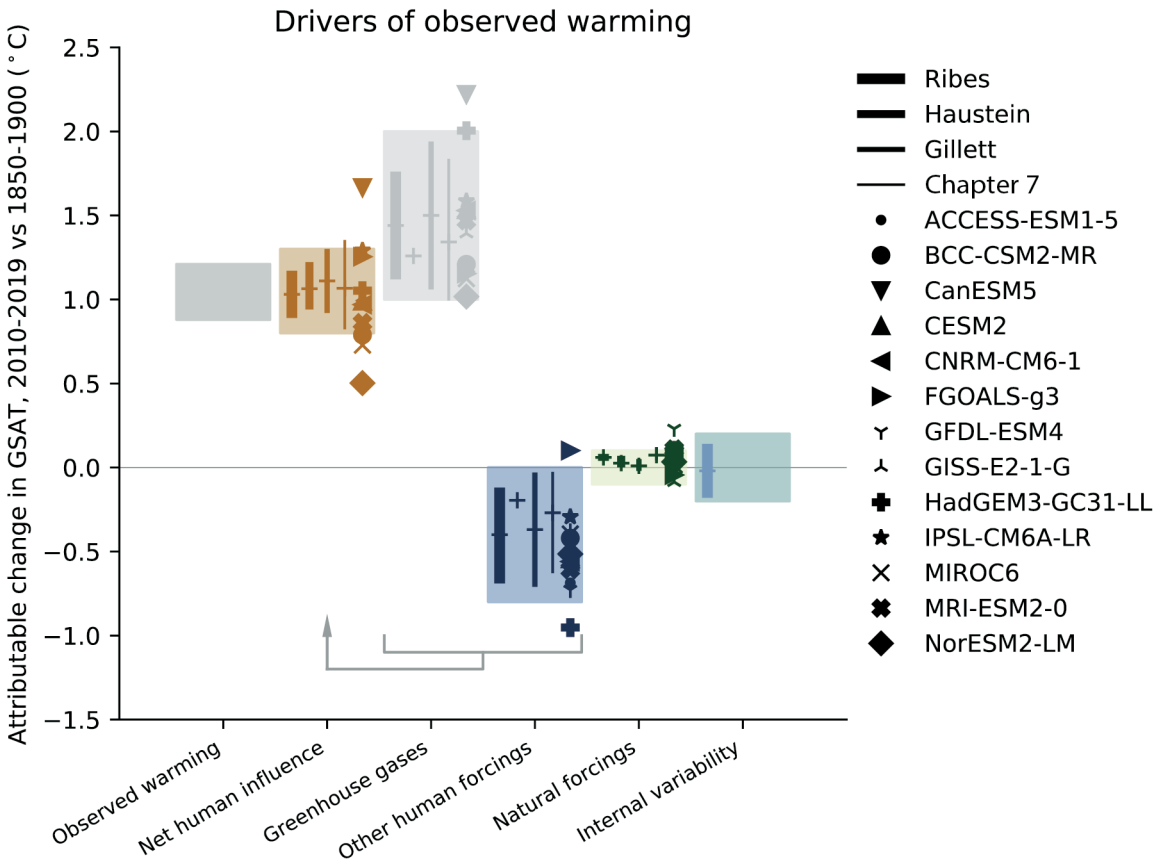

FIGURE 2.13: Changement attribuable de la GSAT (en °C) en 2010-2019 par rapport à la période 1850-1900 pour différents groupes de forçages, dans les observations et pour la variabilité interne évalués pour différentes méthodes avec des intervalles de confiance à 90 %. Les symboles colorés montrent les réponses simulées par des modèles climatiques. Tiré de Eyring et al., 2021

2.5 Réseau de neurones

2.5.1 Définition

Un réseau de neurones artificiel (ANN pour *artificial neural networks*) est un modèle statistique dont le fonctionnement a été inspiré par le fonctionnement du cerveau humain. Un réseau de neurones artificiel est une fonction mathématique (noté f) non linéaire qui transforme des variables d'entrées en variables cibles (Bishop, 1994). Cette fonction vise à approximer une fonction $f*$ qui relie les entrées aux sorties. La transformation décrite par la fonction du réseau de neurones artificiel est gouvernée par des paramètres (noté θ) nommés "poids" et "biais" (Goodfellow et al., 2016). Par exemple, pour une entrée x correspondant à une sortie, nous avons la relation $y = f*(x)$. Un ANN cherche à trouver les "bons" paramètres θ pour retrouver cette relation : $y = f(x, \theta)$. La fonction $f*$ que le réseau de neurones cherche à approcher sert à répondre à une tâche liée au problème étudié. Il existe un très grand nombre de tâches possibles dont les deux principales sont la classification qui vise à catégoriser l'input dans différentes catégories et la régression qui vise à prédire une ou plusieurs valeurs numériques. Les tâches effectuées lors de cette thèse furent uniquement des tâches de régression.

L'unité de base d'un ANN est le neurone dont le fonctionnement rappelle les neurones du cerveau humain. Un neurone prend des inputs en entrée et procède à une combinaison linéaire de ceux-ci. Les poids et biais du neurone sont les constantes associées au neurone. Une fonction non-linéaire dite d'activation est appliquée au résultat de la combinaison linéaire. Nous illustrons dans la Fig 2.14 le schéma du fonctionnement d'un neurone. Diverses fonctions d'activation peuvent être utilisées selon les données utilisées ou l'objectif recherché. Dans cette thèse, nous avons principalement utilisé la tangente hyperbolique. Les poids et biais d'un réseau de neurone

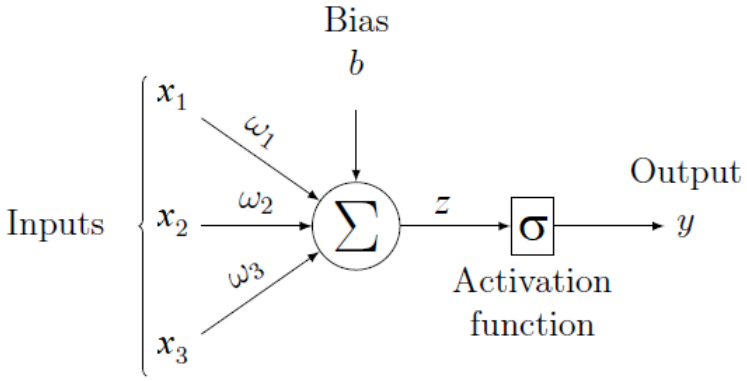

FIGURE 2.14: Schéma du fonctionnement d'un neurone

sont donc l'ensemble de poids et biais associés à tous les neurones le composant. Ce nombre de paramètres dépend de l'architecture du réseau utilisé, allant de plusieurs centaines à plusieurs milliards. De nombreuses architectures de réseaux de neurones existent, qui diffèrent par l'organisation de leurs couches, de leurs neurones, ou par des opérations supplémentaires. Ces opérations supplémentaires dépendent du réseau de neurones et désignent un très grand nombre d'opérations possible : concaténation, boucles, *pooling*, *padding*, etc.

2.5.2 Entraînement d'un réseau de neurones

L'entraînement d'un réseau de neurones consiste à fixer les valeurs de ses poids et biais à l'aide d'une base de données constituée de couples d'entrées et de sorties désirées. Lors de l'entraînement, le but est de minimiser une fonction de coût (noté c) entre les résultats de l'ANN (noté $\hat{y} = f(x)$) et la sortie désirée pour une telle entrée (noté y). Diverses fonctions de coûts sont envisageables selon les problèmes et données utilisés. Durant cette thèse, nous avons utilisé l'erreur quadratique. Pour minimiser la fonction de coût,

nous utilisons le fait que la fonction décrite par le réseau de neurones est entièrement dérivable. La rétropropagation est le nom de l'algorithme permettant de calculer avec une grande efficacité la dérivation de la fonction composé que constitue f. Avec cet algorithme, le gradient du coût en fonction des poids et biais peut donc être calculé en fonction de tous les paramètres θ de l'ANN. Les poids et biais de l'ANN peuvent alors être actualisés par descente du gradient. Soit :

$$\theta' = \theta + \epsilon \delta c(y, \hat{y}) \tag{2.4}$$

avec θ' les nouveaux paramètres de l'ANN et ϵ un scalaire positif nommé le taux d'apprentissage qui détermine la taille du pas en cours.

L'entraînement parcourt plusieurs fois les couples entrées/sorties de la base de données d'entraînement. Parcourir une fois ces couples forme une itération d'apprentissage. Ces couples sont parcourus plusieurs à la fois en les regroupant en *"batch"*. La taille de ces *batch* est un hyper-paramètre de l'apprentissage du réseau de neurone.

D'autres bases de données regroupant des entrées/sorties non utilisées sont toutefois nécessaires. Une base de données dite de validation est nécessaire afin de fixer les hyper-paramètres de l'ANN. Les hyper-paramètres sont les variables qui décrivent l'architecture de l'ANN employé ainsi que les paramètres fixant l'entraînement de celui-ci. On peut citer comme paramètres régissant l'entraînement le nombre d'itérations d'apprentissage, la valeur du taux d'apprentissage ou l'algorithme de calcul de la rétropropagation.

Fixer ces hyper-paramètres de façon appropriée est vital pour éviter que l'ANN se trouve dans un état de sous-apprentissage ou de sur-apprentissage. Le sous-apprentissage se produit quand l'ANN est incapable d'obtenir une erreur d'apprentissage suffisamment basse. Le réseau n'a alors pas réussi à apprendre les liens logiques et statistiques entre les entrées et les sorties.

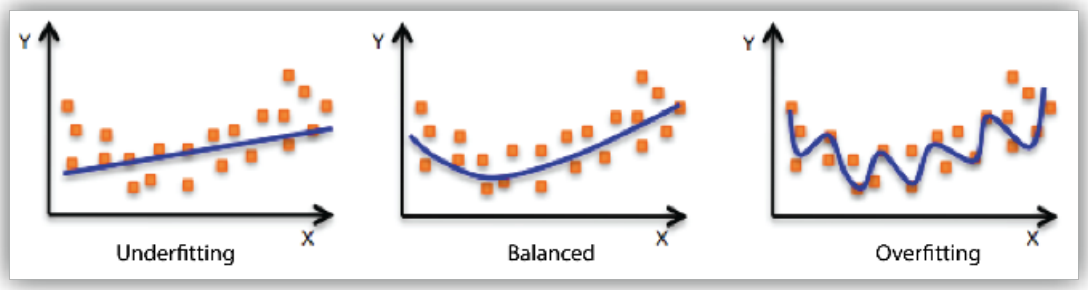

FIGURE 2.15: Illustration du (gauche) sous-apprentissage, d'un (milieu) apprentissage robuste et du (droite) sur-apprentissage. La ligne bleue désigne la fonction décrite par le réseau et les points orange la base de données d'entraînement.

Le sur-apprentissage se produit quand l'écart entre l'erreur obtenue sur les données d'entraînement et celle obtenue sur les données de validation est trop important. L'ANN a alors "mémorisé" les bonnes solutions au lieu de comprendre les liens entre les entrées et les sorties. Nous illustrons dans la Fig 2.15 une représentation graphique du sous-apprentissage et du sur-apprentissage en régression.

Un dernier set de couples entré/sortie désiré dit "données de test" est nécessaire. Il sert à évaluer les performances de l'ANN à l'aide de données nouvelles non vu lors de l'apprentissage de l'ANN ou lors du choix de ses hyper-paramètres.

Construire ces bases de données peut se retrouver problématique quand il n'y a que peu de données disponibles. Une méthode alternative pour procéder à la validation est de couper la base de données totale en seulement deux blocs d'entraînement et de test. La base d'entraînement est alors subdivisée en k blocs. Le réseau est alors successivement entraîné sur $k-1$ blocs et validé sur le dernier bloc. Cette procédure est appelée "*k-fold cross-validation*" et permet de ne pas avoir à constituer d'ensemble de données de validation, mais nécessite de fixer le nombre k de blocs et demande plus de ressources informatiques.

2.5.3 Réseaux perceptron multicouches

Les réseaux perceptrons multicouches sont les types de réseaux de neurones les plus fondamentaux et représentatifs du *machine learning* (Goodfellow et al., 2016). Ils sont à la base de beaucoup d'applications commerciales d'importance et beaucoup d'autres architectures de réseaux de neurones comme celles utilisées lors de cette thèse. Il est nécessaire d'expliciter leur fonctionnement pour pouvoir décrire les réseaux de neurones utilisés lors de la thèse.

Les neurones d'un perceptron sont organisés en couches où les neurones agissent de manière indépendante les unes des autres. Les résultats des neurones sont alors transmis à la couche suivante. La première couche est la couche d'entrée où sont mises les données d'entrée au perceptron (notés x), la dernière couche est la couche de sortie donnant y, les couches intermédiaires sont les couches dites cachées. Un perceptron peut être considéré comme une chaîne de fonctions composées qui représente chacune une couche. Par exemple, si notre ANN est composé de trois couches, on peut noter $f(x) = f3(f2(f1(x)))$ avec $f1$ la première couche, $f2$ la seconde et $f3$ la dernière. Le nombre de couches utilisées est appelé la profondeur du perceptron et c'est de là que vient le nom d'apprentissage profond. Nous illustrons dans la Fig 2.16 le schéma d'un perceptron constitué de cinq couches dont trois couches cachées. Un des problèmes courant des perceptrons est le nombre de poids et biais nécessaires : en effet chaque neurone de chaque couche est relié à tous les neurones de la couche précédente. Cela peut représenter un nombre massif de poids et biais si la taille de la couche d'entrée et de sortie, soit le nombre de dimensions de chaque entrée et sortie du réseau, sont importantes, comme c'est le cas pour des images par exemple. Cela peut mener à des réseaux trop lourds pour les ressources informatiques disponibles ou trop sensibles au sur-apprentissage.

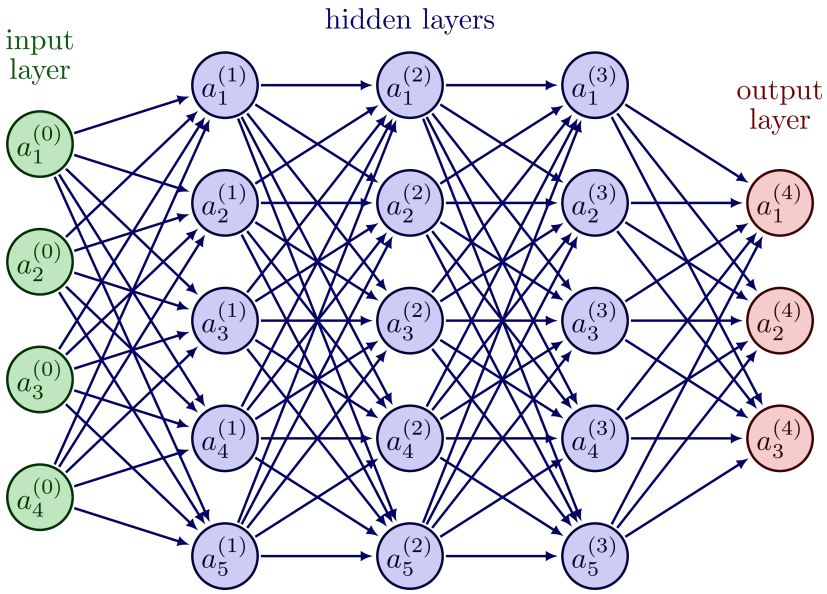

FIGURE 2.16: Schéma d'un perceptron à cinq couches : une d'entrée, trois cachés et une de sortie.

2.5.4 Réseaux convolutionnels

Nous nous sommes dans cette thèse concentrés sur les réseaux convolutionnels (CNN pour *convolutionnal neural networks*; (LeCun et al., 1989)) adaptés à la gestion de données ayant une topologie connue sous forme de grille comme les séries temporelles ou les images. Ils ont été inspirés par le fonctionnement du cortex visuel. À l'instar des perceptrons ils sont organisés en couches, mais se caractérisent par l'utilisation de couches convolutionnelles qui sont des filtres passant sur les données et opérant une convolution. Nous illustrons dans la Fig 2.17 une telle couche dans le cas de données en deux dimensions. Les poids et biais de cette couche convolutionnelle sont ainsi partagés entre les différentes parties de l'input. L'utilisation de tels filtres diminue la taille de la sortie. Si une sortie de même taille que l'entrée est désirée, il est nécessaire de recourir au padding qui consiste à étendre l'entrée par des valeurs à ses bords.

De nombreuses variantes aux couches convolutionnelles existent et peuvent être employé dans les CNN. Par exemple, bien qu'elles soient surtout utilisées pour traiter des images, les couches convolutionnelles peuvent être

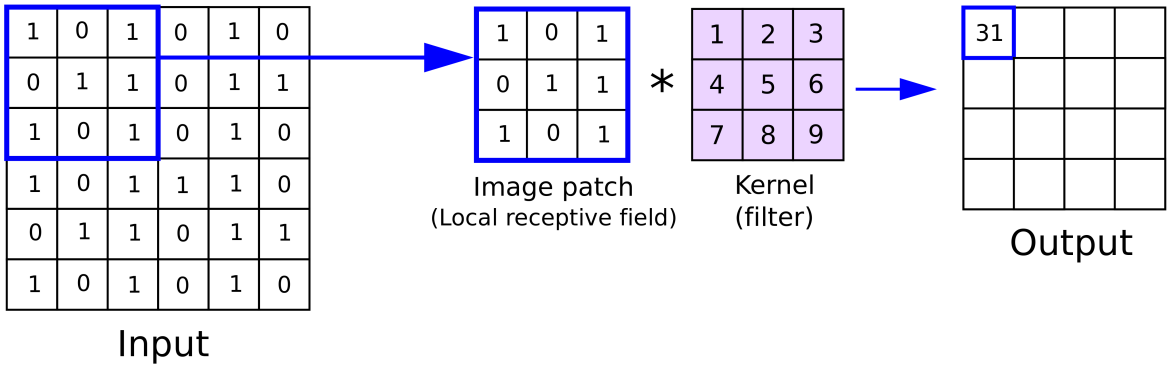

FIGURE 2.17: Schéma d'un filtre convolutionnel en deux dimensions de taille 3*3 appliqué à une image binaire.

employées pour un nombre quelconque de dimensions dans les données. Une autre variante couramment utilisée, y compris lors cette thèse, sont les couches convolutionnels transposées (aussi appelées couches de déconvolution) qui se composent également d'un filtre. Mais, au lieu de faire passer le filtre sur les données d'entrée, on fait passer les données d'entrée sur le filtre. Le résultat obtenu est donc d'une taille plus grande que l'input.

D'autres opérations peuvent avoir lieu dans un réseau convolutionnel comme le max-pooling (Zhou and Chellappa, 1988) qui consiste à sélectionner la sortie maximale dans un voisinage rectangulaire.

Un CNN est donc classiquement composé d'un ensemble de couches convolutionnelles avec potentiellement d'autres opérations simples comme le max-pooling. Nous illustrons dans la Fig 2.18 l'architecture d'un simple CNN composé de trois couches pour la classification en quatre classes.

Les poids d'un CNN étant liés au nombre et à la taille des filtres utilisés et non pas à la taille des données d'entrées, ils peuvent être utilisés pour faire des réseaux relativement simples ne contenant qu'un nombre réduit de poids. Ils peuvent être néanmoins utilisés pour faire des réseaux beaucoup plus complexes comme l'U-Net (Ronneberger et al., 2015). Ce

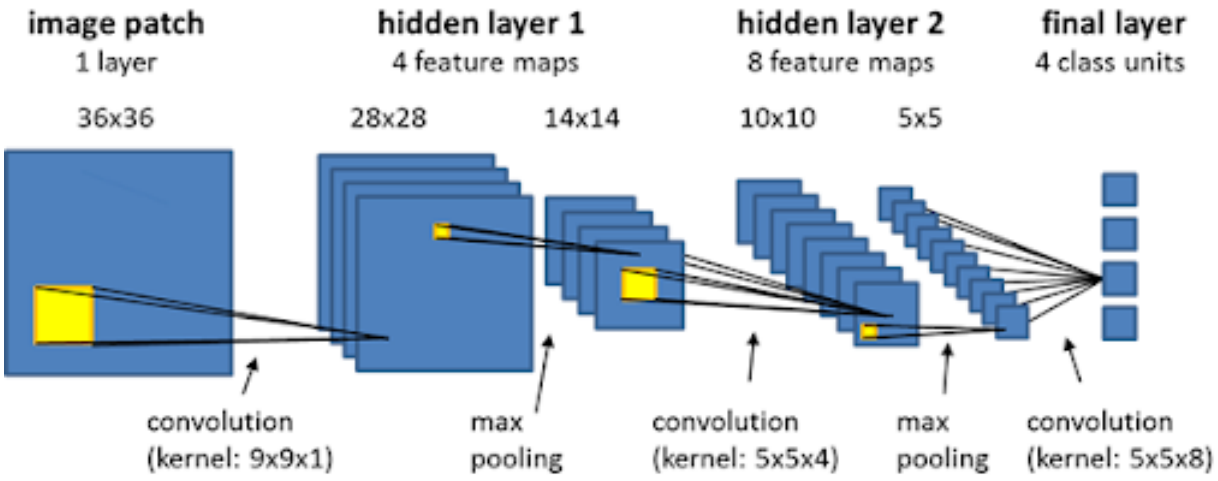

FIGURE 2.18: Schéma d'un CNN de classification constitué de quatre couches. Une d'entrée et de sortie faite d'une convolution et deux cachés constitués chacune d'une convolution et d'une opération de max-pooling.

réseau, utilisé dans le chapitre 3 ci-après, a été développé pour la segmentation d'images médicales, qui consiste à classifier les pixels d'une image selon certains critères afin de former des régions uniformes. Il fut appliqué à de nombreuses autres applications et est devenu une référence dans les applications d'images à images. Diverses modifications et versions alternatives à ce réseau sont utilisées, nous illustrons dans la Fig 2.19 l'architecture originelle de l'U-Net de Ronneberger et al., 2015. Son architecture se caractérise par une forme de U qui lui a donné son nom, alternant un chemin contractant, composé de couches convolutionnelles et d'opérations de max-pooling, et un chemin expansif composé de couches convolutionnelles standard et de couches convolutionnelles transposées. Des informations sont transmises directement entre ces deux chemins par des concaténations.

2.5.5 Intelligence artificielle explicable

Si les opérations d'un ANN sont individuellement simples, leur très grand nombre rend impossible une compréhension humaine du fonctionnement interne du réseau de neurone ou du rôle particulier d'un neurone ou d'une

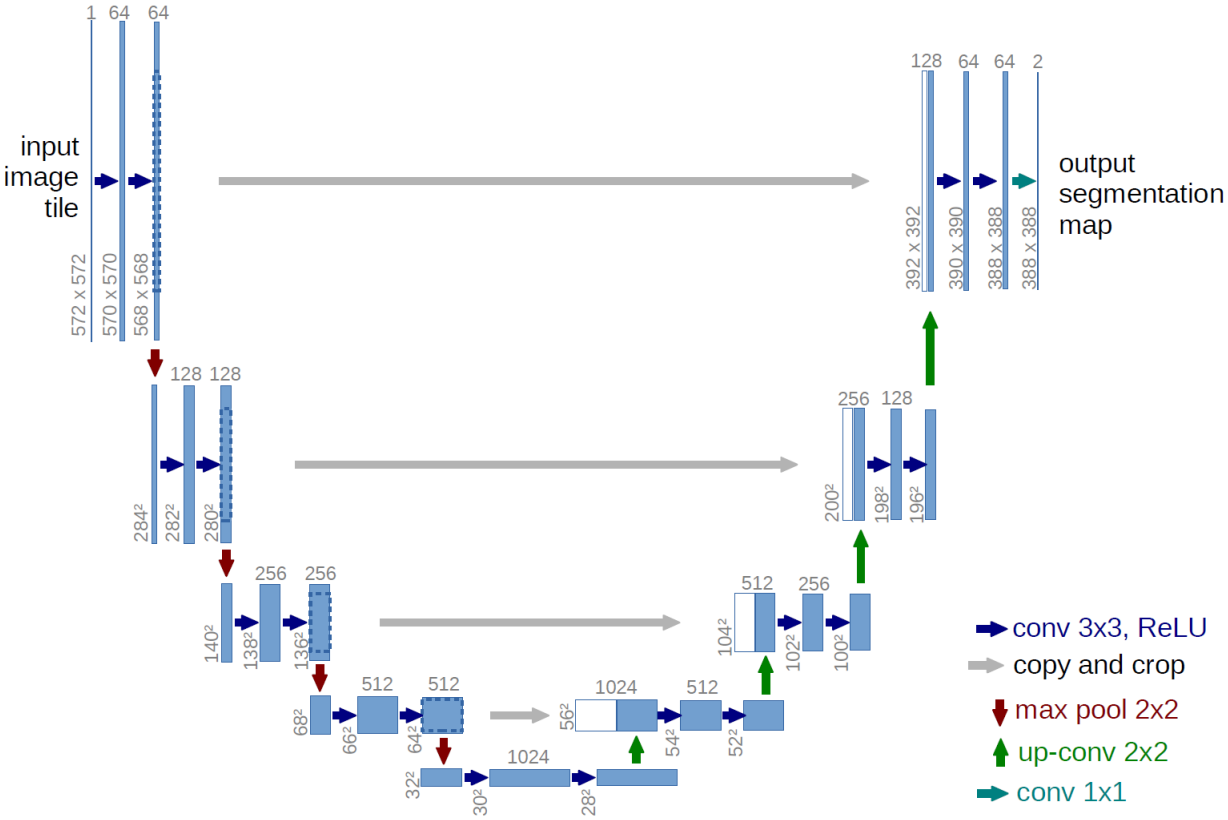

FIGURE 2.19: Architecture U-net. Chaque boîte bleue correspond à une carte de caractéristiques multicanale. Le nombre de canaux est indiqué en haut de la boîte. La taille x-y est indiquée sur le bord inférieur gauche de la boîte. Les cases blanches représentent des cartes de caractéristiques copiées. Les flèches indiquent les différentes opérations. Tiré de Ronneberger et al., 2015

couche. Cet effet est dit celui de la "boîte noire" et empêche l'utilisateur d'un réseau de neurones d'obtenir facilement une interprétabilité des résultats obtenus par l'ANN. Une des critiques majeures faite au machine learning fut le manque d'outil à disposition afin d'obtenir une telle interprétabilité (Fan et al., 2021). Le manque d'interprétation des résultats des ANNs peuvent les rendre moins dignes de confiance pour des applications nécessitant une compréhension et une grande confiance dans le modèle comme les applications médicales, la conduite de véhicule... Les sciences du climat ne font pas exception à ce besoin d'interprétabilité (Toms et al., 2020) pour vérifier les résultats obtenus.

Ce besoin d'interprétabilité à conduit au développement de l'intelligence artificielle explicable qui rassemble des techniques donnant une compréhension au fonctionnement des réseaux de neurones. Une partie de ces méthodes se concentrent sur le problème de la contribution des parties de l'input d'une entrée $x = (x1, x2, ..., xn)$ d'un ANN sur chaque composante de la sortie $y = (y1, y2...yn)$. La résolution de ce problème permet de comprendre quelle partie de l'input fut déterminante pour l'ANN dans le calcul de l'output.

Deux grandes familles de méthodes existent pour ce problème. La première famille de méthodes sont les méthodes de perturbation. Elles ont été appliquées notamment aux réseaux convolutionnel (Zeiler and Fergus, 2014) et fonctionnent en appliquant successivement des modifications aux différentes composantes de x pour évaluer l'impact sur la sortie de l'ANN. Ces méthodes ont le défaut de pouvoir être très lente si le nombre de composantes de x est grand comme pour une image. La seconde famille de méthode repose sur la rétropropagation qui permet de calculer pour chaque composante de $\hat{y}$ le gradient en fonction de tous les composants de x.

Ces méthodes ont été abordées dans un grand nombre de travaux scientifiques (Simonyan and Zisserman, 2014; Zeiler and Fergus, 2014; Bach et

al., 2015; Springenberg et al., 2014; Shrikumar et al., 2017; Sundararajan et al., 2017; Montavon et al., 2017; Zintgraf et al., 2017) et furent utilisés à différentes étapes de la thèse. Une variante de ces techniques basée sur la rétropropagation fut à la base de notre travail sur l'attribution du changement climatique (voir chapitre 4). Il s'agit de l'optimisation inverse. Cette technique vise à créer, pour un ANN entraîné, une entrée correspondante à une sortie donnée grâce à la rétropropagation.

Chapitre 3

Séparation de la variabilité interne et forcé

3.1 Introduction

Nous avons voulu, au cours de la thèse, développer une méthode pour séparer la variabilité interne de la variabilité forcée à l'aide d'un réseau de neurones. Cette étude s'est effectuée avec des sorties de modèles climatiques et les observations de la température de surface. Dans la partie suivante, nous résumons la méthodologie ainsi que les résultats principaux de cette méthode. Une présentation plus exhaustive de celle-ci se trouve dans un manuscrit soumis au *Journal of Advances in Modelling Earth Systems*.

Nous utilisons un réseau de neurones entrainé avec des sorties de température de surface de l'air provenant de simulations historiques et scénarios provenant de modèles climatiques issues de CMIP5, CMIP6 et SMILE.
Nous assimilons les simulations et observations de température de surface de l'air à des images. En effet, les réseaux de neurones ont été très utilisés dans le domaine de l'analyse d'image (Egmont-Petersen et al., 2002). Un des domaines de l'analyse d'image où les réseaux de neurones ont particulièrement été utilisés est le "débruitage" qui consiste à éliminer le bruit dans des

FIGURE 3.1: Exemple de pairs d'images d'entraînement dans une méthode *Noise to Noise*. Tiré de Lehtinen et al., 2018

images pour les « restaurer » (Ilesanmi and Ilesanmi, 2021; Tian et al., 2020). Le champ tridimensionnel (temps, latitude et longitude) des simulations ou des observations de température de surface est donc assimilé à une image tridimensionnelle. La variabilité interne est, elle, assimilé à un bruit semblable à celui pouvant se trouver sur des images et s'ajoutant à la variabilité forcée qui est, elle, assimilé à la « vraie image ». Nous avons donc recouru à une méthodologie de débruitage utilisant un réseau de neurones et créée pour les images que nous adaptons à notre problème climatique.

Cette méthodologie de débruitage se nomme *Noise to Noise* (Lehtinen et al., 2018) et possède la particularité d'utiliser uniquement des images bruitées. Cela est intéressant pour notre cas, car toutes les simulations climatiques et les observations sont affectées par la variabilité interne. Elle consiste à utiliser des images bruitées de différents objets. Chaque objet possède un certain nombre d'images bruitées qui lui est associé. La différence entre ces images tient donc au bruit et pas au « vrai signal ». L'entraînement du réseau de neurone se fait alors en utilisant comme couples *input/output* toutes les combinaisons d'images du même objet. Nous illustrons dans la figure 3.1 des exemples de paires d'images d'entraînement dans le contexte de vraies images. Chaque image est affectée par un bruit différent (un texte aléatoire superposé

à l'image). La tâche est par nature impossible : le réseau de neurones ne peut pas apprendre à transformer une instance aléatoire de bruit en une autre. À la place, le réseau de neurones va être conduit à reproduire l'espérance mathématique de l'objet qui est par hypothèse proche de l'objet sans bruit. Nous illustrons dans la figure 3.2 un exemple de ce que le débruitage *Noise to Noise* peut atteindre sur de vraies images affectées par un bruit gaussien. Nous pouvons voir que qualitativement, malgré un bruit important se trouvant sur l'image d'entrée, l'image retrouvée par la méthodologie *Noise to Noise* est très proche de la véritable image.

Dans notre cas, les objets sont les variabilités forcées des modèles clima-

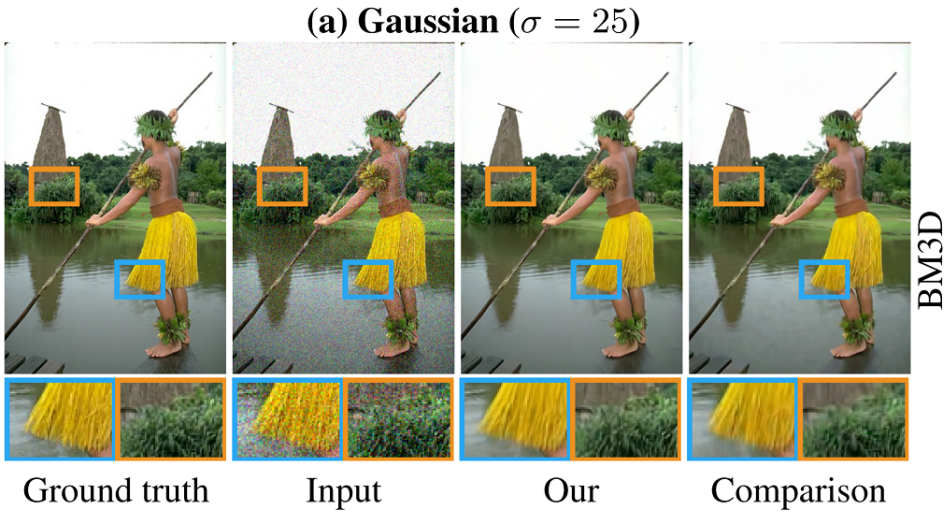

FIGURE 3.2: Exemple de résultats de la méthode *Noise to Noise* sur une image flouté par un bruit gaussien. Tiré de Lehtinen et al., 2018

tiques ou du climat réel. Les images bruitées sont les membres historiques ou les observations de la température de surface de l'air. Le réseau de neurones utilisé est un U-Net (comme décrit dans le Chapitre 2) modifié pour traiter des images tridimensionnelles. La capacité de filtrage de la variabilité interne de l'U-net est évalué en utilisant deux modèles climatiques de SMILE non utilisés dans l'entraînement de l'U-Net : FGOALS-g3 et MPI-ESM ainsi qu'en utilisant la méthode de la moyenne d'ensemble.

L'U-Net réduit l'effet de la variabilité interne d'un facteur $\sim$4 dans les modèles FGOALS-g3 et MPI-ESM, bien que de fortes disparités spatiales existent. En effet, l'U-Net montre une grande capacité à retirer ENSO mais se montre moins performant dans l'hémisphère nord où les variabilités interne et forcée sont plus importantes. De plus, l'U-Net capture moins bien la variabilité forcée de l'Atlantique Nord pour le modèle FGOALS-g3. Ceci peut s'expliquer car le motif spatial retrouvé par l'U-Net est proche de la moyenne multimodèle des modèles climatiques de la base d'entraînement. Cela rend donc plus complexe d'identifier un motif spécifique de variabilité forcée spécifique, et donc peu représenté dans le consensus multimodèle. Quand l'U-Net est appliqué aux observations, on observe une nette réduction de la variabilité interannuelle, bien que les effets des différentes éruptions volcaniques soient bien retranscrits.

Cet article présente une méthode de séparation de la variabilité interne et forcé basé sur la méthode « *Noise to Noise*». Cette méthode est à notre connaissance la première utilisation du *machine learning* pour ce problème. Cette méthode doit donc être comparée aux autres méthodes de séparation de la variabilité interne et forcée du climat présentées en section 2.3 et utilise à la fois les modèles climatiques et les observations. Cette méthode peut être vue comme une méthode d'attribution estimant l'effet de tous les forçages sans pouvoir les distinguer entre eux. Les travaux suivants porteront sur l'attribution de ces différents forçages.

Separation of internal and forced variability of climate using a U-Net

Constantin Bône[1,2], Guillaume Gastineau[1], Sylvie Thiria[1], Patrick Gallinari[2,3] and Carlos Mejia[1]

[1]UMR LOCEAN, IPSL, Sorbonne Université, IRD, CNRS, MNHN
[2]UMR ISIR, Sorbonne Université, CNRS, INSERM
[3]Criteo AI Lab

Key Points:

- We present a new method to separate the forced and internal variability of the surface air temperature.
- We utilise a U-Net trained with global climate models outputs and implement a noise to noise methodology to eliminate internal variability.
- The results are assessed through the utilisation of very large ensemble simulations of two distinct climate models.

Corresponding author: Constantin Bône, constantin.bone@sorbonne-universite.fr

Abstract

The internal variability pertains to fluctuations originating from processes inherent to the climate component and their mutual interactions. On the other hand, forced variability delineates the influence of external boundary conditions on the physical climate system. A methodology is formulated to distinguish between internal and forced variability within the surface air temperature. The noise-to-noise approach is employed for training a neural network, drawing an analogy between internal variability and image noise. A large training dataset is compiled using surface air temperature data spanning from 1901 to 2020, obtained from an ensemble of Atmosphere-Ocean General Circulation Model (AOGCM) simulations. The neural network utilized for training is a U-Net, a widely adopted convolutional network primarily designed for image segmentation. To assess performance, comparisons are made between outputs from two single-model initial-condition large ensembles (SMILEs), the ensemble mean, and the U-Net's predictions. The U-Net reduces internal variability by a factor of four, although notable discrepancies are observed at the regional scale. While demonstrating effective filtering of the El Niño Southern Oscillation, the U-Net encounters challenges in areas dominated by forced variability, such as the Arctic sea ice retreat region. This methodology holds potential for extension to other physical variables, facilitating insights into the enduring changes triggered by external forcings over the long term.

Plain Language Summary

To comprehensively grasp future climate change, it becomes imperative to differentiate between forced variability and internal climate variability. Internal variability refers to the climate's variations driven by the chaotic nature of geophysical fluids. Conversely, forced variability denotes changes prompted by external forcings, predominantly alterations in radiative forcing, primarily due to anthropogenic activities. Here, a novel approach is introduced for filtering internal variability through the utilisation of a convolutional neural network. This neural network is trained using a noise-to-noise methodology, targeting the filtration of internal variability from surface air temperature outputs of climate models or observational data. Internal variability is treated analogously to noise within an image, which is removed to restore the "true image," corresponding to forced variability in our case. This method capitalises on the data generated by state-of-the-art climate models through the coupled model intercomparison project (CMIP). To val-

idate this methodology, we assess its performance using very large ensembles of climate model simulations, enabling precise estimation of forced variability. Our findings demonstrate a reduction in internal variability by a factor of four, accompanied by notable regional variations.

1 Introduction

The phenomenon of climate warming is characterized by an elevated surface air temperature, notably reaching a pivotal juncture during the latter half of the twentieth century (Eyring et al., 2021). Nevertheless, the observed anomalies in surface air temperature arise from a dual spectrum of variabilities. The first source of variability is due to the effect of the external forcings, such as the increase in the greenhouse gases concentration, the variations of concentration in anthropogenic and natural aerosols, the fluctuations in solar variability or volcanic eruptions and the land-use changes. The related variability is designated as the forced variability. The second source of variability is coming from processes internal to the atmosphere, oceans, cryosphere and land or the interactions between them (Cassou et al., 2018). Subsequently, this form of variability is referred to as 'internal variability,' encapsulating its inception within the climate system and its persistence even without alterations in external forcings. Despite the overarching dominance of forced variability in shaping the broad-scale and long-term trajectory of surface air temperature across the 1900-2020 timeframe (Deser et al., 2012; Kay et al., 2015), a comprehensive understanding of the distinct contributions of internal and forced variability remains elusive. Internal variability takes center stage in briefer temporal scales and smaller spatial dimensions. For instance, the leading mode of internal variability in global air surface temperature manifests as the El Niño Southern Oscillation (ENSO), characterized by significant anomalies in the equatorial Pacific Ocean, accompanied by distant teleconnections, and a prevailing cycle spanning two to seven years (Wang & Picaut, 2004). Additionally, the interdecadal Pacific variability (Newman et al., 2016) and the Atlantic Multidecadal variability (Zhang et al., 2019) wield the capacity to influence climate dynamics across the decadal to multidecadal spectrum. A notable example involves the deceleration in the global warming rate experienced during 2002-2012, commonly referred to as the global warming hiatus, which has been robustly linked to Interdecadal Pacific Variability (Meehl et al., 2013; Kosaka & Xie, 2013; England et al., 2014). Lastly, internal variability exercises influence even over centennial and

multi-centennial spans (Jiang et al., 2021; S. Li & Huang, 2022) exerting substantial impact on trends within the 1900-2015 interval (Bonnet et al., 2022).

The distinction between forced variability and internal variability is essential for conducting detection and attribution studies, enabling accurate estimation and simulation of the climate's reaction to alterations in radiative forcing. Moreover, this differentiation aids in recognizing and comprehending internal climate variability. Nevertheless, the availability of instrumental observations is limited to the period since 1850, and the relatively brief duration of these observations presents challenges in effectively and confidently discerning internal variability.

For identifying both internal and forced variability, linear trends (Swart et al., 2015; Vincent et al., 2015) or quadratic trends (Enfield & Cid-Serrano, 2010) have been employed to characterize forced variability. However, linear or quadratic trends inadequately capture the temporal evolution of temperature, particularly failing to account for the abrupt cooling subsequent to significant volcanic eruptions, which hold significant climate impact (Schmidt et al., 2018). Additional approaches include the application of Empirical Orthogonal Functions (EOF) analysis (Parker et al., 2007), low-frequency pattern filtering (Wills et al., 2020), and linear inverse models (Marini & Frankignoul, 2014). These techniques deconstruct forced variability into a combination of modes featuring distinct patterns and corresponding time series. Regression analysis of the global mean surface temperature (GMST) has also been employed, although this may inadvertently establish misleading links between the Atlantic and Pacific basins (Frankignoul et al., 2017; Deser & Phillips, 2023). However, a comprehensive and systematic examination of these methodologies remains notably absent.

Climate model simulations have been employed to overcome the limitations of sparse observation sampling. Conducting an ensemble of climate model simulations with diverse initial conditions enables estimation of forced variability via the ensemble mean. This approach effectively mitigates the variance linked to internal variability by a factor of n, where n signifies the ensemble's size (Harzallah & Sadourny, 1995; Hawkins & Sutton, 2009; Ting et al., 2009; Solomon et al., 2011; Deser et al., 2014; Frankcombe et al., 2015). As a result, modeling centers have undertaken substantial ensembles with over 20 or 30 ensemble members (Jeffrey et al., 2013; Rodgers et al., 2015; Sun et al., 2018; Deser et al., 2020). These large ensembles are commonly referred to as Single-Model Initial-

Condition Large Ensembles (SMILE; Deser et al. (2020)). Multiple SMILE initiatives have been undertaken using models such as CCSM3 (Collins et al., 2006), CCSM4 (Gent et al., 2011), CESM (Kay et al., 2015), MPI-ESM (Maher et al., 2019), FGOALS-g3 (Li et al., 2020), CanESM2 (Chylek et al., 2011), and IPSL-CM6A-LR (Bonnet et al., 2021), among others. This offers a valuable dataset for crafting methodologies dedicated to the disentanglement of forced and internal variability. Notably, employing members of a large ensemble model as surrogate observations allows for a comparison of results with the ensemble mean. Differences primarily mirror residual internal variability or limitations inherent in the method.

Nevertheless, the forced variability estimated through an ensemble mean remains contingent upon the specific climate model employed. These climate models carry substantial uncertainties, particularly in terms of their climate sensitivity (Sherwood et al., 2020), often attributed to factors like uncertain cloud retroaction which significantly impact equilibrium climate sensitivity (Zelinka et al., 2016). Additionally, significant uncertainties surround historical emissions and the linked radiative forcing from aerosols (Menary et al., 2020; C. J. Smith & Forster, 2021). Moreover, the internal variability exhibited by different models also varies significantly (Parsons et al., 2020).

Several methodologies have been devised to harness data from diverse climate models, as employing a multi-model approach holds the potential to alleviate the uncertainties inherent in individual climate models. Multi-model ensemble means are widely adopted for estimating the forced signal (Steinman et al., 2015). Notably, techniques such as the signal-to-noise-maximizing empirical orthogonal functions (Ting et al., 2009; Wills et al., 2020) and the discriminant analysis and maximization of the average predictability time (DelSole et al., 2011) have been put forth to extract forced variability with superior efficacy compared to ensemble means. Furthermore, scaling techniques that adjusts the forced signal from models using observational data have been proposed. Among these methodologies are fingerprinting methods grounded in linear regression, commonly applied for detecting and attributing climate change with a unified forcing that encapsulates the influence of all external forcings (Hasselmann, 1993; Allen & Tett, 1999; Allen & Stott, 2003). More recently, the use of scaling factors was also proposed by Frankcombe et al. (2015).

This paper introduces an alternative approach to distinguishing internal and forced variability using climate model data, employing a non-linear method that takes into account the spatio-temporal data covariances. This method is rooted in a neural network trained on data from Atmosphere-Ocean General Circulation Models (AOGCMs). Among the areas where neural networks have excelled is image analysis (Egmont-Petersen et al., 2002). One of the prominent applications of neural networks in image processing is image denoising, involving the elimination of noise from an image to restore its true form (Ilesanmi & Ilesanmi, 2021; Tian et al., 2020). In this context, internal variability is treated as noise. It is demonstrated that machine learning image denoising methodologies can subsequently isolate forced variability. The internal variability is eliminated, leaving behind a quantifiable residue. This method leverages the temporal and spatial information inherent in climate models to establish the weights and biases of a neural network. With these parameters in place, the neural network is also employed with observations to delve into and attribute the progression of climate change since 1905 to 2016. To the best of our knowledge, this represents the pioneering application of a dedicated neural network for the purpose of disentangling internal and forced variability.

The structure of this paper is as follows: Section 2 outlines the data utilized. Section 3 introduces the method anchored in a neural network. Section 4 assesses the method's performance. In Section 5, the neural network method is applied to observations. Lastly, Section 6 offers the conclusion and discussion.

2 Data

2.1 Observations

The gridded monthly Surface Air Temperature anomaly (SAT) from 1901 to 2020, as provided by GISS Surface Temperature Analysis version 4 (GISTEMP; Hansen et al. (2010); Lenssen et al. (2019)), is employed in this study. GISTEMP amalgamates meteorological station data over land (NOAA GHCN v4) with sea surface temperature (SST) estimates from ERSST v5. This data is available on a consistent 2°x2° grid. The monthly values are aggregated to calculate annual means, and the SAT anomalies are determined using the reference period 1950-2014.

2.2 Climate model simulations

The monthly SAT data is sourced from historical simulations within the Coupled Model Intercomparison Project Phase 5 (CMIP5; Taylor et al. (2012)) and the Coupled Model Intercomparison Project Phase 6 (CMIP6; (Eyring et al., 2016)), along with several Single-Model Initial-Condition Large Ensembles (SMILEs) from distinct models: MPI-ESM (Maher et al., 2019), CSIRO-Mk3-6-0 (Collier et al., 2011), EC-Earth (Döscher et al., 2021), and FGOALS-g3 (Li et al., 2020). For the historical simulations, spanning 1901 to 2005 (2014) for CMIP5 (CMIP6), all external forcings are integrated. These forcings encompass the effects of historical greenhouse gas concentrations, anthropogenic and natural aerosols, stratospheric ozone, solar activity, and land-use changes. Each climate model delivers multiple realizations referred to as ensemble members, generated through distinct initial conditions. From 2005 (2014 for CMIP6) until 2020, the outputs under the pessimistic Representation Concentration Pathway 8.5 (RCP8.5) scenario for CMIP5 (Van Vuuren et al., 2011) and the intermediate Shared Socio-economic Pathway 2 4.5 (SSP2-4.5) for CMIP6 (Tebaldi et al., 2020) are employed. These simulations utilize socio-economic assumptions to project future external forcing patterns. Additionally, several SMILEs are incorporated, employing distinct historical forcings or scenario simulations of CMIP5 or CMIP6 (elaborated in Table S3). While minor differences are anticipated in external forcing between CMIP5 and CMIP6 simulations, notable uncertainties arise in aerosol emissions (C. J. Smith et al., 2020; Fyfe et al., 2021). Modest differences may also emerge between the RCP8.5 (strong) and SSP2-4.5 (moderate) scenarios, particularly until 2020, where actual forcings mirror observed forcings to a considerable extent (Masson-Delmotte et al., 2021).

The count of members accessible for scenario simulations is fewer compared to the historical counterparts. Therefore, we extended the outputs from historical experiments using the scenario ensemble member of the same model with the same number identification. In case the number identification is lacking, we select randomly an scenario ensemble member of the same climate model.

All monthly data are aggregated into annual means. Subsequently, the SAT anomalies are computed for each ensemble member using 1950-2014 as a reference period. This furnishes a multi-model ensemble comprising 801 members derived from 47 AOGCMs. Subsequently, the concatenated historical and scenario members are harnessed within

the 1901-2020 timeframe. All model data is regridded using bilinear interpolation on the horizontal grid from GISTEMP. The details pertaining to the climate model names, ensemble sizes, and the names of the employed scenario simulations are elucidated in Tabs. S1, S2, and S3.

2.3 Validation of the data set

The forced variability simulated within the multi-model ensemble is succinctly examined for two specific data subsets. We investigate the MPI-ESM and FGOALS-g3 climate models from SMILE, as they have a very large size of 100 and 115 members, respectively, which largely exceed the size of other model ensembles. Anticipatedly, the estimated forced variability derived from the ensemble mean for each of these models is expected to be accurate, as the reduction in variance attributed to internal variability reaches 100 and 115, respectively. For instance, Deser et al. (2012, 2014) demonstrated that identifying regional climate responses on time scales of several decades may necessitate between 10 to 40 members. Specifically, to detect a change in SAT between the decades 2005-2014 and 2028-2037 on a global scale, the use of 3 to 6 members is requisite. This requirement can surge beyond 10 for local analyses such as in North America. Subsequently, the data originating from these two models is subsequently employed to appraise the outcomes of the neural network model in section 4.1.

We utilize the ensemble mean to characterize the forced variability and employ the standard deviations from the ensemble members for evaluating the internal variability. Figure 1 illustrates the standard deviation of the SAT deviation from the ensemble mean for FGOALS-g3 and MPI-ESM. The variability in SAT is more pronounced over land surfaces ($\sim$0.3°C) compared to oceans ($\sim$0.1°C), consistent with the lower thermal inertia of land. Notably, substantial variability (ranging from approximately 1.5°C to 2.5°C) is observed over regions coinciding with the sea ice edge, such as the Bering Sea and Nordic Seas in the Northern Hemisphere, as well as the Amundsen and Weddell Seas in the Southern Hemisphere. Additionally, a marked variability is observed in the equatorial Pacific Ocean, with a standard deviation of 0.8°C, and this variability is more prominent in MPI-ESM compared to FGOALS-g3. A localized peak of variability is situated over the subpolar North Atlantic, especially notable for FGOALS-g3 (reaching up to 2°C). These outcomes coherently reflect a significant internal variability stemming from extratropical weather fluctuations over land surfaces, exhibiting local maxima around regions adjacent to the

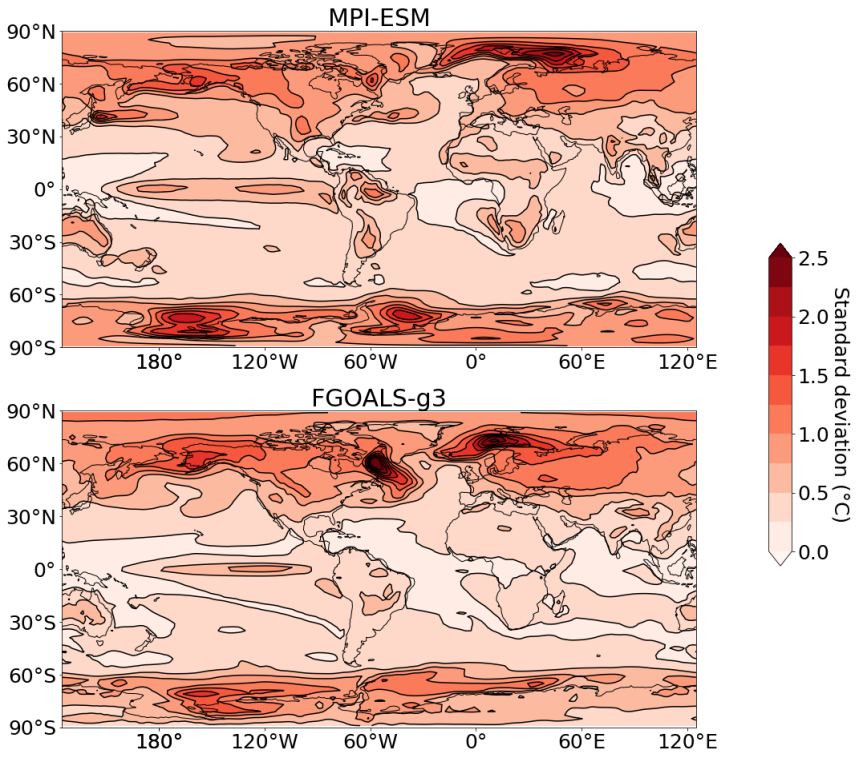

Figure 1. Standard deviation of the SAT deviations from the ensemble mean for (top) MPI-ESM and (bottom) FGOALS-g3.

sea ice edge. Moreover, the variability observed in the equatorial Pacific mirrors the phenomenon of El Nino Southern Oscillation (Neelin et al., 1998).

The forced variability is estimated through the ensemble mean of each model. Subsequently, the multi-model mean (MMM) is computed by averaging the ensemble means across all models, ensuring equal weight for each model. Nonetheless, MPI-ESM and FGOALS-g3 are excluded from this computation, as the intention is to later compare them to the MMM. To assess the prominent impact of greenhouse gas forcing, Figure 2 (a, c, e) illustrates the ensemble mean SAT anomaly for MPI-ESM, FGOALS-g3, and the MMM throughout the 2010-2020 interval. Furthermore, Figure 2 (b, d, f) presents the temporal standard deviation of the ensemble means across the period from 1901 to 2020. As anticipated, all climate models project more substantial warming over land (up to 0.8°C) than over oceans (approximately 0.3°C). Notably, the Arctic exhibits an amplification of global warming, with warming exceeding 2°C north of 60°N. The MMM showcases an average warming of 0.8°C for the 2010-2020 period, surpassing MPI-ESM (0.64°C) and

FGOALS-g3 (0.69°C). This aligns with the comparatively lower equilibrium climate sensitivity (ECS) of these two models (3.6°C for MPI-ESM and 2.8°C for FGOALS-g3) when compared to other models employed in this study (Zelinka et al., 2020). Within the subpolar Atlantic, the SAT anomalies exhibit a minimum, with negative temperatures anomalies observed in FGOALS-g3 over the Labrador Sea, or in MPI-ESM over the subpolar gyre. This phenomenon, known as the North Atlantic warming hole (Keil et al., 2020), is associated with a deceleration of the Atlantic meridional overturning circulation (He et al., 2022). It is worth noting that such a minimum is less pronounced in the MMM, presumably due to considerable uncertainties regarding the precise location of this warming hole and the linked processes. An equivalent spatial pattern can be derived using standard deviations, revealing values of approximately 0.3°C for the majority of global regions and higher values over land (∼0.6°C). Grid points located north of 60° also exhibit elevated values, peaking at around 2°C in the Barents Sea for MPI-ESM or the Labrador Sea for FGOALS-g3.

The forced variability exhibited by MPI-ESM and FGOALS-g3 diverges from that of the MMM, revealing a comparatively weaker global warming trend and standard deviation pattern. This divergence is particularly evident north of 60°N, where the warming exhibits greater amplification (refer to Fig. 2), amounting to 1.54°C for MPI-ESM and 1.45°C for FGOALS-g3. Local variations are also observed in regions such as the Labrador Sea, Barents and Kara Sea, the Canadian archipelago, and the Bering Sea in the case of FGOALS-g3. Notably, MPI-ESM similarly presents notable differences in the Barents Sea. These discrepancies may arise from biases related to sea ice representation. Specifically, FGOALS-g3 depicts an excessive extent of Arctic sea ice (Li et al., 2020), which in turn leads to inaccuracies in simulating the location of the sea ice edge. This discrepancy can account for spurious SAT variability attributed to the misplaced sea ice edge within the Labrador Sea. The mean standard deviation of the ensemble mean registers as 0.34°C for MPI-ESM and 0.43°C for FGOALS-g3, exceeding the mean standard deviation of the SAT deviations of the members to the ensemble mean which is of 0.51°C for MPI-ESM and 0.46°C for FGOALS-g3. This underscores that the internal variability is marginally more pronounced than the forced variability.

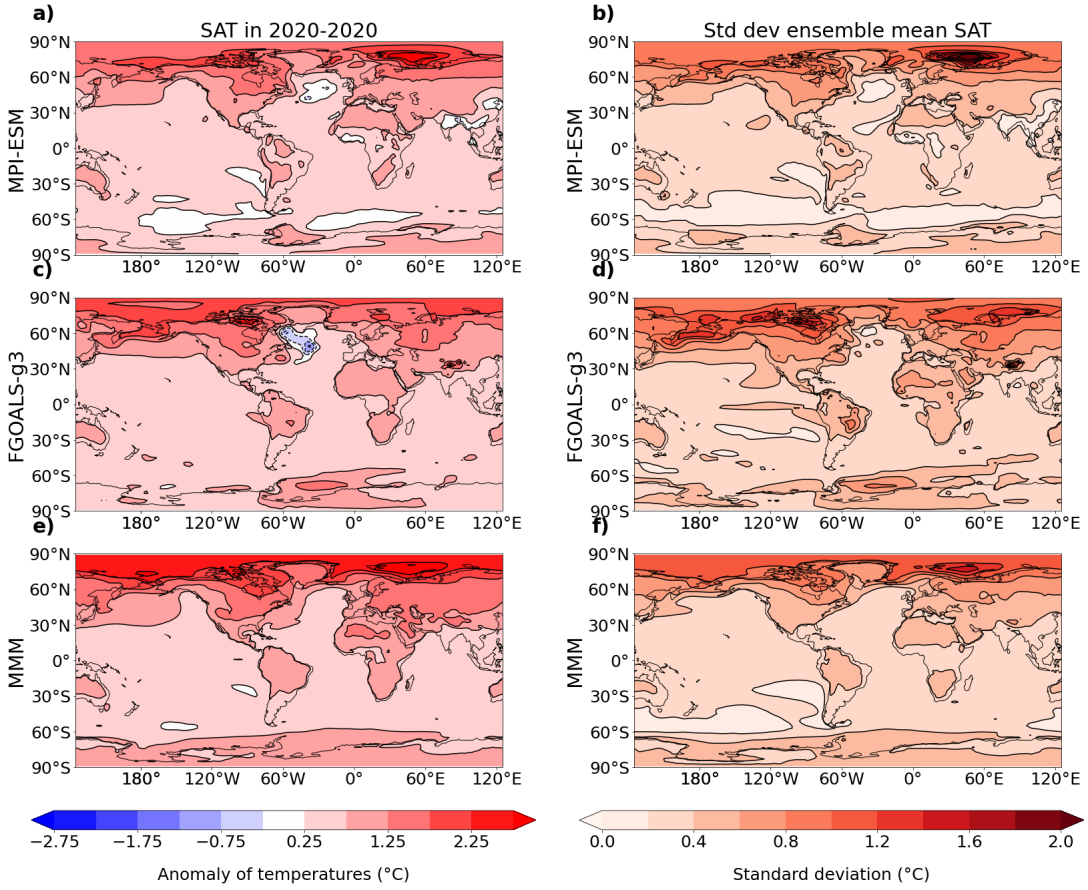

Figure 2. a) Ensemble mean of the air surface temperature (°C) in MPI-ESM in 2010-2020. c) Same as a) but for FGOALS-g3. e) Same as a) but for the MMM. b) Standard deviation of the ensemble mean surface air temperature (°C) in 1901-2020 for MPI-ESM. d) Same as b) but for FGOALS-g3. f) Same as b) but for the MMM.

3 Methods

3.1 Neural network

We design a neural network to remove the internal variability from the SAT. The input data is structured with dimensions (120, 90, 180), corresponding to time spanning from 1901 to 2020, latitude, and longitude, respectively. On the other hand, the output holds dimensions of (112, 90, 180), encompassing the years 1905 to 2016, while maintaining the latitude and longitude dimensions intact. Notably, the output's temporal span is truncated compared to the input, by excluding the initial and final four years. This reduction addresses the substantial uncertainty typically observed at the dataset's endpoints, an aspect that will be elaborated upon later.

A neural network's characteristics are shaped by its hyperparameters, which dictate both its architecture and training process. Our approach involves utilizing three distinct datasets, each composed of input and desired output pairs. The training dataset serves the purpose of establishing the neural network's weights and biases. Meanwhile, the validation dataset comes into play for estimating the hyperparameters. Finally, the test dataset is employed to assess the neural network's performance.

3.2 Constitution of the database

To construct the training dataset, we adapt a noise-to-noise methodology originally introduced in Lehtinen et al. (2018). This approach was initially designed to train a neural network in denoising images. In this method, the network is exclusively trained on noisy images depicting various objects. Each object has more than one noised image depicting it. In the noise to noise method, we create an input/output training database that comprises pairs of noisy image combinations for identical objects. It's essential to note that the network cannot effectively learn to transform a random noise realization into another. Instead, the configuration is designed to approximate the mathematical expectation of all noisy images associated with the same object, culminating in an estimate that closely resembles the noise-free image.

For our application, we consider the forced spatio-temporal SAT anomalies from each climate model as distinct objects. These anomalies, inherent to each member, can be likened to noisy images, where the internal variability introduces the noise compo-

nent. The ensemble members' mathematical expectation equates to the forced variability, which can be approximated through the ensemble mean.

To create the training dataset, we follow a procedure wherein we compute pairs of members for each climate model, except for MPI-ESM, FGOALS-g3, and MIROC6, which are reserved for testing and validation purposes. Adopting an approach similar to Lehtinen et al. (2018), we augment the dataset by introducing the ensemble mean of the climate model's members as an additional member. This inclusion serves to expedite the training process without introducing any other influences. In this process, each pair of members becomes an input/output pair. If we denote the number of ensemble members obtained from a specific climate model as n, this approach yields $n(n+1)$ input/output pairs per model. By accumulating such pairs from all models, the resulting training dataset primarily comprises simulations characterized by the most extensive ensemble sizes (namely IPSL-CM6A-LR, CanESM5, CNRM-CM6-1, and ACCESS-ESM1-5).

To create the validation set, we employ the ensemble simulation data from the MIROC6 model, which ranks as the third-largest ensemble in terms of size (with $n = 50$ members). For this purpose, we designate the ensemble members as inputs, while the ensemble mean spanning the period from 1905 to 2016 serves as the desired output.

To form the test dataset, we draw upon data derived from the FGOALS-g3 and MPI-ESM models, leveraging their extensive ensemble sizes of $n = 110$ and $n = 100$ respectively. Subsequently, we proceed to make comparisons between the outputs of the neural network obtained from ensemble members and their corresponding ensemble means for both of these models.

The conclusions drawn from these tests and validation processes may exhibit some dependence on the specific model being analyzed, as alternative models could yield varying outcomes. Nevertheless, this approach has been chosen due to its simplicity and its potential to mitigate the impact of any remaining internal variability.

3.3 U-Net

Convolutional neural networks (CNNs, Yamashita et al. (2018)) constitute a category of non-linear neural networks, notably applied in tasks related to imagery (O'Shea & Nash, 2015). A distinctive attribute of CNNs is their utilization of convolutional layers, which incorporate a trainable kernel that slides across the input data.

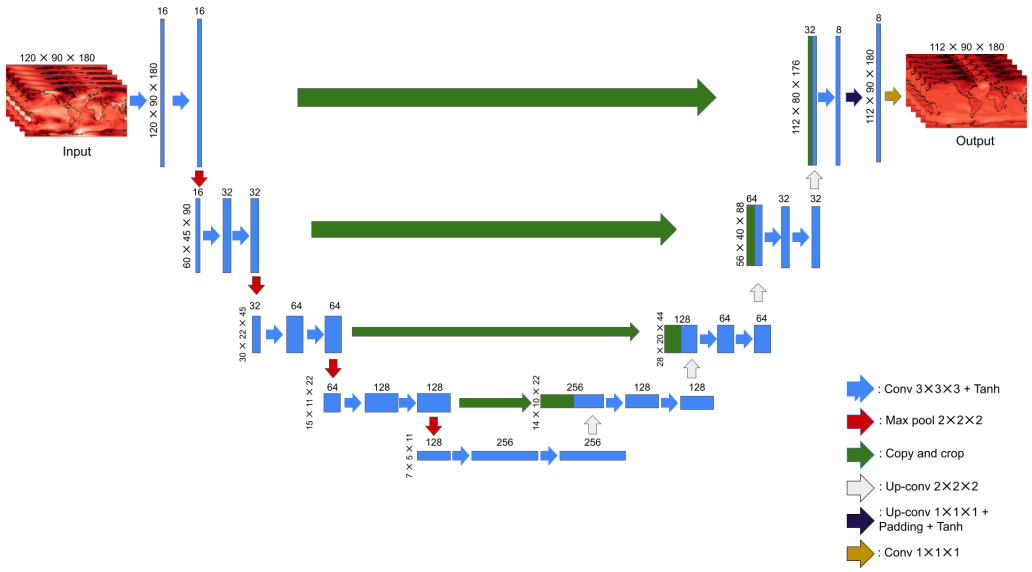

Figure 3. Schematic of the U-Net. The arrows represent the operations within the network. The numbers shows the dimension of the data and the number of filters used.

In this context, a U-Net architecture is employed, which falls within the realm of CNNs. Originally introduced by Ronneberger et al. (2015) for image segmentation, the U-Net structure has gained widespread popularity in image-related analyses such as denoising (Ilesanmi & Ilesanmi, 2021; Tian et al., 2020). The U-Net architecture is characterized by its inclusion of a contracting path and an expansive path, which collectively give rise to its characteristic U shape (refer to Fig. 3). The contracting path adheres to a conventional design of a convolutional network, featuring numerous convolutional layers, each followed by an activation function and a max-pooling operation. As the contracting path advances, spatial information is diminished while feature information is enriched. Conversely, the expansive path amalgamates feature and spatial information through a sequence of up-convolutions and concatenations with high-resolution features derived from the contracting path.

The U-Net architecture employed in this study shares similarities with the design proposed by Ronneberger et al. (2015). However, a modification is made by replacing the 2-dimensional convolutional layers with 3-dimensional counterparts. This alteration is introduced to encompass not only the spatial dimension but also the temporal dimension of the data. The selected activation function is the hyperbolic tangent. Additionally, adaptations have been made to the output layer to accommodate an output com-

prising 112 time steps. The neural network is comprised of a total of 5,659,009 trainable parameters.

A batch size of 8 is chosen, and the optimization process employs the Adam optimizer with a learning rate of 0.001. To ensure proper application of the CNN to the data, padding is introduced. This involves extending the image by appending zero values at its edges. For the longitudinal dimension, which is periodic, the zero padding only results in a slight discontinuity at 180°E, the edge of the data. Indeed, due to the nature of convolutional layers, a U-Net has more difficulty processing information located at the edge of the data. This is the reason why we excluded the initial and final four years (1901-1904 and 2017-2020) in the U-Net's outputs. The chosen cost function is the root mean squared error (RSME), calculated using an area-weighted mean of the gridded data.

The validation dataset is utilized to determine the optimal values for two key hyperparameters: the number of epochs and the number of filters used in the convolutional layers. The term "number of filters" pertains to the thickness of the convolutional layers. The number of epochs refers to how many times the training dataset is processed during the training phase. These hyperparameters are selected to minimize the root mean squared error (RMSE) using the validation dataset. Examination of the validation RMSE for different values of epochs and layer thickness reveals a consistent pattern (see Fig. S1): a significant reduction in RMSE occurs in the initial epochs, followed by a gradual increase. As a result, we settle on a layer thickness of 16 for the first layer (as shown in Fig. 3) and a total of 32 epochs.

3.4 Example

Figure 4 provides an illustrative example featuring two randomly selected ensemble members from MPI-ESM and FGOALS-g3. The comparison focuses on the SAT at the year 2016, depicted in the top panels, as well as the resulting output generated by the neural network in 2016 (centre panels), juxtaposed against the ensemble mean anomaly for the same year (bottom panels). The anticipated impact of elevated greenhouse gas concentrations in 2016 is evident in the SAT of both MPI-ESM and FGOALS-g3 members, which exhibit warm anomalies. However, the internal variability introduces anomalies that surpass those of the ensemble mean in numerous regions, accompanied by some negative anomalies in other areas. To elaborate, an instance of cooling is simulated across

the Equatorial Pacific Ocean, possibly linked to a La Niña event in the case of MPI-ESM. The same ensemble member displays cooling over land in equatorial Africa, South-Eastern Asia, and Australia, as well as in extratropical zones like the North Atlantic Ocean and the Weddell Sea. In the example from FGOALS-g3, cold anomalies emerge over the Nordic Seas and the Labrador Sea. Such cooling diverges from the ensemble average, which exhibits a relatively uniform warming pattern across the globe, with a more pronounced effect over landmasses. Notably, the Arctic and its environs experience heightened warming compared to other global regions, due to polar amplification. Conversely, minimal warming is observed in the Southern Ocean and the subpolar North Atlantic Ocean, and even a cooling tendency is noted in the Northern Atlantic warming hole.

The SAT obtained from the U-Net's output, utilizing the same ensemble member as input, exhibits a pattern strikingly similar to that of the ensemble mean (compare centre and bottom panels). In both instances, the pattern is relatively uniform, albeit with heightened warming observed over land areas, coupled with an Arctic Amplification phenomenon. This suggests that the internal variability—such as the influence of ENSO events or the effects of prolonged weather patterns over continents—has been successfully eliminated. The regions displaying subdued warming or cooling tendencies are replicated, although the exact positioning and intensity might not precisely match those of the ensemble mean in certain areas, particularly the Southern and subpolar North Atlantic. It's worth noting a minor discontinuity at 180°E resulting from the padding process.

The performance of the method is quantified more systematically in the next section.

4 The U-Net as an internal variability filter

The U-Net was applied to every member of FGOALS-g3 and MPI-ESM. We then compare the results obtained with the respective ensemble mean of these two climate models.

Figures 5a and 5b illustrate the root mean squared error (RMSE) between the outcomes generated by the U-Net and the corresponding ensemble mean for the time period of 1905-2016. Notably, the discrepancies in U-Net's predictions are not uniformly distributed across space. The RMSE values fall within the range of 0.05°C to 0.5°C. The discrepancies generally remain below 0.2°C in tropical regions, except for instances over

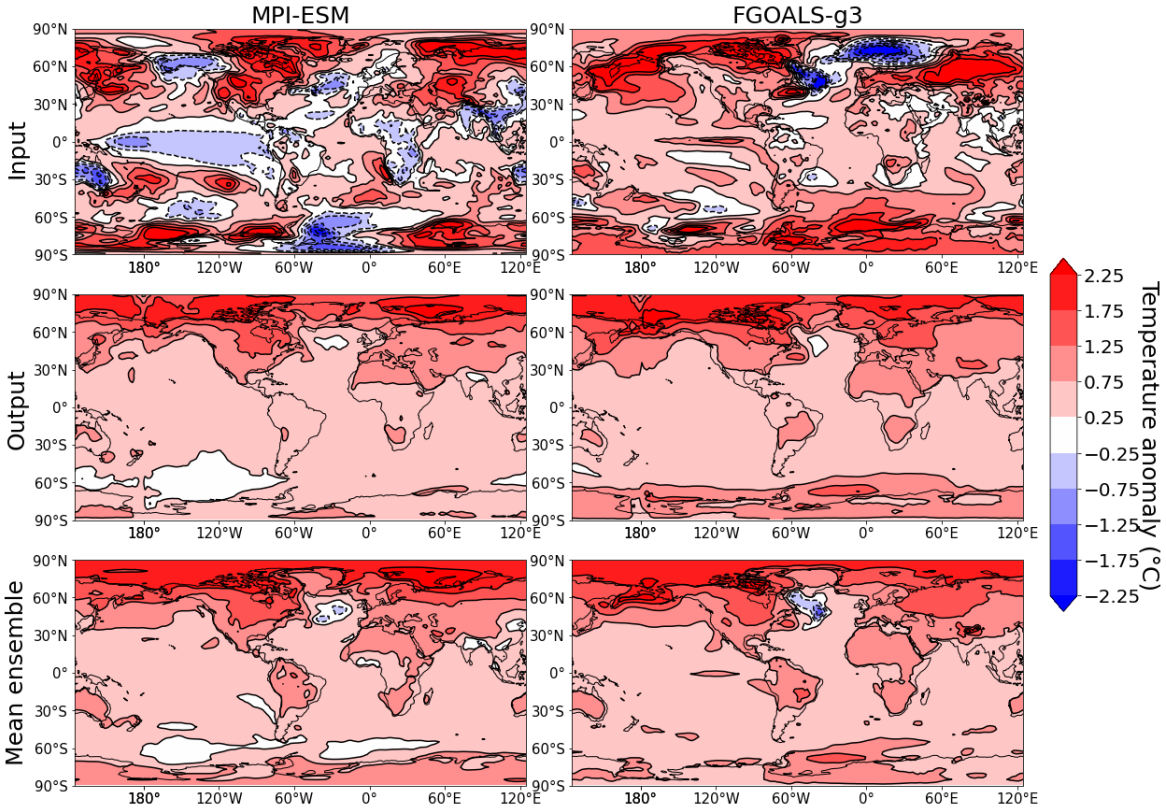

Figure 4. (First column) Anomalies of SAT in a randomly chosen member of MPI-ESM, the associated U-Net output and ensemble mean in 2016. (Second column) Same as the first column but for a randomly chosen ensemble member for FGOALS-g3.

Western Africa in the MPI-ESM model. In contrast, the largest errors are concentrated in polar areas, encompassing the Nordic Seas, Labrador Sea, and Bering Sea. Moreover, sizable errors are also evident over the Southern Ocean and the continents of the Northern Hemisphere situated above 45°N. These high-error regions correspond to locales characterized by substantial internal variability (refer to Figure 1). Nevertheless, it is noteworthy that the errors produced by the U-Net are approximately five times smaller than the actual internal variability. Between the years 1996 and 2016, both ensemble results exhibit a warming trend that is roughly 0.1°C lower in the U-Net results when compared to the ensemble mean (as observed in Figs. 5cd). This difference is indicated by the nearly consistent negative divergence situated between latitudes 45°N and 45°S.

The prevailing trend of systematic underestimation is, however, disrupted by an exception involving the subpolar Atlantic and the Southern Ocean, where an overestimation of warming is observed. This overestimation is particularly conspicuous in the FGOALS-g3 model, with warming anomalies extending to approximately 1°C over the Labrador Sea and 0.5°C over the Bering Sea. This divergence from the ensemble mean highlights the limited capacity of the neural network to accurately predict forced changes within the subpolar North Atlantic, which is a region that exhibits inconsistent surface temperature shifts across models (Swingedouw et al., 2021). The neural network's performance is restricted due to this discrepancy among models, which hampers its ability to discern the specific features of each climate model. For example, in the case of FGOALS-g3, the extensive anomalies in the Labrador and Bering Seas are not mirrored in the multi-model mean (see Figure 2). It's also plausible that the substantial internal variability observed in these regions poses a challenge for accurate removal by the neural network (refer to Figure 1). This underestimation extends to the continents, with a greater impact on South America, Africa, and Australia in the tropics, as well as North America and Northern Siberia in boreal regions. The degree of underestimation reaches 0.15°C for MPI-ESM and 0.13°C for FGOALS-g3 in these regions.

Figures 6c and 6d illustrate the temporal evolution of the global surface air temperature (GSAT) for both the MPI-ESM and FGOALS-g3 models, before and after applying the U-Net correction. The range of data variability is portrayed by a 90% confidence interval assuming an Gaussian distribution. The forced variability's temporal trend extracted via ensemble mean (depicted by the red line) is effectively captured by the U-Net outputs (represented by the blue line and blue shading).

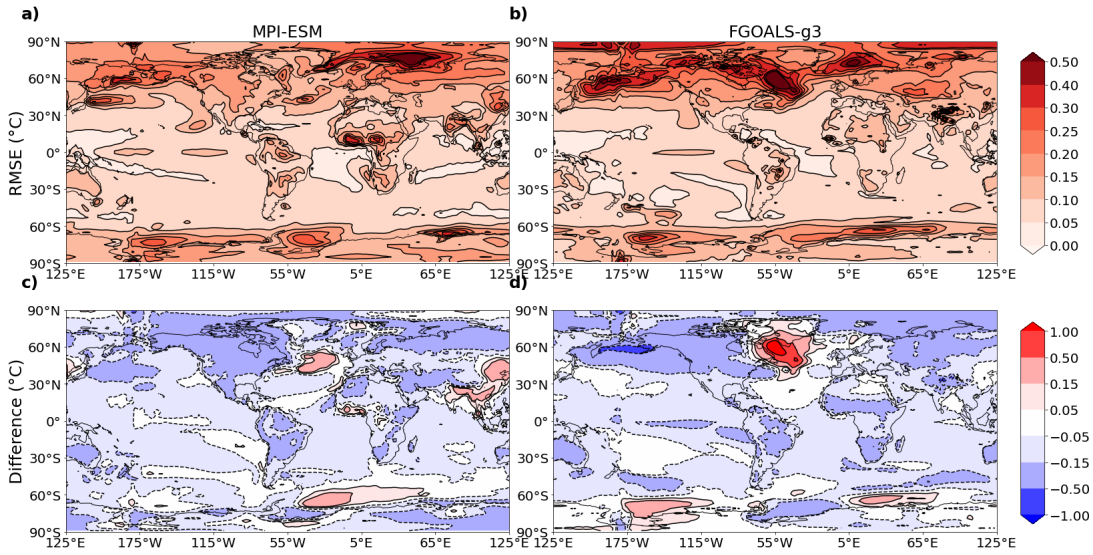

Figure 5. a) Root mean square difference of the surface air temperature, in C°, between the outputs of the U-Net and the mean ensemble in MPI-ESM, calculated across the members and all years in 1905-2016 b. b) Same as a) but for FGOALS-g3 c) Difference of the time mean SAT anomaly during 1996-2016, in °C, between the mean output of the U-Net and the corresponding ensemble mean, for MPI-ESM. d) Same as c) but for FGOALS-g3

From 1905 to 2016, a GSAT rise is observed, aligning with the anticipated shifts in radiative forcing (Gulev et al., 2021). Additionally, a cooling pattern emerges a few years subsequent to the significant volcanic eruptions of Agung (1963), El Chichón (1982), and Pinatubo (1991), a phenomenon accurately estimated by the U-Net. This outcome aligns with expectations based on climate models incorporating volcanic aerosol emissions. Impressively, the U-Net's outputs exhibit a marginal spread, reduced approximately tenfold, indicating a substantial removal of internal variability.

Nonetheless, the U-Net results exhibit anomalies with a slightly diminished amplitude compared to the ensemble mean. The spread of the U-Net outputs is also approximately twice as wide at the time series' beginning and end. The distribution of spatially averaged RMSE values within 90°S-90°N, comparing all U-Net outputs to the ensemble mean (depicted in Fig. 6a and 6b as blue histograms), reveals errors of around 0.12°C in MPI-ESM and 0.13°C in FGOALS-g3. Additionally, we examine the RMSE values when averaging within 60°N-90°N, as Fig. 5ab suggests that errors are most pronounced in this region (illustrated in Fig. 6ab as red histograms). Errors north of 60°N

are approximately twice as substantial as global averages, with an average error of around 0.23°C in MPI-ESM and 0.26°C in FGOALS-g3. In Fig. 6ef, the internal variability observed when averaging the SAT north of 60°N (as depicted by the red shading) is considerable in the raw model outputs (around 0.8°C). The ensemble mean SAT anomalies in this region increase from approximately -1°C in the early twentieth century to about 1.2°C in 2010. The temporal evolution of the SAT north of 60°N demonstrates notable similarity between the ensemble mean and the ensemble mean of U-Net outputs, with a roughly 10-fold reduction in spread. However, the amplitude of the anomalies is slightly underestimated, with a reduction of around 0.3°C in negative anomalies in the U-Net output between 1905 and 1930 in MPI-ESM. For FGOALS-g3, the SAT is underestimated by around 0.2°C during 1970-1990.

In Figure S2, the quadratic errors between the mean ensemble members and the U-Net output are presented for each year, with global (90°S-90°N) and north of 60°N averages considered for both MPI-ESM and FGOALS-g3. Notably, the RMSE exhibits elevated values during the initial and final years, characterized by peaks around the years 1975-1985 in both models. This pattern underscores the presence of substantial uncertainties at the data's onset and conclusion. When applying the 1900-2020 period for the output (without excluding the first and last four years), the errors actually surpass those portrayed in Figure S2, a fact that elucidates the rationale for excluding the endpoints in the ongoing analysis, as detailed in the methods (section 2). Moreover, the notable error peak during 1975-1985 lacks a definitive explanation, although it's plausible that this discrepancy could be linked to uncertainties associated with the implementation of aerosol forcings, notably CMIP5 for MPI-ESM and CMIP6 for FGOALS-g3.

The errors exhibited by the U-Net in relation to data from FGOALS-g3 are more prominent compared to those arising from the use of MPI-ESM data. This discrepancy can be attributed to the fact that MPI-ESM's simulated forced variability aligns more closely with the training data's characteristics, on average. Specifically, the training data's forced variability is in line with that of the MMM, and MPI-ESM demonstrates a smaller root mean squared difference from the MMM compared to FGOALS-g3 (as illustrated in Fig. 2).

To assess the reduction in internal variability achieved by the U-Net, we can quantitatively measure the number of ensemble members needed to surpass the U-Net's in-

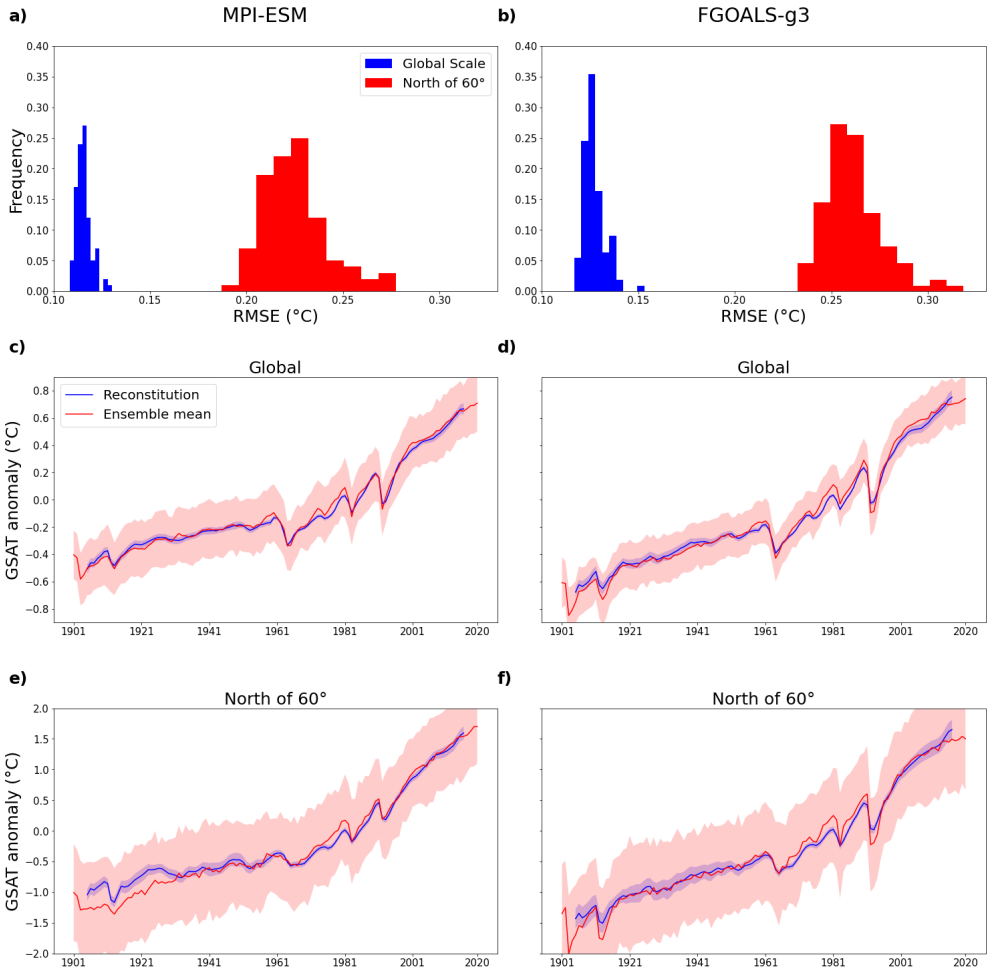

Figure 6. a) Histogram showing the distribution of the RMSE between the mean ensemble and the U-Net outputs of MPI-ESM. b) Same as a), but for FGOALS-g3. c) Time evolutions of the global mean surface air temperature, in °C, for the ensemble mean and the mean U-Net outputs for MPI-ESM. Color shade shows the spread of the time series, with 90% the ensemble members uncertainty assuming a gaussian distribution. d) Same as c) but for FGOALS-g3. e) and f) are the same as c) and d) but when averaging the SAT, in °C, north of 60°N.

dividual member results using a basic ensemble mean approach. This evaluation is conducted through a random subsampling process involving 500 sets of m members, where m varies from 1 to 40, for both the FGOALS-g3 and MPI-ESM ensembles. Within each subset, ensemble means are calculated. The RMSE between these subsample ensemble means and the actual ensemble mean obtained from all members is then determined (depicted by vertical red and blue lines in Figure 7). This RMSE computation is performed across all grid points and is spatially averaged. The 90% intervals, assuming an Gaussian distribution, of the 500 subsamples are also illustrated. This analysis is done for both the MPI-ESM and FGOALS-g3 ensembles across distinct geographical regions: global (90°S-90°N), North Atlantic (60°W-0°E, 0°N-60°N), North Pacific (120°E-100°W, 20°N-60°N), Niño3 (5°N-5°S, 150°W-90°W), as well as polar regions north of 60°N and south of 60°S. These chosen regions exhibit considerable forced and internal variability, as visually demonstrated in Fig. 1 and Fig. 2. Additionally, this evaluation is extended to encompass both oceanic and terrestrial areas in the 60°S-60°N band, allowing for a more comprehensive understanding of the U-Net's performance. The horizontal lines in the illustration correspond to the same RMSE values but for the U-Net output from each individual member. The accompanying color shade represents the spread of 90% uncertainty assuming an Gaussian distribution.

Figure 7a visually illustrates the progression of errors within the subset of members as the size of the subset increases. This pattern aligns with expectations, as a larger subset size leads to better estimations of forced variability and a corresponding reduction in residual internal variability by a factor of $\sqrt{n}$. The distribution of U-Net outputs mirrors the histograms presented in Figure 6, showing a high degree of similarity across both climate models. The U-Net effectively diminishes internal variability in GSAT by approximately a factor of slightly more than four, which is analogous to the residual variability observed within subsets containing around 17 members for FGOALS-g3 and 20 members for MPI-ESM. When focusing on regions spanning oceans and land between 60°N and 60°S, the outcomes remain largely consistent, showcasing a reduction in error magnitude by a factor of approximately four. This reduction corresponds closely to that achieved by using a subset of 15 to 20 members.

The U-Net's efficacy stands out prominently over the equatorial Pacific region, as depicted in panel 7f. This region is known for being heavily influenced by the ENSO, which dominates internal variability. The U-Net achieves a substantial reduction in variabil-

ity, amounting to a factor of 5.5. This reduction is akin to the outcome of utilizing an ensemble mean derived from around 30 members for both MPI-ESM and FGOALS-G3.

In other regions, the variability reduction is quite similar to that found globally. For instance, this consistency is observed in the North Pacific and polar regions, where the required number of members for equivalent outcomes remains relatively steady. However, in terms of removing internal variability, the U-Net showcases higher efficiency in the context of MPI-ESM for most scenarios. This pattern holds true except for the North Atlantic, where a notable deviation is observed: a set of 15 members is necessary in MPI-ESM to achieve results equivalent to the U-Net (∼4-fold reduction in residual variability), while merely 5 members suffice for FGOALS-g3 (halving of the residual variability).

The variation in performance between FGOALS-g3 and MPI-ESM might arise from dissimilarities in their internal variability, particularly over multi-decadal timescales, or due to differences in forced variability compared to the training data. Having completed this method evaluation, our focus now shifts to examining the outcomes when the U-Net is employed with observational data.

4.1 Filtering of the observations

The U-Net is now employed to process SAT observations derived from GISSTEMP. By utilizing observed data as input, the U-Net provides an estimate of the forced variability. In the interval from 1996 to 2016, the U-Net-derived forced SAT (depicted in Figure 8a) illustrates a fairly uniform warming, with amplified warming evident over the Arctic region, consistent with Arctic amplification. Furthermore, this warming effect is slightly more pronounced over land compared to oceans. Conversely, the Southern Ocean experiences less warming in comparison to other global regions. The spatial distribution of standard deviations (Figure 8b), computed from 1905 to 2016 using U-Net output, mirrors the anomalies observed in the 1996-2016 period. This agreement indicates the prevailing influence of increasing anthropogenic forcing. Notably, this pattern closely resembles the changes observed in the multi-model mean (MMM) (as depicted in Fig. 2). This underscore the significant contribution of the training dataset in determining the identified forced changes.

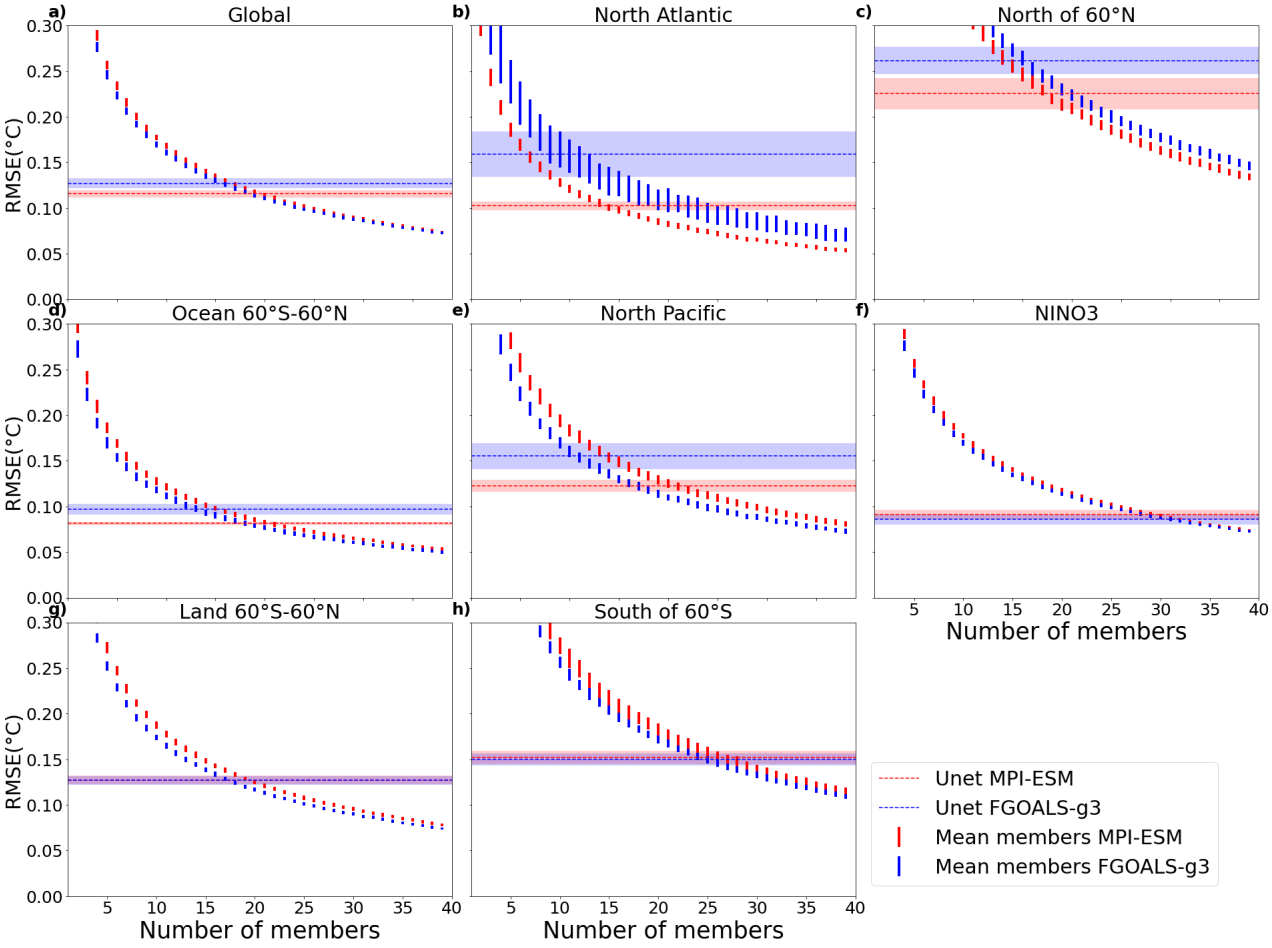

Figure 7. Spatial average of the RMSE for the forced variability estimated with the U-Net outputs obtained from each ensemble member, and the forced variability obtained with ensemble averages subsampling ensemble of size 1 to 40; for (red) MPI-ESM and (blue) FGOALS-g3. The RMSE calculated from the U-Net and each ensemble member is given by (color shade) the interval including 90% of the distribution, assuming a gaussian distribution, and (horizontal dashed line) the mean RMSE. The RMSE calculated from 500 subsample of size between 1 to 40 is illustrated with (vertical lines) the intervals including 90% of the ensemble member distribution, also assuming a gaussian distribution.

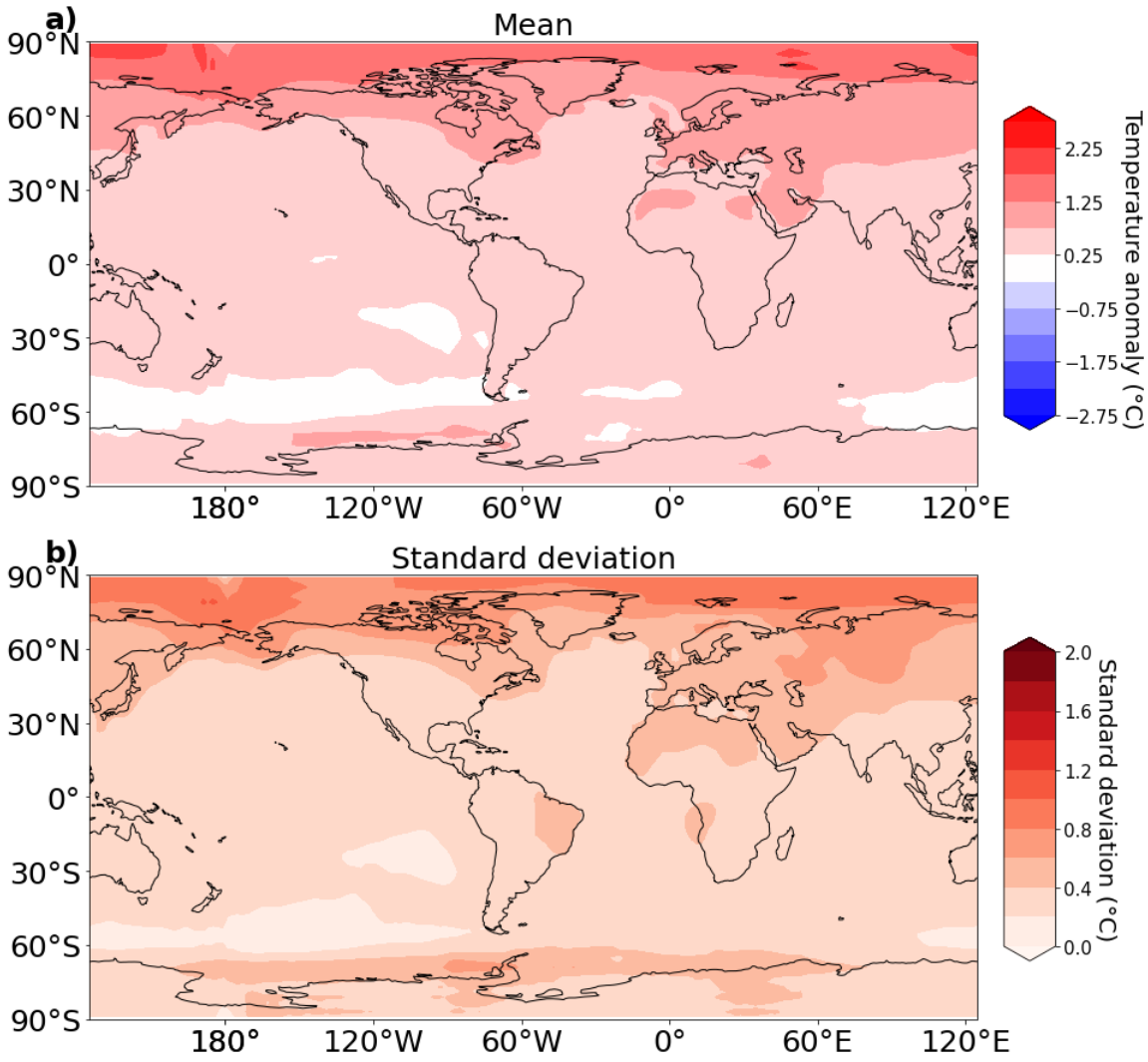

Figure 8. Forced surface air temperature (in °C) anomaly when applying the U-Net to GISTEMP observation : a) time average in 1996-2016; b) standard deviation in 1905-2016.

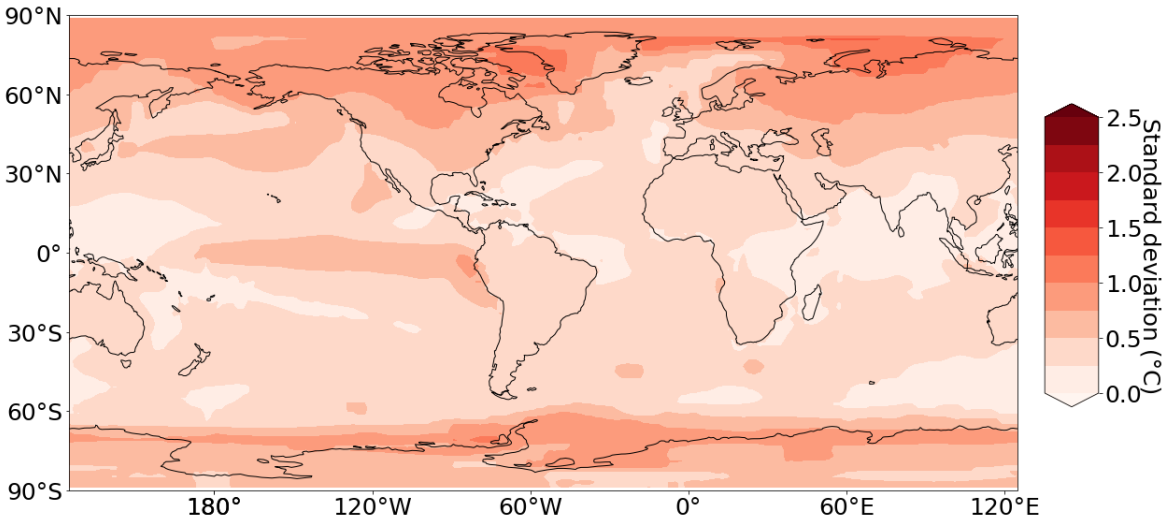

Figure 9. Standard deviation of the SAT deviations from the forced SAT, as estimated using the U-Net, in 1905-2016.

To quantify internal variability within the observations, we compute the deviations of observed SAT anomalies from the estimated forced changes. The resulting internal variability pattern, illustrated by the time standard deviation of these deviations shown in Figure 9, mirrors the model-derived pattern (Fig. 1). Higher internal variability values are observed over land areas, as well as regions near the boundaries of sea ice, such as the Labrador Sea and the Nordic Seas in the Northern Hemisphere, and the Southern Ocean. Notably, a local maximum of internal variability emerges in the equatorial Pacific, corresponding to the El Niño-Southern Oscillation region. This similarity in the spatial distribution of internal variability between observations and models underscores the consistency of our findings.

We now shift our focus to the GSAT and the Niño 3.4 region (5°N-5°S, 170°W-120°W), with a particular emphasis on Niño 3.4 due to its notably improved performance in our study. In the global context (Figure 10a), the forced variability reveals a consistent warming trend, which becomes more pronounced during the 1960s. Notably, the major volcanic eruptions of Agung (1963), El Chichón (1982), and Pinatubo (1991) are associated with temporary cooling patterns. By 2016, the GSAT anomaly reaches 0.7°C. As expected, the forced variability time series exhibits a significant reduction in inter-annual variability. This reduction is particularly striking within the Niño 3.4 region (Figure 10b), where variability at 2 to 7 years is almost entirely eliminated. The U-Net estimates the Niño

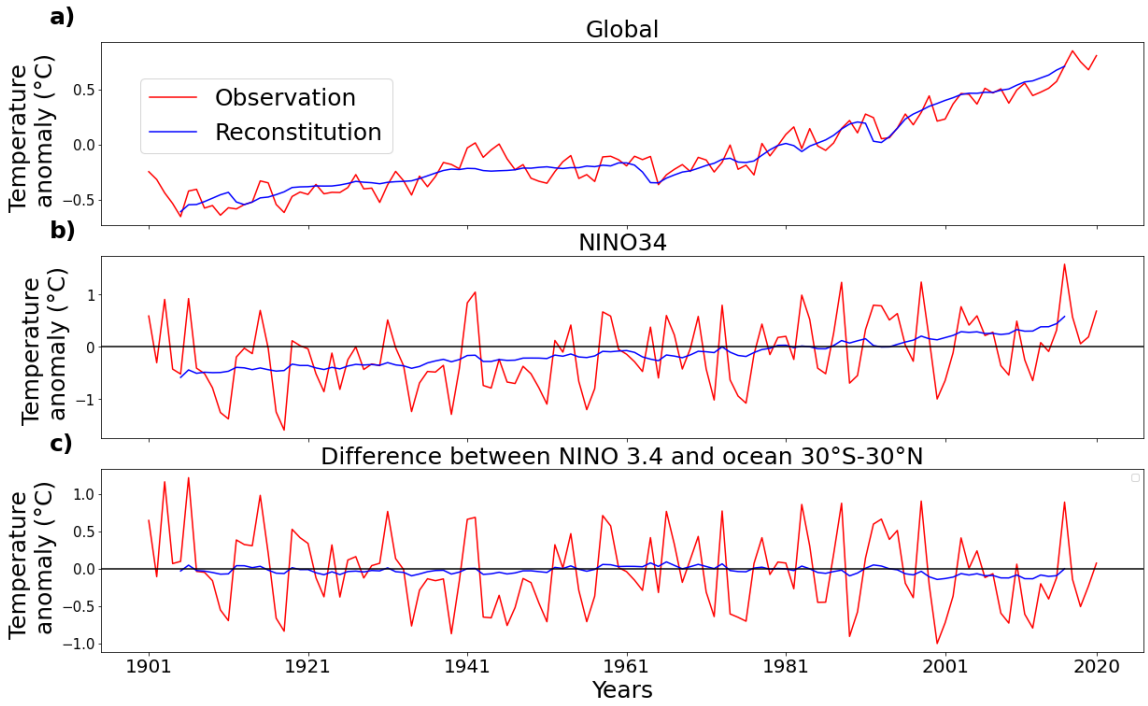

Figure 10. Time series of (red) the observed SAT anomaly and (blue) the forced SAT anomaly estimated by the U-Net for a) the global mean b) NINO 3.4 and c) the relative SAT, calculated as the difference between the averaged SAT in Niño 3.4 region and the tropical ocean SAT (30°S-30°N).

3.4 forced variability, depicting a steady warming trend. To quantify the changes of SAT in Niño 3.4 relative to the tropics, we calculate also the relative SAT, defined as the difference between the average SAT on the NINO 3.4 region and the average SAT on ocean grid between 30°S-30°N. The relative SST shows that the warming over the Niño 3.4 follows that of the tropics, so that no clear El Niño-like reponse is found, unlike climate models (Fig. 2). Some authors (Clement et al., 1996; Heede et al., 2020) have suggested that a forced cooling could exists in the relative SAT, called thermostat effect. Here the relative SAT shows a very small cooling (see Fig. 10c). In addition the SAT in the Niño 3.4 region are not affected by the forcing from the main volcanic eruptions. Therefore, no evidence of a Niño-like response to volcanic eruption (as in Khodri et al. (2017)) is found.

5 Conclusion

A novel approach is introduced in this study to effectively eliminate internal variability from a time-evolving two-dimensional dataset, specifically focusing on surface air temperature. The method employs a U-Net neural network and draws inspiration from the noise-to-noise technique. This framework treats internal variability as an analogous noise superimposed on the underlying forced variability. The U-Net model is trained using outputs from a diverse ensemble of climate models obtained from the CMIP simulations. Subsequently, this trained network is applied to observational data to unveil the forced variability signal by attenuating internal variability. The validation of this method involves utilizing large ensemble simulations from individual models, specifically the MPI-ESM and FGOALS-g3, to gauge its effectiveness. The forced variability derived from the ensemble mean is then contrasted with the outcomes from the U-Net application. To quantitatively assess the U-Net's efficacy in reducing internal variability, an "equivalent ensemble size" is computed. This metric indicates the ensemble size that would be required to achieve the same level of precision in capturing forced changes as the U-Net which is applied to a single member. The U-Net outputs for these two climate models' test data exhibit an error equivalent to an internal variability reduction of a factor of more than 4. This magnitude corresponds to the internal variability one could expect from an ensemble averaging 17 to 20 members. Furthermore, when the U-Net is applied to surface air temperature observations, the inferred forced changes align closely with the multi-model mean in terms of spatial patterns. The U-Net's results do not suggest an El Niño-like response to global warming. We observe that the U-Net encounters greater challenges in accurately estimating forced variability over the Arctic region. This discrepancy can be attributed to the significant forced and internal variability associated with changes in sea-ice extent in that area. Additionally, the U-Net's performance in capturing forced variability in the North Atlantic is less successful for the FGOALS-g3 model. This limitation might be linked to uncertainties stemming from the multi-decadal variability prevalent in these regions (Menary & Wood, 2018; Zhang, 2007).

In the pursuit of enhancing the U-Net methodology, several avenues for future improvements have been identified. One potential approach is to address the U-Net's sensitivity to the multi-model consensus of future variability by employing neural network regularization techniques, such as weights penalisation. Additionally, preprocessing methods like data augmentation could be explored to potentially mitigate such impacts. Im-

proving the evaluation process of the U-Net's performance is also on the horizon. This could involve testing the U-Net on a broader range of climate models to assess its generalizability. Comparing its outcomes with results from alternative methods, such as signal-to-noise filtering, could offer a comprehensive evaluation of the U-Net's effectiveness. To broaden the scope of application, the U-Net's performance might be further investigated using additional climate variables beyond surface air temperature (SAT). Variables such as sea level surface pressure and precipitation could be explored, capitalizing on potential correlations among these variables to provide more comprehensive insights. Lastly, the proposed method holds the potential for wider applications, including its deployment on simulations from projects like the Detection and Attribution Model Intercomparison Project (Gillett et al., 2016) or the Large Ensemble Single Forcing Model Intercomparaison Project (D. M. Smith et al., 2022). By leveraging transfer learning, the U-Net trained on historical simulations could be adapted to these datasets. This adaptation could facilitate the evaluation of specific forcing effects in individual climate models, offering a valuable tool for studying the impact of different external factors on the climate system. Such extensions of the method could contribute significantly to our understanding of climate attribution and variability.

Acknowledgments

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).

6 Open Research

Data Availability Statement

The CMIP5 and CMIP6 data is available through the Earth System Grid Federation 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://zenodo.org/record/8233743`.

References

Allen, M. R., & Stott, P. A. (2003). Estimating signal amplitudes in optimal fingerprinting, Part I: Theory. *Climate Dynamics*, *21*, 477–491.

Allen, M. R., & Tett, S. F. (1999). Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, *15*, 419–434.

Bonnet, R., Boucher, O., Deshayes, J., Gastineau, G., Hourdin, F., Mignot, J., ... Swingedouw, D. (2021). Presentation and evaluation of the ipsl-cm6a-lr ensemble of extended historical simulations. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002565.

Bonnet, R., Boucher, O., Vrac, M., & Jin, X. (2022). Sensitivity of bias adjustment methods to low-frequency internal climate variability over the reference period: an ideal model study. *Environmental Research: Climate*, *1*(1), 011001.

Bône, C. (2023). *Codes for "Separation of internal and forced variability of climate using a U-Net" [Software].* Retrieved from `https://zenodo.org/record/8233743`

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.-S., & Caltabiano, N. (2018). Decadal Climate Variability and Predictability: Challenges and Opportunities. *Bulletin of the American Meteorological Society*, *99*(3), 479 - 490. Retrieved from `https://journals.ametsoc.org/view/journals/bams/99/3/bams-d-16-0286.1.xml`

Chylek, P., Li, J., Dubey, M., Wang, M., & Lesins, G. (2011). Observed and model simulated 20th century Arctic temperature variability: Canadian earth system model CanESM2. *Atmospheric Chemistry and Physics Discussions*, *11*(8), 22893–22907.

Clement, A. C., Seager, R., Cane, M. A., & Zebiak, S. E. (1996). An ocean dynamical thermostat. *Journal of Climate*, *9*(9), 2190–2196.

Collier, M. A., Jeffrey, S. J., Rotstayn, L. D., Wong, K., Dravitzki, S., Moseneder, C., ... others (2011). The CSIRO-Mk3. 6.0 Atmosphere-Ocean GCM: participation in CMIP5 and data publication. In *International congress on modelling and simulation–modsim* (pp. 2691–2697).

Collins, W. D., Rasch, P. J., Boville, B. A., Hack, J. J., McCaa, J. R., Williamson, D. L., ... Zhang, M. (2006). The formulation and atmospheric simulation of the Community Atmosphere Model version 3 (CAM3). *Journal of Climate*, *19*(11), 2144–2161.

DelSole, T., Tippett, M. K., & Shukla, J. (2011). A significant component of unforced multidecadal variability in the recent acceleration of global warming. *Journal of Climate*, *24*(3), 909–926.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., ... others (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*(4), 277–286.

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: the role of internal variability. *Climate dynamics*, *38*, 527–546.

Deser, C., & Phillips, A. S. (2023). A range of outcomes: the combined effects of internal variability and anthropogenic forcing on regional climate trends over Europe. *Nonlinear Processes in Geophysics*, *30*(1), 63–84.

Deser, C., Phillips, A. S., Alexander, M. A., & Smoliak, B. V. (2014). Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *Journal of Climate*, *27*(6), 2271–2296.

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arneth, A., Arsouze, T., ... others (2021). The EC-earth3 Earth system model for the climate model intercomparison project 6. *Geoscientific Model Development Discussions*, *2021*, 1–90.

Egmont-Petersen, M., de Ridder, D., & Handels, H. (2002). Image processing with neural networks—a review. *Pattern recognition*, *35*(10), 2279–2301.

Enfield, D. B., & Cid-Serrano, L. (2010). Secular and multidecadal warmings in the North Atlantic and their relationships with major hurricane activity. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *30*(2), 174–184.

England, M. H., McGregor, S., Spence, P., Meehl, G. A., Timmermann, A., Cai, W., ... Santoso, A. (2014). Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nature climate change*, *4*(3), 222–227.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958.

Eyring, V., Gillett, N., Rao, K. A., Barimalala, R., Parrillo, M. B., Bellouin, N., ... Zho, B. (2021). Human Influence on the Climate System. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Pres*.

Frankcombe, L. M., England, M. H., Mann, M. E., & Steinman, B. A. (2015). Separating internal variability from the externally forced climate response. *Journal of Climate*, *28*(20), 8184–8202.

Frankignoul, C., Gastineau, G., & Kwon, Y.-O. (2017). Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic multidecadal oscillation and the Pacific decadal oscillation. *Journal of Climate*, *30*(24), 9871–9895.

Fyfe, J. C., Kharin, V. V., Santer, B. D., Cole, J. N., & Gillett, N. P. (2021). Significant impact of forcing uncertainty in a large ensemble of climate model simulations. *Proceedings of the National Academy of Sciences*, *118*(23), e2016549118.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., ... others (2011). The community climate system model version 4. *Journal of climate*, *24*(19), 4973–4991.

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., ... Tebaldi, C. (2016). The detection and attribution model intercomparison project (DAMIP v1. 0) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3685–3697.

Gulev, S. K., Thorne, P. W., Ahn, J., Dentener, F. J., Domingues, C. M., Gerland, S., ... others (2021). Changing state of the climate system.

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, *48*(4).

Harzallah, A., & Sadourny, R. (1995). Internal versus SST-forced atmospheric variability as simulated by an atmospheric general circulation model. *Journal of*

Climate, *8*(3), 474–495.

Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, *6*(10), 1957–1971.

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108.

He, C., Clement, A. C., Cane, M. A., Murphy, L. N., Klavans, J. M., & Fenske, T. M. (2022). A North Atlantic warming hole without ocean circulation. *Geophysical research letters*, *49*(19), e2022GL100420.

Heede, U. K., Fedorov, A. V., & Burls, N. J. (2020). Time scales and mechanisms for the tropical Pacific response to global warming: A tug of war between the ocean thermostat and weaker Walker. *Journal of Climate*, *33*(14), 6101–6118.

Ilesanmi, A. E., & Ilesanmi, T. O. (2021). Methods for image denoising using convolutional neural network: a review. *Complex & Intelligent Systems*, *7*(5), 2179–2198.

Jeffrey, S., Rotstayn, L., Collier, M., Dravitzki, S., Hamalainen, C., Moeseneder, C., ... Syktus, J. (2013). Australia's CMIP5 submission usingthe CSIRO-Mk3. 6 model. *Australian Meteorological and Oceanographic Journal*, *63*(1), 1–13.

Jiang, W., Gastineau, G., & Codron, F. (2021). Multicentennial variability driven by salinity exchanges between the Atlantic and the Arctic Ocean in a coupled climate model. *Journal of Advances in Modeling Earth Systems*, *13*(3), e2020MS002366.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... others (2015). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349.

Keil, P., Mauritsen, T., Jungclaus, J., Hedemann, C., Olonscheck, D., & Ghosh, R. (2020). Multiple drivers of the North Atlantic warming hole. *Nature Climate Change*, *10*(7), 667–671.

Khodri, M., Izumo, T., Vialard, J., Janicot, S., Cassou, C., Lengaigne, M., ... others (2017). Tropical explosive volcanic eruptions can trigger El Niño by cooling tropical Africa. *Nature communications*, *8*(1), 778.

Kosaka, Y., & Xie, S.-P. (2013). Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature*, *501*(7467), 403–407.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., & Aila, T. (2018). Noise2Noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*.

Lenssen, N. J., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., & Zyss, D. (2019). Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6307–6326.

Li, Yu, Y., Tang, Y., Lin, P., Xie, J., Song, M., ... others (2020). The flexible global ocean-atmosphere-land system model grid-point version 3 (FGOALS-g3): description and evaluation. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002012.

Li, S., & Huang, P. (2022). An exponential-interval sampling method for evaluating equilibrium climate sensitivity via reducing internal variability noise. *Geoscience Letters*, *9*(1), 1–10.

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., ... others (2019). The Max Planck Institute Grand Ensemble: enabling the exploration of climate system variability. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2050–2069.

Marini, C., & Frankignoul, C. (2014). An attempt to deconstruct the Atlantic multidecadal oscillation. *Climate dynamics*, *43*, 607–625.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., ... others (2021). Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, *2*.

Meehl, G. A., Hu, A., Arblaster, J. M., Fasullo, J., & Trenberth, K. E. (2013). Externally forced and internally generated decadal climate variability associated with the Interdecadal Pacific Oscillation. *Journal of Climate*, *26*(18), 7298–7310.

Menary, M. B., Robson, J., Allan, R. P., Booth, B. B., Cassou, C., Gastineau, G., ... others (2020). Aerosol-forced AMOC changes in CMIP6 historical simulations. *Geophysical Research Letters*, *47*(14), e2020GL088166.

Menary, M. B., & Wood, R. A. (2018). An anatomy of the projected north atlantic

warming hole in cmip5 models. *Climate Dynamics*, *50*(7-8), 3063–3080.

Neelin, J. D., Battisti, D. S., Hirst, A. C., Jin, F.-F., Wakata, Y., Yamagata, T., & Zebiak, S. E. (1998). ENSO theory. *Journal of Geophysical Research: Oceans*, *103*(C7), 14261–14290.

Newman, M., Alexander, M. A., Ault, T. R., Cobb, K. M., Deser, C., Di Lorenzo, E., ... others (2016). The Pacific decadal oscillation, revisited. *Journal of Climate*, *29*(12), 4399–4427.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Parker, D., Folland, C., Scaife, A., Knight, J., Colman, A., Baines, P., & Dong, B. (2007). Decadal to multidecadal variability and the climate change background. *Journal of Geophysical Research: Atmospheres*, *112*(D18).

Parsons, L. A., Brennan, M. K., Wills, R. C., & Proistosescu, C. (2020). Magnitudes and spatial patterns of interdecadal temperature variability in CMIP6. *Geophysical Research Letters*, *47*(7), e2019GL086588.

Rodgers, K. B., Lin, J., & Frölicher, T. L. (2015). Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences*, *12*(11), 3301–3320.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–miccai 2015: 18th international conference, munich, germany, october 5-9, 2015, proceedings, part iii 18* (pp. 234–241).

Schmidt, A., Mills, M. J., Ghan, S., Gregory, J. M., Allan, R. P., Andrews, T., ... others (2018). Volcanic radiative forcing from 1979 to 2015. *Journal of Geophysical Research: Atmospheres*, *123*(22), 12491–12508.

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., ... others (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678.

Smith, C. J., & Forster, P. M. (2021). Suppressed late-20th century warming in CMIP6 models explained by forcing and feedbacks. *Geophysical Research Letters*, *48*(19), e2021GL094948.

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., ...

others (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16), 9591–9618.

Smith, D. M., Gillett, N. P., Simpson, I. R., Athanasiadis, P. J., Baehr, J., Bethke, I., ... others (2022). Attribution of multi-annual to decadal changes in the climate system: The Large Ensemble Single Forcing Model Intercomparison Project (LESFMIP). *Frontiers in Climate*, *4*.

Solomon, A., Goddard, L., Kumar, A., Carton, J., Deser, C., Fukumori, I., ... others (2011). Distinguishing the roles of natural and anthropogenically forced decadal climate variability: implications for prediction. *Bulletin of the American Meteorological Society*, *92*(2), 141–156.

Steinman, B. A., Mann, M. E., & Miller, S. K. (2015). Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures. *Science*, *347*(6225), 988–991.

Sun, L., Alexander, M., & Deser, C. (2018). Evolution of the global coupled climate response to Arctic sea ice loss during 1990–2090 and its contribution to climate change. *Journal of Climate*, *31*(19), 7823–7843.

Swart, N. C., Fyfe, J. C., Hawkins, E., Kay, J. E., & Jahn, A. (2015). Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change*, *5*(2), 86–89.

Swingedouw, D., Bily, A., Esquerdo, C., Borchert, L. F., Sgubin, G., Mignot, J., & Menary, M. (2021). On the risk of abrupt changes in the north atlantic subpolar gyre in cmip6 models. *Annals of the New York Academy of Sciences*, *1504*(1), 187–201.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American meteorological Society*, *93*(4), 485–498.

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., ... others (2020). Climate model projections from the scenario model intercomparison project (ScenarioMIP) of CMIP6. *Earth System Dynamics Discussions*, *2020*, 1–50.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., & Lin, C.-W. (2020). Deep learning on image denoising: An overview. *Neural Networks*, *131*, 251–275.

Ting, M., Kushnir, Y., Seager, R., & Li, C. (2009). Forced and internal twentieth-

century SST trends in the North Atlantic. *J. Climate*, *22*, 1469–1481.

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., ... others (2011). The representative concentration pathways: an overview. *Climatic change*, *109*, 5–31.

Vincent, L., Zhang, X., Brown, R., Feng, Y., Mekis, E., Milewska, E., ... Wang, X. (2015). Observed trends in Canada's climate and influence of low-frequency variability modes. *Journal of Climate*, *28*(11), 4545–4560.

Wang, C., & Picaut, J. (2004). Understanding ENSO physics—A review. *Earth's Climate: The Ocean–Atmosphere Interaction, Geophys. Monogr*, *147*, 21–48.

Wills, R. C., Battisti, D. S., Armour, K. C., Schneider, T., & Deser, C. (2020). Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations. *Journal of Climate*, *33*(20), 8693–8719.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, *9*, 611–629.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., ... Taylor, K. E. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782.

Zelinka, M. D., Zhou, C., & Klein, S. A. (2016). Insights from a refined decomposition of cloud feedbacks. *Geophysical Research Letters*, *43*(17), 9259–9269.

Zhang, R. (2007). Anticorrelated multidecadal variations between surface and subsurface tropical north atlantic. *Geophysical Research Letters*, *34*(12).

Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y.-O., Marsh, R., Yeager, S. G., ... Little, C. M. (2019). A review of the role of the Atlantic meridional overturning circulation in Atlantic multidecadal variability and associated climate impacts. *Reviews of Geophysics*, *57*(2), 316–375.

12

Supporting Information for "Separation of internal and forced variability of climate using a U-Net"

Constantin Bône[12], Guillaume Gastineau[1], Sylvie Thiria[1], Patrick Gallinari[23]and Carlos Mejia[1]

[1]UMR LOCEAN, IPSL, Sorbonne Université, IRD, CNRS, MNHN

[2]UMR ISIR, Sorbonne Université, CNRS, INSERM

[3]Criteo AI Lab

Contents of this file

1. Figures S1 to S2

2. Tables S1 to S3

Table 1. List of climate CMIP5 climate model used. Nb indicates the ensemble size of each simulation.

Table 1.

Name	Nb Historical	Nb RCP8.5
GISS-E2-R-p1	2	2
MPI-ESM-LR	3	3
CanESM2	5	5
CESM1-CAM5	3	3
FIO-ESM	3	3
CNRM-CM5	10	5
CSIRO-Mk3-6-0	10	10
FGOALS-g3-s2	3	3
GISS-E2-H-p1	6	3
GISS-E2-H-p3	2	2
HadGEM2-ES	3	3
IPSL-CM5A-LR	6	4
GISS-E2-R-p3	3	2

Table 2. List of climate CMIP6 climate model used. Nb indicates the ensemble size of each simulation.

Name	Nb Historical	Nb SSP2-4.5
ACCESS-CM2	3	3
CanESM5	25	25
CESM2	11	3
CanESM5-CanOE	3	3
GISS-E2-1-G	27	27
EC-Earth3	16	16
MIROC-ES2L	23	23
HadGEM3-GC31-LL	5	4
GFDL-ESM4	3	3
FIO-ESM-2-0	3	3
KACE-1-0-G	3	3
GISS-E2-1-G-p3	9	4
ACCESS-ESM1-5	40	30
CAS-ESM2-0	4	2
NESM3	5	2
MPI-ESM1-2-HR	10	2
NorESM2-LM	3	3
GISS-E2-1-G-p5	9	9
IPSL-CM6A-LR	33	11
GISS-E2-1-H	14	5
CESM2-WACCM	3	3
CNRM-CM6-1	30	10
CAMS-CSM1-0	3	2
UKESM1-0-LL	19	17
MPI-ESM1-2-LR	10	10
MRI-ESM2-0	9	9
CNRM-ESM2-1	10	6
FGOALS-f3	6	4
CanESM5	40	25
MIROC6	50	50

Table 3. List of climate SMILE climate model used. Nb indicates the ensemble size of each simulation.

Name of the model	Nb of historical members	Nb scenario members	Origin of forcings	Scenario
CSIRO-Mk3-6-0	29	29	CMIP5	RCP8.5
EC-EARTH	15	15	CMIP6	SSP2-4.5
MPI-ESM	100	100	CMIP6	SSP2-4.5
FGOALS-g3	110	110	CMIP6	SSP2-4.5

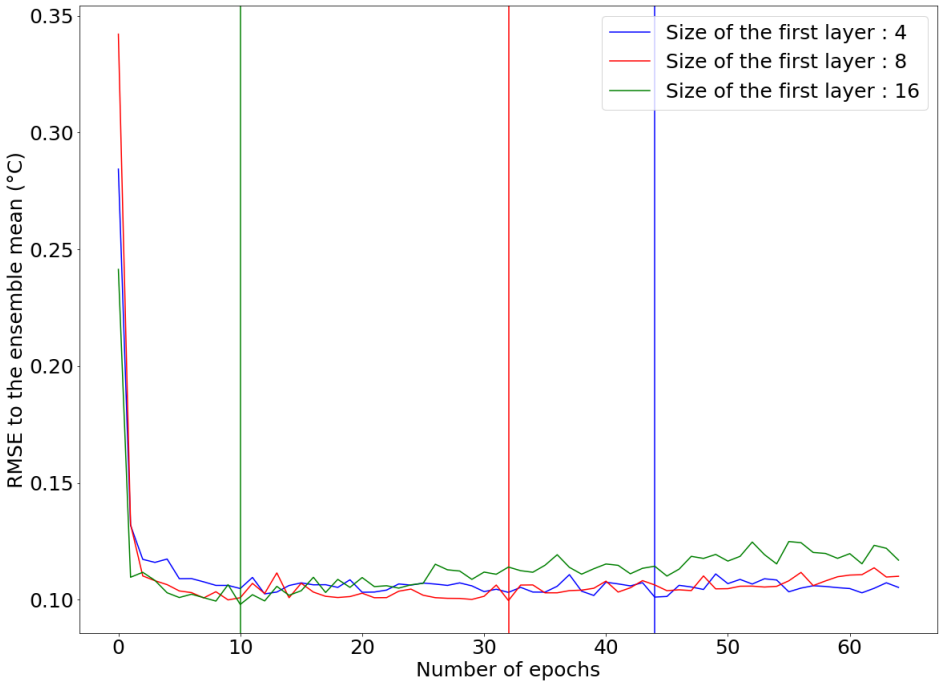

Figure S1. Validation RMSE (in °C) using the ensemble mean of MIROC6 outputs as a target, and each member as inputs for different epochs and when varying the numbers of filters for each convolutionnal layer of the U-Net. Vertical line of the same colour shows the epoch where the minimum RMSE is obtained for the three changes in the number of filters.

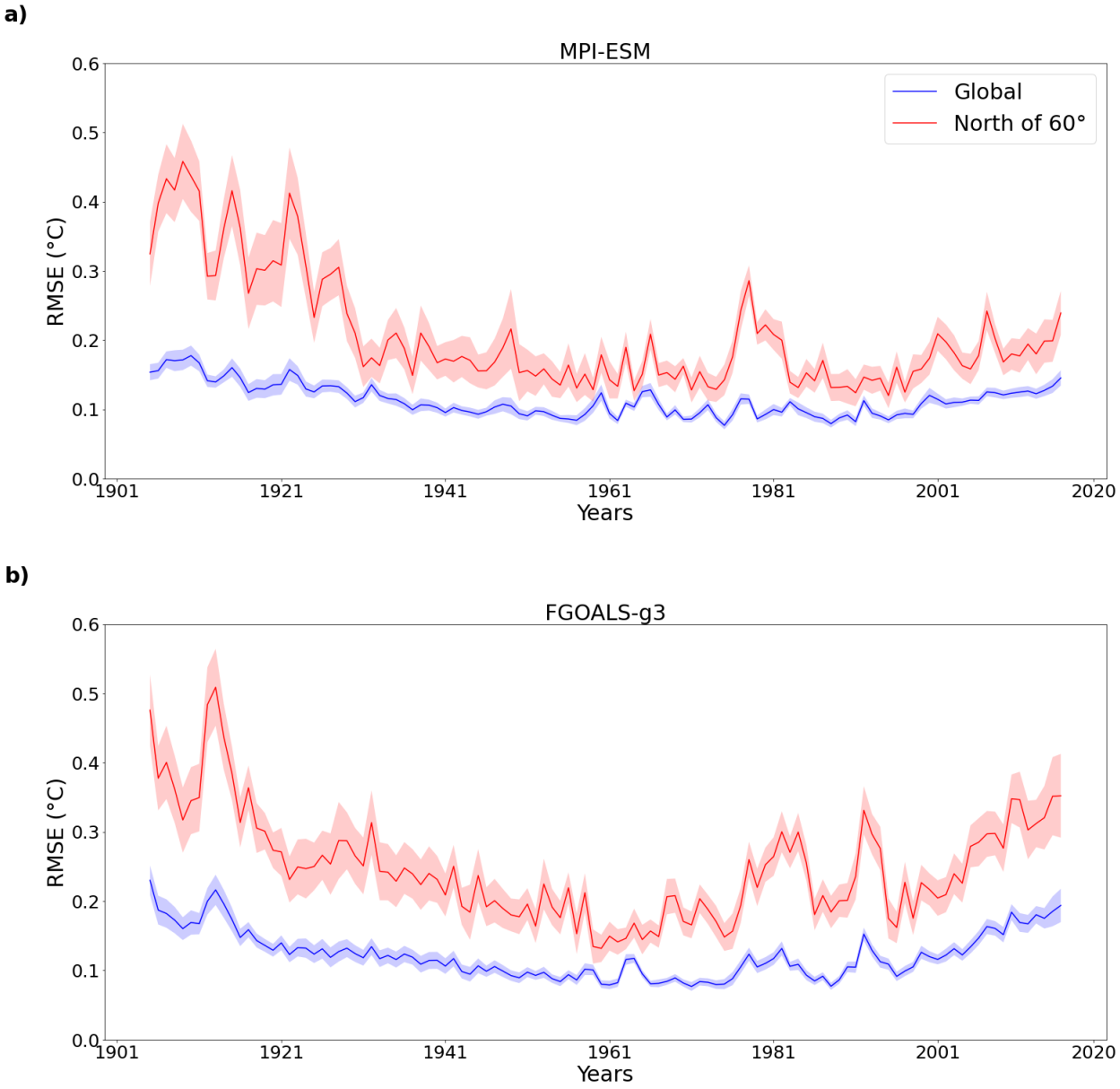

Figure S2. a) Spatial average of the RMSE, in °C, between the U-Net output obtained from each member of MPI-ESM and the ensemble average (blue) over 90°S-90°N and (red) over 60°N-90°N. The line provides the ensemble mean error obtained with an average from the errors of all U-Net outputs. Colour shade shows one standard deviation among the error of the outputs from all members. b) Same as a) but for FGOALS-g3 members.

Chapitre 4

Attribution de la température globale de l'air en surface

4.1 Introduction

Une grande partie de la thèse fut de créer une nouvelle méthode pour procéder à la détection et attribution du changement climatique qui prenne en compte les non-additivités possibles dans le comportement des forçages. Pour cela, une nouvelle méthode reposant sur un CNN et une méthode issue de l'IA explicable, l'optimisation inverse, a été développée. Lors de la première année de thèse, cette nouvelle méthode a été présentée lors d'une conférence nommée "*Climate Informatics*". Cette conférence donnait accès à une publication automatique dans la revue *Environmental Data Sciences*. Ce travail fut poursuivi dans un second manuscrit accepté dans la revue *Journal of Advances in Modelling Earth Systems*.

4.2 Article de *Environmental Data Sciences*

La méthode d'optimisation inverse est une méthodologie issue de l'IA explicable, mais dont le fonctionnement général est semblable à des méthodologies utilisées sur des modèles météorologiques (Brajard et al., 2012). Dans ce

dernier contexte, la méthode est appelée *variationnal inversion*. Cette méthode fut également citée par Toms et al., 2020 comme une des méthodes prometteuses et physiquement interprétables issues du *machine learning* pour les géosciences. Selon Toms et al., 2020 une méthode interprétable est une méthode se focalisant sur la manière dont un réseau de neurone a appris plutôt que juste la sortie de celui-ci. Cette méthode consiste à retrouver l'entrée (noté x) d'un réseau de neurones entraîné qui correspondrait à un résultat de sortie donné (noté y). Cela est fait par une méthode de descente de gradient en minimisant une fonction de coût mesurant la différence entre y et $CNN(x)$. La méthodologie d'optimisation inverse décrite par Toms et al., 2020 est illustré dans la Figure 4.1.

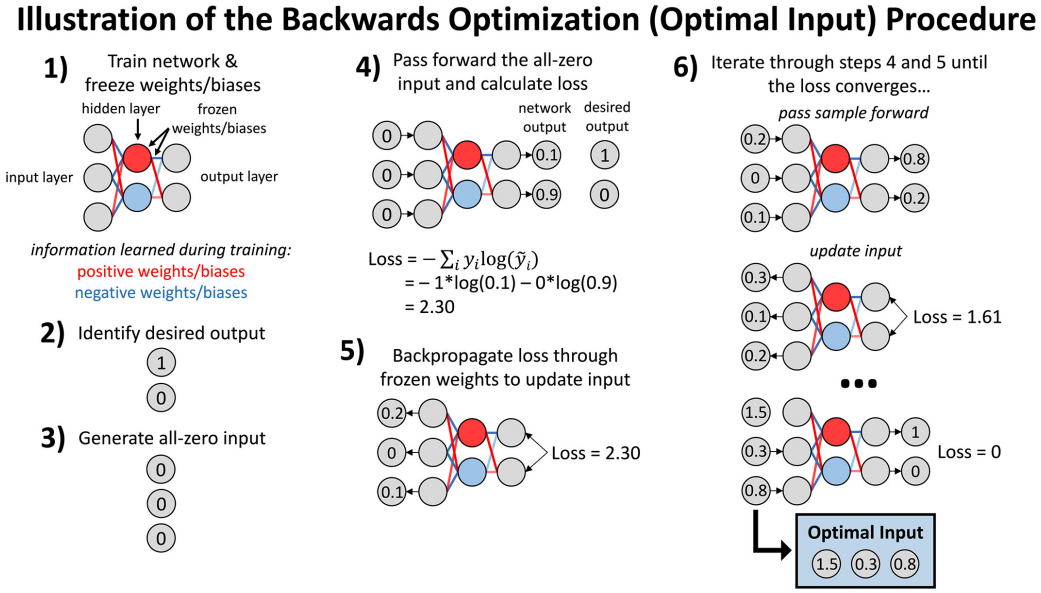

FIGURE 4.1: Méthodologie d'optimisation inverse. Tiré de Toms et al., 2020

Cette méthode d'attribution par optimisation inverse (appelé *"variationnal inversion"* dans l'article) n'utilise pas, contrairement au travail sur la séparation de la variabilité interne et forcée, des données tridimensionnelles (temps, longitude et latitude) mais monodimensionnelle (temps uniquement). Nous utilisons en effet uniquement la GSAT car nous souhaitions efficacement comparer l'optimisation inverse aux autres méthodologies d'attribution

existantes. En effet, l'étude de l'attribution de la GSAT est a été très étudié, en particulier par la méthode des empreintes optimisées.

Nous utilisons un réseau de neurones CNN (voir Chapitre 2) entraîné avec des sorties de simulations historiques et DAMIP de douze modèles climatiques issues de CMIP6. Nous n'utilisons pas de modèles climatiques issues de CMIP5 car pas assez de simulations correspondant à celles de DAMIP ont été réalisées. Le CNN vise à reproduire les changements de température globale de surface dû à l'ensemble des forçages avec comme entrée les changements de température dû à l'effet individuel des forçages. Les forçages utilisés sont les gaz à effet de serre, les aérosols anthropiques et les forçages naturels. Contrairement au chapitre 3 nous n'utilisons pas un U-Net, mais un CNN, car la dimension et le volume de données est bien plus faible. Le CNN utilisé n'est fait que de trois couches et a un relativement faible nombre de poids et de biais.

Une fois le réseau entraîné et ses poids et biais fixés, une méthode d'optimisation inverse est utilisée. Cette méthodologie fut toutefois modifiée afin de mieux correspondre à notre problème. Contrairement à l'étape 3 de la méthodologie d'optimisation inverse « classique » (voir Fig. 4.1) nous n'utilisons pas un point de départ constant fixé à 0, mais un ensemble de points de départs constitués de triplets de simulations DAMIP venant de tous les modèles climatiques utilisés. En effet, la méthode d'optimisation inverse est sensible au point de départ choisi. Cette façon de faire permet de donner au réseau un « a priori » ayant une cohérence physique. Cette possibilité était déjà discutée dans Toms et al., 2020. La fonction de coût est également altérée pour rajouter un terme dit de « rappel » qui pénalise une solution s'éloignant trop du point de départ. Cette modification est faite afin que la solution trouvée ne s'éloigne pas trop d'une solution « physiquement cohérente ». Nous utilisons cette méthode d'optimisation inverse modifié sur les observations de changement de température globale sur la période 1900-2014 et trouvons des

estimations de l'impact individuel de chaque forçage.

Nous validons cette méthode avec une approche en modèle parfait. Cela consiste à retirer alternativement chaque modèle climatique utilisé dans l'étude et d'entraîner le CNN en leurs absences. On utilise alors un membre historique du modèle retiré comme pseudo-observations. Les résultats obtenus sont alors comparés à la moyenne d'ensemble des simulations DAMIP du modèle retiré.

Avec cette méthode de modèle parfait, nous trouvons pour sept des douze modèles climatiques utilisés un bon accord entre les résultats de la *variationnal inversion* et des simulations DAMIP. Cette méthode obtient également des résultats physiquement cohérents sur les observations. En l'effet attribué pour les gaz à effet de serre est un réchauffement prononcé estimé dans un intervalle de confiance à 90 % à [0,8 °C ; 1,9 °C] en 2014. L'effet des aérosols anthropique est lui un refroidissement moins marqué ([-0,7 °C ; -0,1 °C] en 2014) et celui des forçages naturels n'est que peu marqué ([-0,1 °C ; 0,3 °C] en 2014) excepté dans les années suivantes des éruptions volcaniques comme celle de Pinatubo en 1991. En effet, en 1993, l'effet des forçages naturels est de [-0,5 °C,°0,1 °C], ce qui est cohérent avec le refroidissement attendu lors de l'émission d'aérosols naturels. Les résultats obtenus sont similaires à ceux obtenus dans d'autres études de détection et attribution comme illustrés dans la Figure 2.13 avec cependant des intervalles de confiance plus élevés.

Cette nouvelle méthode d'attribution dite par « optimisation inverse » diffère d'autres méthodes d'attribution par son utilisation du *machine learning* et par sa prise en compte des effets de non-additivité présents dans les forçages. Ces non-additivités ne sont que peu présentes dans les GSAT (Marvel et al., 2015; Shiogama et al., 2013) mais sont bien plus présentes sur d'autres variables physiques comme les précipitations (Marvel et al., 2015), ou à des échelles régionales (Good et al., 2015) comme l'Arctique (Deng et al., 2020)

ou dans l'hémisphère Sud (Pope et al., 2020). Cependant, cette étude souffrait de limites très nettes. Le réseau de neurones utilisé et la méthodologie d'« optimisation inverse » ont des hyper-paramètres dont l'impact ne sont pas proprement étudiés. La méthodologie d'optimisation inverse pouvait également être améliorée afin de mieux tenir compte des spécificités du problème d'attribution. Par exemple, l'impact estimé pour chaque forçage montrait une variabilité haute fréquence très élevée alors que l'effet des aérosols ou des gaz à effet de serre ne montrent que des variations multidécennales ou plus longues (Gulev et al., 2021). Enfin, la méthode d'optimisation inverse n'est pas dans ce travail comparé rigoureusement à d'autres méthodes d'attribution. Son principal avantage de ne pas faire l'hypothèse d'additivité des forçages n'est également pas démontré. Ce travail fut donc poursuivi dans une étude plus longue qui fit également l'objet d'un manuscrit accepté dans le *Journal of Advances in Modelling Earth Systems*.

Environmental Data Science (2022), 1: e27, 1–9
doi:10.1017/eds.2022.17

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Detection and attribution of climate change: A deep learning and variational approach

Constantin Bône[1,2,*], Guillaume Gastineau[1], Sylvie Thiria[1] and Patrick Gallinari[2,3]

[1]UMR LOCEAN, Sorbonne Université, IRD, CNRS, MNHN, Paris, France
[2]UMR ISIR, Sorbonne Université, Paris, France
[3]Criteo AI Lab, Paris, France
*Corresponding author. E-mail: constantin.bone@sorbonne-universite.fr

Received: 03 October 2022; **Accepted:** 25 October 2022

Keywords: Climate change; climate models; convolutional neural network; detection and attribution; variational inversion

Abstract

Twelve climate models and observations are used to attribute the global mean surface temperature (GMST) changes from 1900 to 2014 to external climate forcings. The external forcings are decomposed into the effects of the well-mixed greenhouse gas concentration variation, the effects of anthropogenic aerosol concentration changes, and the effects of natural forcings. First, a convolutional neural network (CNN) is trained to estimate the simulated historical GMST from single-forcing experiments using outputs from the multi-model ensemble. We then use this CNN to solve the attribution problem using an original variational inversion approach. The variational inversion is first validated using historical climate simulations as pseudo-observations. Then we perform an inversion from observations. This provides a distribution of the GMST resulting from the three forcings. For 2014, inversions estimate that the greenhouse gases changes are responsible for a GMST anomaly within [0.8°C,1.9°C], while anthropogenic aerosols and natural forcings anomalies are within [−0.7°C,−0.1°C] and [−0.1°C,0.3°C], respectively. The method designed here can be adapted and extended to attribute the changes of other variables or to focus on the regional scale.

Impact Statement

To devise efficient adaptation policies, it is key to understand the causes of past climate changes. Here, we present a method based on neural networks to estimate the past global mean surface temperature (GMST) anomalies caused by the changes in the greenhouse gas concentration, the variation of anthropogenic aerosols, and the variation driven by naturally occurring phenomena. This method is based on the training of a convolutional neural network using the estimations from 12 state-of-the-art climate models. Then we infer the most likely causes for the observed GMST changes from 1900 to 2014. The methodology presented could be applied in future studies to other variables or at the regional scale.

1. Introduction

Detection and attribution of climate change are key to understanding past climate change and devising adaptation policies. Detection aims to prove the existence of climate change exceeding its internal

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

variability. Internal variability refers to climate variations resulting from processes intrinsic to the climate system. For instance, the global mean surface temperature (GMST) varies by a few tenths of degrees during the phases of the El Niño Southern Oscillation. Similarly, the Atlantic multidecadal variability can also influence the global climate. Boundary conditions of the climate system, known as forcings, can also cause climate change. The dominant forcings in the historical period (i.e., from 1850 to present day) are the increase in the greenhouse gases atmospheric concentration, the variations of the atmospheric aerosol concentration, the variations of solar insolation, the changes in land use, and stratospheric ozone concentrations. Anthropologically driven and naturally occurring forcings are generally considered separately. Attribution then aims to explain and quantify the impacts of the different forcings in the detected change, using both observations and climate models (Stott et al., 2010).

Hasselmann (1993) defined a method to estimate the fingerprints of forced climate change based on the analysis of observation and the climate models. In the reference methods, such fingerprints are used to characterize the climate. The observations are linearly regressed onto the simulated responses of the external forcings using these fingerprints (Ribes et al., 2013). It is often assumed that the impacts of the forcings are additive. Detection and attribution studies often consider reduced dimensional data, using global spatial and temporal means. By doing so, they estimate a pseudo-invertible covariance matrix used for the linear regression (Zhang et al., 2007). These methods have shown that global warming cannot be explained only by internal variability, and it was extremely likely that human activities had caused at least more than half of the observed increase in GMST from 1951 to 2010 (Gulev et al., 2021).

State-of-the-art attribution methods have several limitations such as the additivity assumption of the influence of forcings, and attribute anomalies to more than two forcings are often difficult (Gillett et al., 2021). We aim at exploring an alternative framework based on non-linear predictors to account for more complex interactions between the forcings. We then consider neural network regressors that have shown their ability to exploit spatial and temporal data structures, find patterns, and fuse heterogeneous sources of information efficiently in different domains of earth system sciences (Reichstein et al., 2019). We use convolutional neural networks (CNNs) and a variational inversion method, to perform climate change attribution. This accounts for the non-additivity in the forcings and is used to quantify the uncertainties in the attributed changes. Section 2 presents the data and the methodology. Section 3 is devoted to the results and Section 4 to the conclusions.

2. Data and Methods

2.1. Data

We use monthly air temperature from 1850 to 2014, from climate model simulations and observations. We use the outputs of the CMIP6 (Coupled Model Intercomparison Project 6; Eyring et al., 2016) simulations performed with 12 ocean–atmosphere general circulation models (see Table 1 for details). We denote HIST as the historical simulations, using as varying boundary conditions all the external forcings. These forcings include the estimations of greenhouse gases, aerosols, ozone concentration, and the estimated past variation of solar activity and land use. Each climate model provides several simulation instances called members generated through a macro-perturbation of the initial conditions. We also use single-forcing simulations from the DAMIP (Gillett et al., 2016) panel of CMIP6. These simulations used as varying boundary conditions only one of the external forcings, all the other external forcings being fixed at their value from 1850. We use the single-forcing simulations hist-aer denoted AER, hist-nat denoted NAT, and hist-GHG denoted GHG. They respectively use as varying forcing the anthropogenic aerosols, natural forcing (i.e., volcanic aerosol and solar variations), and greenhouse gases concentration. The effects of stratospheric ozone and land use were not investigated.

We also use observations (denoted OBS) of the 2 m air temperature over the continent from HadCRUT4 (Morice et al., 2012) blended with sea surface temperature from HadISST4 (Rayner et al., 2003). To make HIST and OBS comparable, we correct OBS of its blending effects using a 1.06 multiplier

https://doi.org/10.1017/eds.2022.17 Published online by Cambridge University Press

Table 1. Model and simulation used in this study.

Model	GHG	AER	NAT	HIST
CESM2	3	3	2	11
IPSL-CM6A-LR	10	10	10	32
ACCESS-ESM1-5	3	3	3	30
BCC-CSM2-MR	3	3	3	3
CanESM5	50	50	30	65
CNRM-CM6-1	9	10	10	10
FGOALS-g3	3	3	3	6
HadGEM3	4	4	4	5
MIROC6	3	3	3	50
MRI-ESM2.0	5	5	5	7
NorESM2-LM	3	3	3	3
GISS-E2-1-G	10	12	20	19

Note. The numbers in the columns GHG, AER, and NAT provide the number of members used.

coefficient (Richardson et al., 2018). The missing values have been filled by kriging (Cowtan and Way, 2014).

All monthly data are converted into an annual mean and averaged spatially from 90°S to 90°N. We then estimate the temperature anomalies in the 1900–2014 period (115 years). In each simulation (HIST, GHG, NAT, AER) we compute the mean temperature during the 1850–1900 period and remove it from the temperature. For the GISS-E2-1-G, we compute the temperature anomalies separately for the simulations using two different physics, with different schemes to calculate the aerosols indirect impact (Kelley et al., 2020). The same procedure is applied to OBS. Then, the time series of 115 years are normalized: for each model, we compute the maximum value of the ensemble mean of HIST and divide all the members of all simulation by this maximum. We also divide OBS by its maximum value.

2.2. Methodology

First, we determine the relationship linking the GMST of HIST to that of GHG, AER, and NAT. We train a CNN using the time series of AER, GHG, and NAT as input, with a size of (3,115), and HIST as the target, with a size of (1,115). The resulting CNN estimates the GMST anomaly in the historical period from GHG, AER, and NAT GMST time series.

The CNN consists of 3 one-dimensional convolutional layers. The kernel size for all layers is 11, the input is zero-padded by 5 pixels, and the length of the layers is 10 in the CNN. Hyperbolic tangent is used as an activation function in the hidden layers to add non-linearities. The training phase is made of three steps using the mean square error (MSE): (a) we randomly select a climate model, (b) we randomly select an instance of each simulation (GHG, AER, NAT, and HIST), and (c) we train the network using the corresponding (GHG, AER, NAT) and HIST time series as input and target with a batch size of 100. We iterate this process 5×10^6 times in total separated into 100 epochs of 5×10^4 iterations. The procedure ensures that each model is used equally when training the CNN. The architecture and hyperparameters are chosen using a k-fold cross-validation technique. We considered the 12 models separately, leaving out the data of one climate model. We then train a CNN using the remaining 11 models and use the data from the excluded model as the validation set. The process is iterated by removing each model alternately and the mean validation error is estimated. The selected architecture provides the lowest mean validation

https://doi.org/10.1017/eds.2022.17 Published online by Cambridge University Press

error after varying the number of layers, the kernels sizes, and the lengths of the layers. The CNN is finally trained using the 12 models together.

2.2.1. Variational inversion

The detection attribution problem aims at estimating the input time series (GHG, AER, and NAT) corresponding to the OBS time series. Therefore, we use a variational approach and the trained CNNs estimating the contribution of each forcing in the multi-model dataset. In geophysics, the variational inversion (Diouf et al., 2011; Brajard et al., 2012) considers a physical phenomenon and an associated model M. The variational inversion seeks to infer the physical parameters that led to the observations, according to the geophysical model. It often implies the use of the adjoint model of M which estimates changes in the input in response to a disturbance of the output values calculated by M. The basic idea is to determine the minimum of a cost function J that measures the disagreements between the observations and the model estimations. Due to the complexity of the model, the desired minimum is classically obtained by using gradient methods, to estimate the control parameters. In our case, we use the CNN as a model M so this process is straightforward and the inversion is obtained using the classical back-propagation algorithm. In our case, the parameters we are looking for are the input of the CNN with OBS at output. The inversion is an "ill-posed" problem that has multiple solutions and needs the estimation of the parameter distribution; moreover, the method is sensitive to the initialization. To overcome these (Bauer et al., 2020) problems, we repeat the process using different starting points. We use as a cost function the MSE and add a penalization term. This penalization is needed to keep a physically coherent solution. The cost function is

$$J(X) = MSE(OBS, CNN(X)) + B \times MSE(X, Xst) \tag{1}$$

where B is a scaling factor, set to 0.01, Xst the initial value of the inputs, and X input to be determined. This minimization is iterated until $MSE(OBS, CNN(X))$ is less than 0.05°C/°C. To choose the initial value Xst, we used multiple physically consistent values. For each of the 12 climate models, we randomly select 100 triplets of members of GHG, AER, and NAT for Xst and generate 1,200 variational inversions.

3. Results

3.1. Neural network performance

We evaluate the ability of the proposed approach to estimate the GMST from the HIST simulations using the k-fold cross-validation. For each climate model, we use a CNN trained with the data excluding that model using the 11 other climate models. Table 2 presents the training (validation) root mean square error (RMSE) in the first column (second column) when the CNN has seen the outputs (or not seen the outputs) from the climate model. All the RMSE is provided here for the normalized time series so that the unit is °C per °C. We obtain a mean training RMSE of 0.15°C/°C and a validation RMSE of 0.17 °C/°C across the climate models. The training RMSE is only slightly lower than the validation RMSE so that the CNN avoids overfitting. The RMSE varies among the models, with values from 0.09°C/°C (0.1°C/°C) in CanESM5 to 0.20°C/°C (0.24°C/°C) in NorESM1-LM for the training (validation) RMSE. The reason for these differences remains to be fully investigated, but we suggest that the output of the CNN reflects the similarities among models, and a low performance reflects a singularity of the GMST simulated by one model. We verify that the validation performance of a simple baseline linear network consisting of only a linear layer has a larger mean validation error of 0.21 °C/°C so that non-linearities do improve the performance.

3.2. Variational inversion

To validate the variational inversion, we used the members of HIST as pseudo-observations and the k-fold framework. For each HIST members for each climate models, we produce the variational inversions with 100 randomly chosen starting points using the CNN trained from data excluding that climate model. Then,

https://doi.org/10.1017/eds.2022.17 Published online by Cambridge University Press

Table 2. (First column) Training RMSE (°C/°C) computed on the outputs of the climate model when that model is seen by the CNN; (Second column) Validation RMSE (°C/°C) computed on the outputs of the climate model when the model is not seen by the CNN.

Model	Train CNN	Validation CNN	GHG Inversion	AER inversion	NAT inversion
CESM2	0.11	0.12	0.13	0.16	0.09
IPSL-CM6A-LR	0.13	0.14	0.23	0.22	0.08
ACCESS-ESM1-5	0.13	0.13	0.14	0.16	0.07
BCC-CSM2-MR	0.17	0.18	0.26	0.19	0.08
CanESM5	0.09	0.1	0.11	0.07	0.05
CNRM-CM6-1	0.18	0.19	0.11	0.11	0.07
FGOALS-g3	0.09	0.11	0.31	0.36	0.12
HadGEM3	0.14	0.19	0.34	0.26	0.06
MIROC6	0.19	0.2	0.14	0.13	0.11
MRI-ESM2.0	0.2	0.21	0.13	0.1	0.09
NorESM2-LM	0.19	0.24	0.15	0.24	0.12
GISS-E2-1-G	0.19	0.2	0.12	0.13	0.1

Note. (Last three columns) Mean RMSE (°C) between the mean inversion of the effects of greenhouse gases (anthropogenic aerosols and natural forcings) and the GMST from the ensemble mean of GHG (AER and NAT).

the inversions are denormalized. As an illustration, one HIST instance was chosen randomly and is shown in Figure 1, black line. The mean inversion with the forcing of greenhouse gases, anthropogenic aerosols, and natural forcings (Figure 1, red, blue, and green lines) is then compared to the ensemble mean GMST of GHG, AER, and NAT (Figure 1, purple, dark blue and beige lines in Figure 1). The color shades in Figure 1 quantify the spread among the inversions found from the different starting points, with one standard deviation. The spread of the simulations GHG, AER, and NAT illustrates one standard deviation across the available ensemble members. The greenhouse gases influence is variable among models, as the inversion results simulate in 2014 a GMST varying from 1°C to 2°C. This reflects the different sensitivity of climate models. The result of the anthropogenic aerosols inversion also scale with the climate sensitivity, with a large cooling for sensitive models. Nevertheless, FGOALs-g3, IPSL-CM6A-LR, or BCC-CSM2-MR simulate weak anomalies for anthropogenic aerosols compared to AER but large anomalies for greenhouse gases. Conversely, HadGEM3 or NorESM2-LM simulates rather large anomalies for anthropogenic aerosols and for HadGEM3 rather weak anomalies for greenhouse gases. The inversions provide realistic values for the natural forcings that agree with the outputs from NAT, with time series with small anomalies, except for the cooling in 1963, 1982, and 1991, following the major eruptions of Agung, El Chichon, and Pinatubo. For all models, the inversions have a large spread, much larger than the spread of the simulations GHG, AER, or NAT. It reflects the diversity of starting points used in the inversions. In 7 out of 12 models, we found a good agreement between the inversions and the simulated anomalies forcings as the spread of inversions agrees with the mean simulated anomalies. However, for FGOALs-g3, BCC-CSM2-MR, HadGEM3, NorESM2-LM, IPSL-CM6A-LR, the inversion is biased.

The mean RMSEs between the mean inversions results of each HIST instance and the ensemble mean of the corresponding single-forcing simulation are presented in Table 2. RMSEs for BCC-CSM2-MR, IPSL-CM6A-LR, HadGEM3, and FGOALS-g3 RMSEs are large for the influence of greenhouse gases with, respectively, 0.26, 0.23, 0.34, and 0.31 °C. Similarly, the influence of anthropogenic aerosols is not well retrieved for FGOALS-g3 (RMSE of 0.36 °C), HadGEM3 (0.26 °C) and NorESM2-LM (0.24°C), and IPSL-CM6A-LR (0.22°C). This confirms the analysis of Figure 1, where these models all simulate contrasted GMST in the results of anthropogenic aerosols and greenhouse gases inversion results.

https://doi.org/10.1017/eds.2022.17 Published online by Cambridge University Press

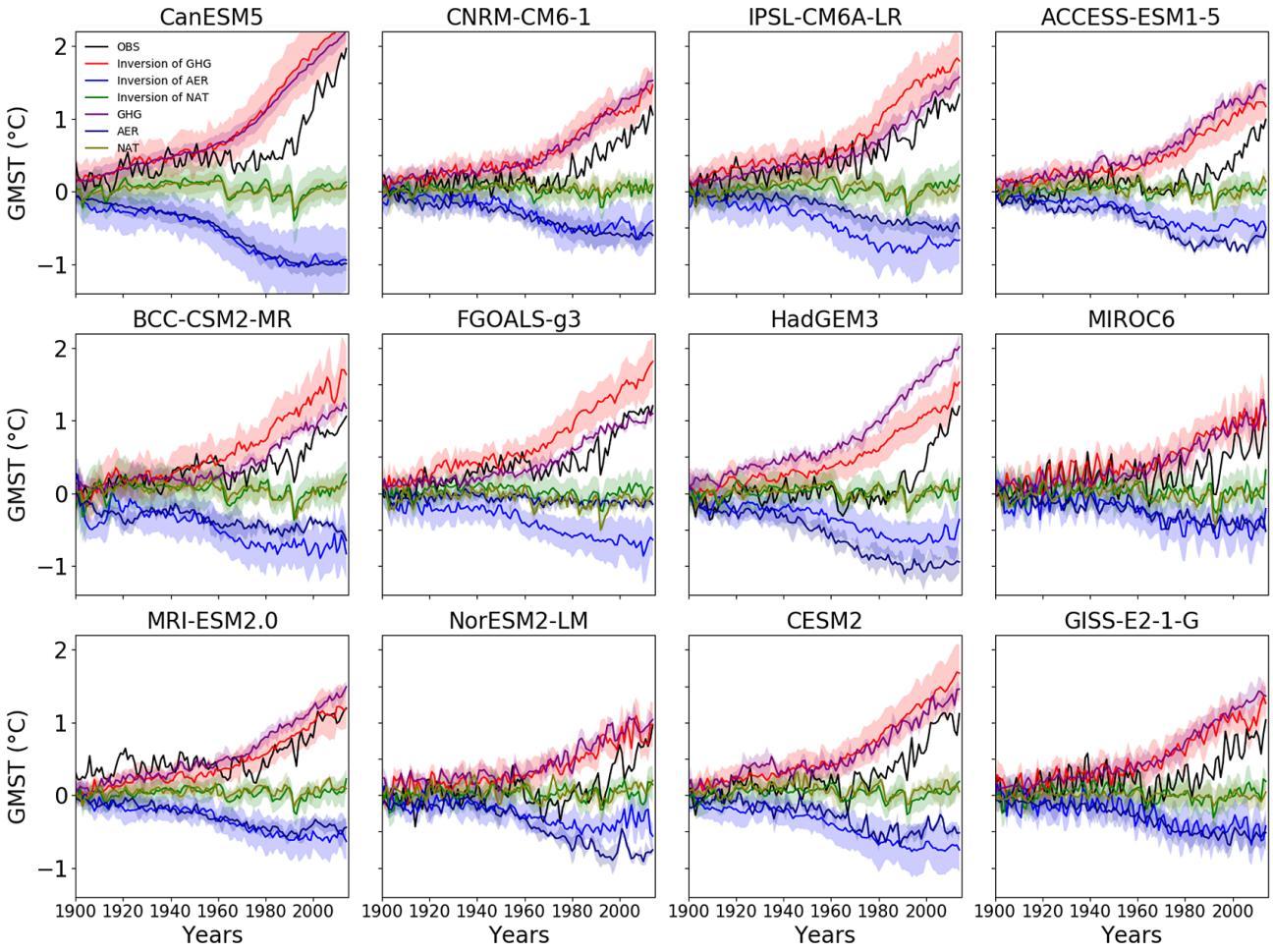

Figure 1. *GMST for an HIST member (black) randomly chosen as pseudo-observation, the mean results of variational inversions from the same member for the (red) greenhouse gases, (blue) anthropogenic aerosols and (green) natural forcings effects and ensemble mean of the (purple) GHG, (dark blue) AER and (beige) NAT. The color shades show one standard deviation across the inversion or across the ensemble members.*

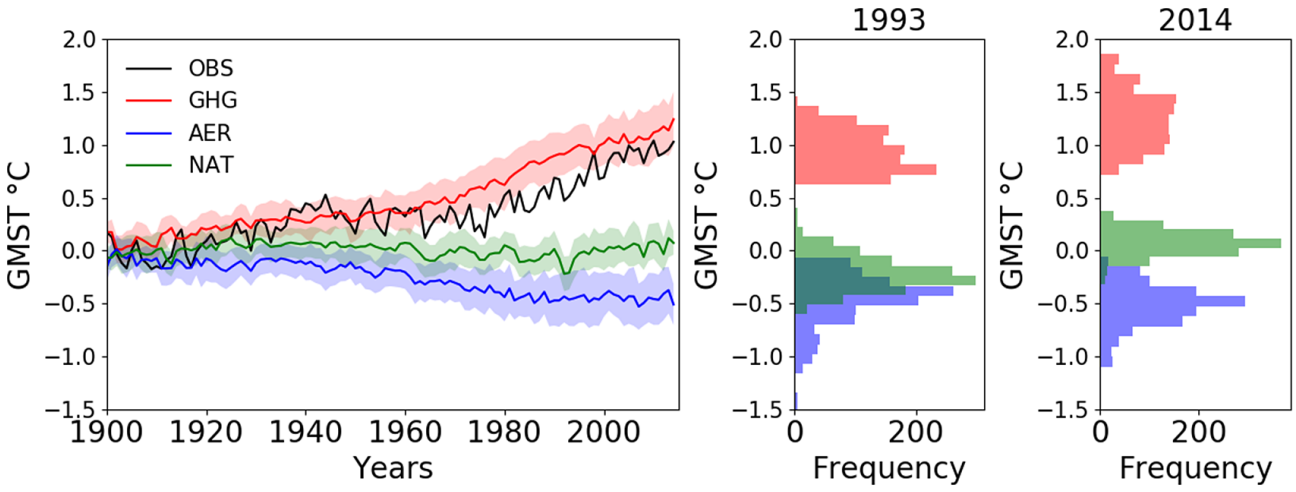

Figure 2. *Left: (black) Observation in °C and variational inversion for (red) greenhouse gases, (blue) anthropogenic aerosols, and (green) natural forcings. Shades show the standard deviation across the 1,200 varational inversion. (Right): Histogram from the inversion for (red) greenhouse gases,(blue) anthropogenic aerosols, and (green) natural forcings for the year 1993 and 2014.*

https://doi.org/10.1017/eds.2022.17 Published online by Cambridge University Press

The validation RMSEs reflect the distance of the climate models from the multi-model average. This distance results from the different climate sensitivity in each model, as well as the implementation of anthropogenic aerosol forcing, which varies between models (Pincus et al., 2016).

To attribute the observed changes, we apply the variational inversion to the observed GMST. The starting points provide 1,200 results of the effects of the greenhouse gases (Figure 2, red line), anthropogenic aerosol (Figure 2, blue line), and natural forcing (Figure 2, green line). The spread is evaluated using the standard deviation across the inversions (shades in Figure 2, left panel). The greenhouse gases and anthropogenic aerosols influence is smaller compared to many of the models illustrated in Figure 1, with variations consistent with that simulated in GHG and AER. The effect of natural forcings remains small compared to the results of Figure 1, with only a small decrease in the GMST in 1992 and 1993 following the Pinatubo eruption. The right panel Figure 2 illustrates the distribution of the results from the inversions in 1993 and 2014. We study in particular the year 1993 following the 1991 Pinatubo eruption and 2014 as it is the last year of the time series. The distribution shows a large spread for the effects of forcings associated with the diversity of the starting points used. When using the 95 percent intervals from the distribution of the inversion, the results show a range of [0.8°C,1.9°C] for greenhouse gases, [−0.7°C,-0.1°C] for anthropogenic aerosols, and [−0.1°C,0.3°C] for natural forcings in 2014. In 1993, the effect of natural forcing is with a 95 percent intervals of [−0.5°C,0.1°C] consistent with well-known effects of volcanic aerosols following the Pinatubo eruption (Gulev et al., 2021). We can compare these results (Figure 2, right) to the results found in Gillett et al. (2021) using the same data but with the regularized optimal fingerprinting method. They found anomalies of [1.2°C,1.9°C] for greenhouse gases, [−0.7°C,-0.1°C] for anthropogenic aerosols, and [0.01°C,0.06°C] for natural forcings in the 2010–2019 decade. This suggests that the variational inversion provides coherent results values but with larger confidence intervals.

4. Conclusion

In this article, we proposed an original solution to the attribution problem that does not rely on the classical forcing additivity assumption. The estimation relies on a non-linear forcing combination model that is learned from climate models simulations using GMST as an example. Spatial information of data will be included in future works. We chose however to not use it in this study to compare our results to previous studies, using mostly the GMST. The results found are coherent with a previous study using the same dataset and a classic fingerprinting method (Gillett et al., 2021) but with a larger uncertainty due in part to the different starting points of the inversion. For 4 of the 12 climate models, the validation score is lower. Other choices of architectures could be tested to combine more realistically the effects of forcings like recurrent neural networks more adapted to time series. Alternative variational inversion frameworks like other cost functions or starting points could also be tested. This new method could be applied in case a large non-additivity is expected in other variables (precipitation for example) or at the regional scale (Lehner and Coats, 2021).

Abbreviations

AER	hist-aer simulations
CMIP6	Coupled Model Intercomparison Project 6
CNN	convolutionnal neural network
GHG	hist-GHG simulations
GMST	global mean surface temperature
GSAT	global surface air temperature
HIST	historical simulations
MSE	mean square error
NAT	hist-nat simulations
OBS	observations
RMSE	root mean square error

https://doi.org/10.1017/eds.2022.17 Published online by Cambridge University Press

Acknowledgments. We are grateful for the technical assistance of Carlos Mejia, Alexandre Bône, and Michel Crépon. This work was performed using HPC ressources from GENCI-IDRIS (grant 2021-AD011013295).

Author Contributions. Conceptualization, methodology, and visualization: G.G, S.T., and C.B,; Software and formal analysis: C.B.; Writing: G.G., S.T., C.B., and P.G; Resources: G.G.; Investigation: C.B. and G.G; all authors approved the final submitted draft.

Competing Interests. The authors declare no competing interests exist.

Data Availability Statement. Data and codes can be found in https://gitlab.com/ConstantinBone/detection-and-attribution-of-climate-change-a-deep-learning-and-variational-approach.

Ethics Statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding Statement. This research was supported by grants from the Sorbonne Center for Artificial Intelligence (SCAI). Part of this work has been supported by project ChairesIA 2019—DL4CLIM ANR-19-CHIA-0018-01. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Provenance. This article is part of the Climate Informatics 2022 proceedings and was accepted in *Environmental Data Science* on the basis of the Climate Informatics peer review process.

References

Bauer S., **Tsigaridis K.**, **Faluvegi G.**, **Kelley M.**, **Lo K.**, **Miller R.**, **Nazarenko L.**, **Schmidt G.**, **Wu and J.** (2020) Historical (1850–2014) aerosol evolution and role on climate forcing using the giss modele2.1 contribution to cmip6. *Journal of Advances in Modeling Earth Systems 12*, e2019MS001978.

Brajard J., **Santer R.**, **Crépon M.**, **Thiria and S.** (2012) Atmospheric correction of meris data for case-2 waters using a neuro-variational inversion. *Remote Sensing of Environment 126*, 51–61.

Cowtan K and Way R (2014) Coverage bias in the hadcrut4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society 140*, 1935–1944.

Diouf D, **Thiria S**, **Niang A.**, **Brajard J. and Crépon M** (2013). Retrieving aerosol characteristics and sea-surface chlorophyll from satellite ocean color multi-spectral sensors using a neural-variational method In Remote Sensing of Environment *130*, 74–86.

Eyring V, **Bony S**, **Meehl G. A.**, **Senior C. A.**, **Stevens B.**, **Stouffer R. J., and Taylor K. E.**(2016) Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development 9*, 1937–1958.

Gillett N, **Kirchmeier-Young M**, **Ribes A.**, **Shiogama H.**, **Hegerl G. C.**, **Knutti R.**, **Gastineau G.**, **John J. G.**, **Li L. and Nazarenko L.** (2021) Constraining human contributions to observed warming since the pre-industrial period. *Nature Climate Change 11*, 207–212.

Gillett N, **Shiogama H**, **Funke B.**, **Hegerl G.**, **Knutti R.**, **Matthes K.**, **Santer B. D.**, **Stone D. and Tebaldi C.** (2016) The detection and attribution model intercomparison project (damip v1.0)contribution to cmip6. *Geoscientific Model Development 9*, 3685–3697.

Gulev T., and Dentener A., **Domingues, Gerland, G., Kaufman, N, Quaas, R, Sathyendranath, S,Trewin, S, & Vose**. (2021) Changing state of the climate system. in climate change 2021: The physical science basis. *contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 287–422.

Hasselmann K (1993) Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate 6*(10), 1957–1971.

Kelley M, **Schmidt GA**, **Nazarenko L. S.**, **Bauer S. E.**, **Ruedy R.**, **Russell G. L.**, **Ackerman A. S.**, **Aleinov I.**, **Bauer M.**, **Bleck R.**, **Canuto V.**, **Cesana G.**, **Cheng Y.**, **Clune T. L.**, **Cook B. I.**, **Cruz C. A.**, **Del Genio A. D.**, **Elsaesser G. S.**, **Faluvegi G.**, **Kiang N. Y.**, **Kim D.**, **Lacis A. A.**, **Leboissetier A.**, **LeGrande A. N.**, **Lo K. K.**, **Marshall J.**, **Matthews E. E.**, **McDermid S.**, **Mezuman K.**, **Miller R. L.**, **Murray L. T.**, **Oinas V.**, **Orbe C.**, **García-Pando C. P.**, **Perlwitz J. P.**, **Puma M. J.**, **Rind D.**, **Romanou A.**, **Shindell D. T.**, **Sun S.**, **Tausnev N.**, **Tsigaridis K.**, **Tselioudis G.**, **Weng E.**, **Wu J. and Yao M.-S.** (2020) Giss-e2.1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems 12*, e2019MS002025.

Lehner F and Coats S (2021) Does regional hydroclimate change scale linearly with global warming? *Geophysical Research Letters 48*, e2021GL095127.

Morice C, **Kennedy J**, **Rayner N. A and Jones P** (2012) Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The hadcrut4 data set. *Journal of Geophysical Research: Atmospheres 117*. D08101.

Pincus R, **Forster PM and Stevens B** (2016) The radiative forcing model intercomparison project (rfmip): Experimental protocol for cmip6. *Geoscientific Model Development 9*, 3447–3460.

Rayner N, **Parker D**, **Horton E.**, **Folland C.**, **Alexander L.**, **Rowell, D.**, **Kent E. and Kaplan A.** (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres 108* D14.

https://doi.org/10.1017/eds.2022.17 Published online by Cambridge University Press

Reichstein M., **Camps-Valls G.**, **Stevens B.**, **Jung M.**, **Denzler J.**, **Carvalhais N.** (2019) Deep learning and process understanding for data-driven earth system science. *Nature 566*(7743), 195–204.

Ribes A, **Planton S and Terray L** (2013) Application of regularised optimal fingerprinting to attribution. Part i: Method, properties and idealised analysis. *Climate Dynamics 41*(11–12), 2817–2836.

Richardson M, **Cowtan K and Millar R** (2018) Global temperature definition affects achievement of long-term climate goals. *Environmental Research Letters 13*, 054004.

Stott P, **Gillet N.**, **Hegerl G.**, **Karoly D.**, **Stone D.**, **Zhang X. and Zwiers F** (2010) Detection and attribution of climate change: A regional perspective. *Wiley Interdisciplinary Reviews: Climate Change 1*, 192–211.

Zhang X, **Zwiers F**, **Hegerl G. C.**, **Lambert F. H.**, **Gillett N. P.**, **Solomon S.**, **Stott P. A. and Nozawa T** (2007) Detection of human influence on twentieth-century precipitation trends. *Nature 448*, 461–465.

Cite this article: Bône C. Gastineau G. Thiria S. and Gallinari P. (2022). Detection and attribution of climate change: A deep learning and variational approach. *Environmental Data Science*, 1: e27. doi:10.1017/eds.2022.17

4.3 Article de *Journal of Advances in Modelling Earth Systems*

Ce second manuscrit est la continuation directe du travail effectué dans l'article de *Environmental Data Sciences*. Il vise toujours à présenter la nouvelle méthode de détection et attribution du changement climatique basé sur l'optimisation inverse. Nous avons poursuivi notre étude en altérant notre méthodologie afin de mieux prendre en compte les spécificités du problème et de mieux évaluer la méthodologie. De nombreuses différences existent entre cette étude et celle de *Environmental Data Sciences*.

La première vient du prétraitement des données. La même base de données issue de CMIP6 est utilisé, mais des simulations Pi-Control sont également utilisées. Elles servent à obtenir des anomalies de températures pour les simulations DAMIP. Dans l'article de *Environmental Data Sciences* pour obtenir de telles anomalies, la valeur moyenne des simulations DAMIP dans la période 1850-1900 était retiré.

La seconde différence provient de différences méthodologiques. L'architecture du réseau de neurones et les hyper-paramètres de l'entraînement de celui-ci ou de l'optimisation inverse ont été testés et évalués au lieu d'être choisi de manière arbitraire. La fonction de coût de l'optimisation inverse comporte également un nouveau terme qui pénalise les variations hautes fréquences des changements attribuables aux aérosols et gaz à effet de serre. En effet, ces deux forçages n'ont qu'un effet de long terme sur le climat (Gulev et al., 2021).

Le dernier et plus important changement concerne l'évaluation de la méthode d'optimisation inverse. La méthodologie de test en modèle parfait est reprise, mais de manière plus exhaustive : tous les membres historiques de tous les modèles climatiques sont utilisés comme pseudo-observations. Un data set synthétique comportant de fortes non-additivités a été également

créé. Il permet de tester ce qui représente la principale plus-value de la méthode d'optimisation inverse : la prise en compte des non-additivités entre les forçages. Pour comparaison, une méthode d'empreintes optimisées (voir partie 2.4.2) est également implémenté avec le même jeu de données que pour l'optimisation inverse et testé également en modèle parfait.

Les résultats obtenus sur les données synthétiques montrent de meilleurs résultats pour la méthodologie d'optimisation inverse que la méthode des empreintes optimisée. Ce résultat illustre la plus-value de la méthode de l'optimisation inverse de prendre en compte les non-additivités des forçages sur la méthode des empreintes optimisées. L'approche en modèle parfait montre des résultats assez équivalents en termes de performance pour les deux méthodes, mais avec des différences concernant les intervalles de confiance provenant des différences de méthodologies de calculs. Les résultats obtenus sur les observations sont globalement similaires : un réchauffement marqué pour les gaz à effet de serre, un refroidissement pour les aérosols anthropiques et un effet globalement nul pour les forçages naturels excepté un bref refroidissement durant les éruptions volcaniques.

Cette méthode d'optimisation inverse est une nouvelle méthode d'attribution comparable à celles décrites dans la section 2.4 et qui donne des évaluations de l'effet des forçages globalement assez similaire quand appliquée à la GSAT (voir Fig. 2.13). Elle permet donc de confirmer une nouvelle fois les conclusions de ces études : l'action de l'Homme est prédominante dans les changements climatiques observés. Les gaz à effet de serre en particulier sont responsables de la quasi-totalité du réchauffement actuel alors que les aérosols refroidissent de moindres manières le climat. Comme sa plus-value de prendre en compte les non-additivités a été démontrée, elle pourra facilement être poursuivie et appliquée à d'autres variables physiques comme les précipitations (Marvel et al., 2015), ou à des échelles régionales (Good et al., 2015).

Detection and attribution of climate change using a neural network

Constantin Bône[1,2], Guillaume Gastineau[1], Sylvie Thiria[1], Patrick Gallinari[2,3]and Carlos Mejia[1]

[1]UMR LOCEAN, IPSL, Sorbonne Université, IRD, CNRS, MNHN
[2]UMR ISIR, Sorbonne Université, CNRS, INSERM
[3]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`

Abstract

A new detection and attribution method is presented and applied to the global mean surface air temperature (GSAT) from 1900 to 2014. The method aims at attributing the climate changes to the variations of greenhouse gases, anthropogenic aerosols, and natural forcings. A convolutional neural network (CNN) is trained using the simulated GSAT from historical and single-forcing simulations of twelve climate models. Then, we perform a backward optimization with the CNN to estimate the attributable GSAT changes. Such a method does not assume additivity in the effects of the forcings. The uncertainty in the attributable GSAT is estimated by sampling different starting points from single-forcing simulations and repeating the backward optimization. To evaluate this new method, the attributable GSAT changes are also calculated using the regularized optimal fingerprinting (ROF) method. Using synthetic non-additive data, we first find that the neural network-based method estimates attribuable changes better than ROF. When using GSAT data from climate model, the attribuable anomalies are similar for both methods, which might reflect that the influence of forcing is mainy additive for the GSAT. However, we found that the uncertainties given both methods are different. The new method presented here can be adapted and extended in future work, to investigate the non-additive changes found at the local scale or on other physical variables.

Plain Language Summary

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

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

attributable to greenhouse gases reported is +1.0°C to +2.0°C, while it is from -0.8°C to 0.0°C for other anthropogenic forcings, and from -0.1°C to +0.1°C for natural forcings.

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

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

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

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

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

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

2 Model and Data

2.1 Climate models simulations

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

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

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

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

All simulations but PI includes multiple realizations called ensemble members and denoted later as members. The members use different initial conditions which are sampled from the PI simulation. We use 12 atmosphere-ocean general circulation models (AOGCMs, see Tab. 1 for details) where at least two members are available for the simulations HIST, 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. σ_{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)	σ_{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)

2.2 Observations

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).

2.3 Pre-processing

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

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

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

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

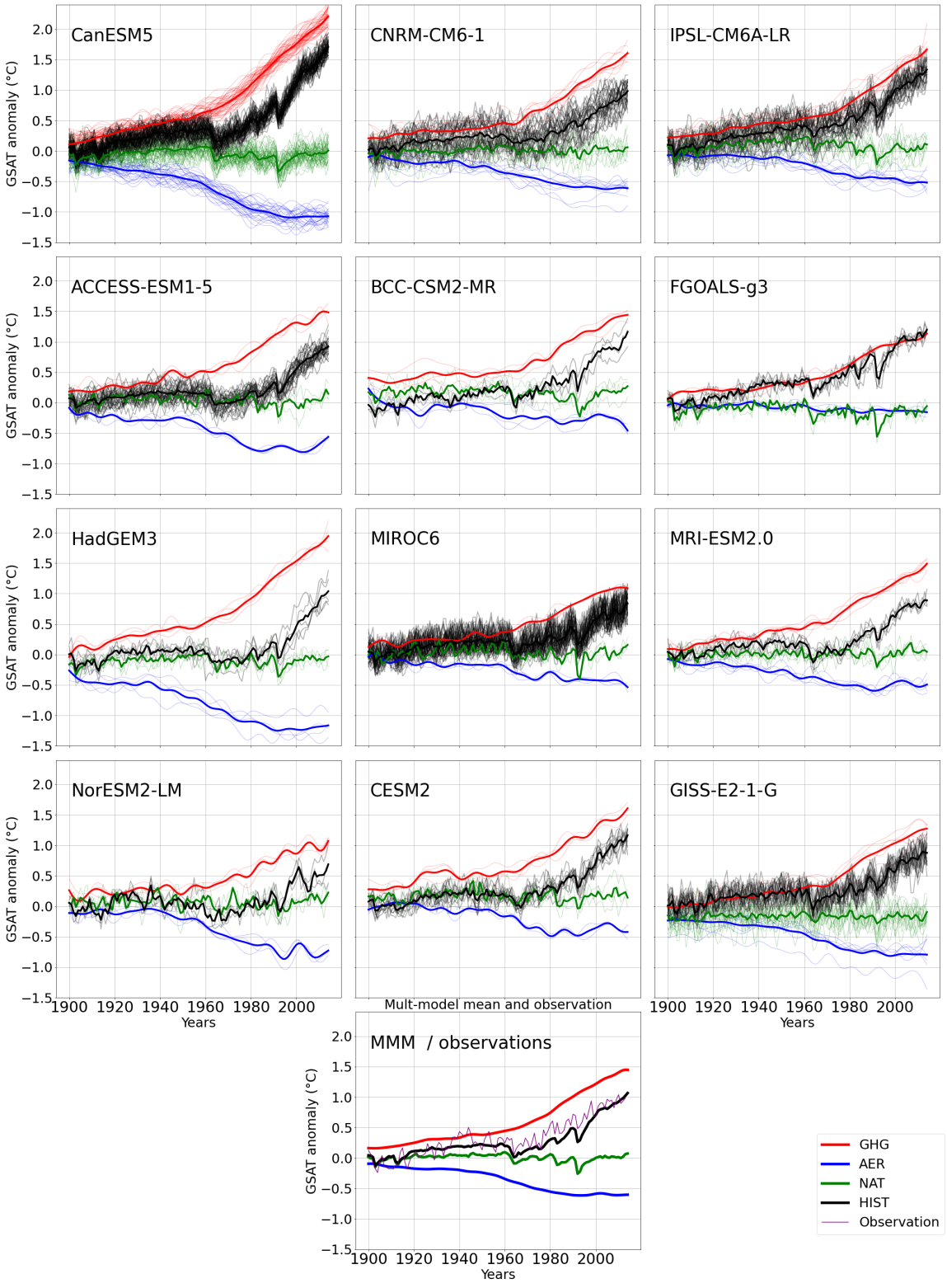

Figure 1. GSAT anomaly simulated by each model and (lower panel only) multi-model mean (MMM) and observed GSAT. Black lines show the HIST members. Red lines show the GHG members. Green lines show the NAT members. Blue lines show the AER members. The purple line shows the observations in the lower panel.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

2.4 Synthetic data

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 :

$$r = f_1 + 0.3f_1^2 + f_2 + f_3 + 0.1f_1f_2 \tag{1}$$

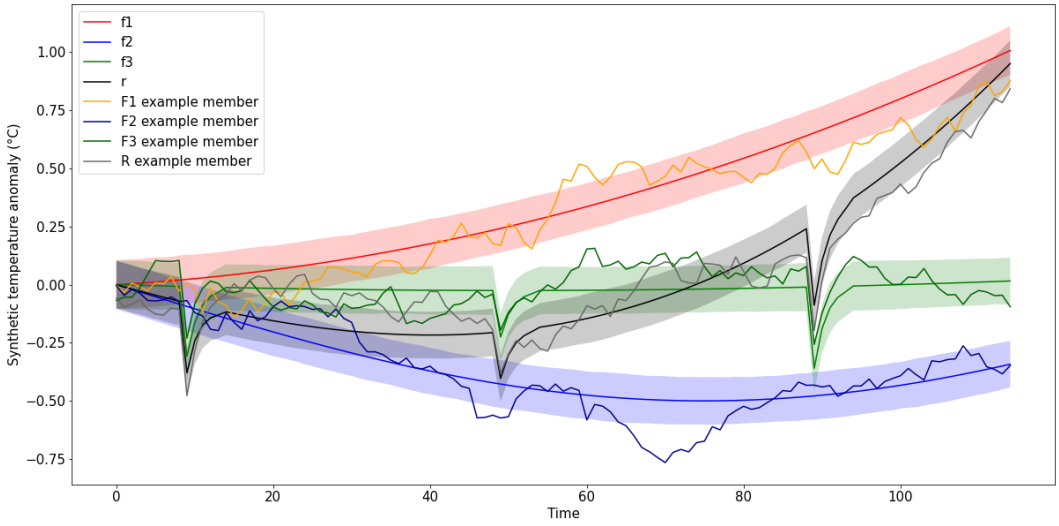

Figure 2. Synthetic time series f_1 (red), f_2 (blue), f_3 (green), and r (black). A randomly chosen time series after adding the variability is illustrated for F_1 (orange), F_2 (dark blue), F_3 (dark green) and R (grey). Colors shades indicate one standard deviation across the 100 surrogate time series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

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-

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

3 Methods

3.1 Backward optimization of a neural network

3.1.1 Neural network

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

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

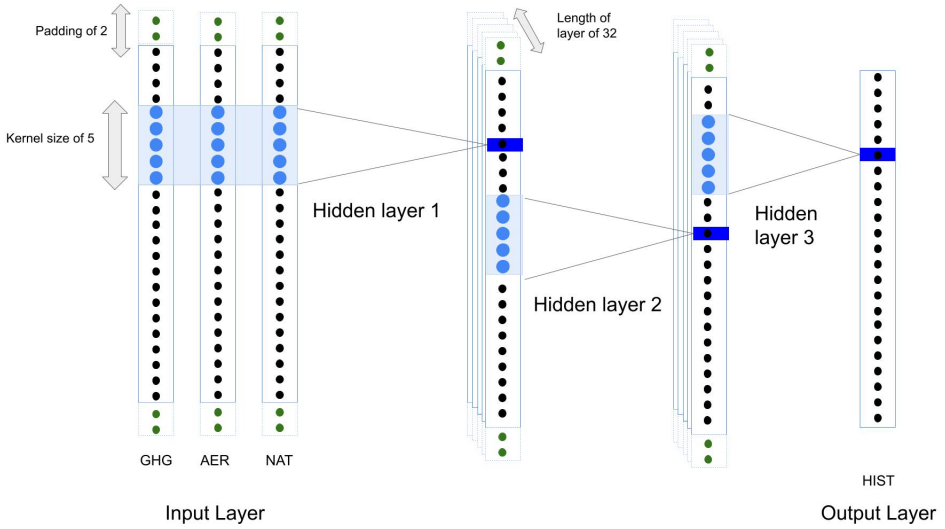

Figure 3. Diagram of the CNN used. Each white-filled blue rectangles represents a time series of 115 years. The input layer is shown on the left, the hidden layers on the middle and the output layer on the right. Light blue-filled rectangles represent the kernels of the different hidden layers. Dark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

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

Three steps are applied. First, a climate model is randomly selected. Secondly, we randomly select one 4-tuple from the chosen climate model. Then the CNN is trained using the three GSAT time series dedicated to (GHG, AER, NAT) as input and the GSAT dedicated to HIST as the target. We iterate this process by repeating it 5.10^6 times. A lower number of iterations was found to degrade the backward optimization results (not shown), but the results are otherwise similar when increasing the number of iterations.

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

3.2 Performance of the CNN

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

The CNN also should produce an estimated GSAT similar to the mean output from the training data, which is expected to be similar to the MMM from HIST. The training RMSE may also reflect a forced signal in the HIST simulations distinct from the other models. The amplitude of the RMSE increases to 0.15°C-0.35°C when using cross-validation. This suggest that the CNN does not overfit. HadGEM3 and, to a lesser extent, FGOALS-g3 and GISS-E2-1-G, show differences much larger than the training RMSE when the data from these models is used for the validation. This might reflect important singularities for these three models, which is probably linked to their singular response to forcings. This might be linked to the equilibrium climate sensitivity which quantifies the ability of a model to warm up when greenhouse gases increase. It depends on the feedbacks acting in the climate system, and remains poorly constrained by observations (Sherwood et al., 2020). GISS-E2-1-G simulate one of the lowest equilibrium climate sensitivity, while HadGEM2 has one of the highest sensitivity. In addition, FGOALS-g3 simulate almost no response to anthropogenic aerosols (see Fig. 1)

3.2.1 Backward optimization

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

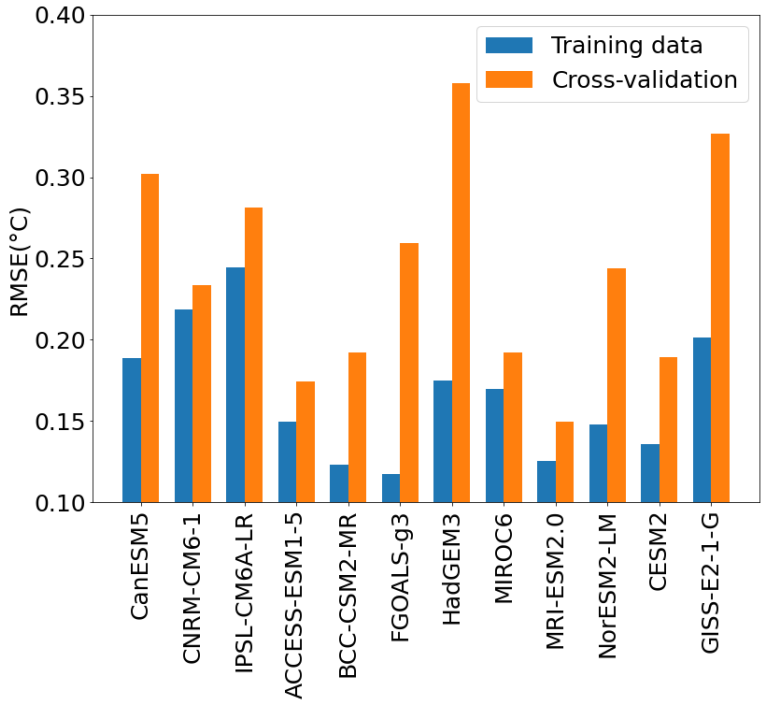

Figure 4. RMSE between the CNN output and the GSAT of HIST, in °C, when using (blue bar) the training data and (cross validation, orange bar) when using the data of a model left out in the training.

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:

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

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. σ_{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,

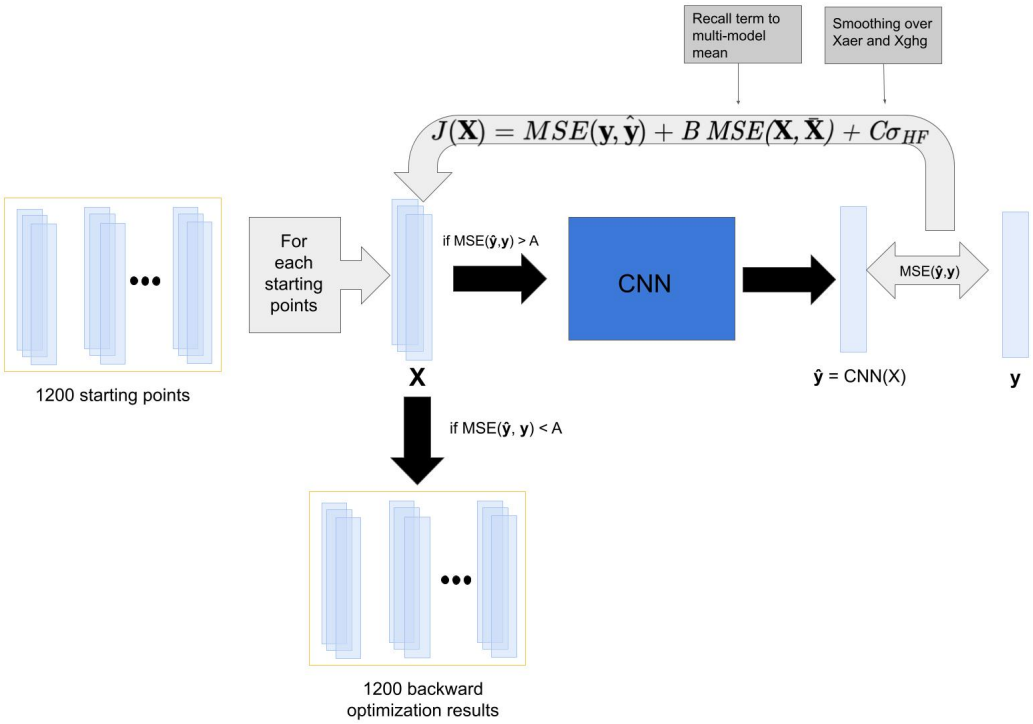

Figure 5. Schematic of the backward optimization attribution process with one entry denoted **y** at the right. The 1200 backward optimization results are at the bottom. The learned CNN is in the middle in dark blue. J(**X**), the cost function of the backward optimization is on the top. **X** denote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

AER, and NAT as first guess for the initial states. These initial states are chosen as they represent physically coherent inputs. This provides 1200 initial physically coherent values for **X** which sample the internal climate variability and the spread among the different models. This generates 1200 backward optimizations. This estimation is empirical and does not account for the internal variability of the target of the backward optimization. For each year, the 90% confidence intervals of the optimal input is then estimated using $\pm$ 1.64 standard deviations among the backward optimization results assuming a Gaussian distribution.

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

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:

$$\mathbf{y} = \beta \mathbf{X} + \epsilon \tag{3}$$

with $\beta = (\beta_{GHG}, \beta_{AER}, \beta_{NAT})$ the scaling factors and and $\mathbf{X} = (X_{GHG}, X_{AER}, X_{NAT})$ the effects of all the forcings on the GSAT. ϵ 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 β_i is significantly different from zero. If the confidence interval of β_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 $\beta_i X_i$

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

$$N = \frac{M^2}{\sum_{i=1}^{M} \frac{1}{n_i}} \tag{4}$$

with M the number of different climate models used (in our case 12) and n_i the number of members available for the i-th climate model. To estimate the uncertainty in the GSAT effect attributable to the i-th forcing, it is necessary to take into account the uncertainty of β_i and the internal variability contained in X_i. For each year and forcing, the uncertainty in the attributable GSAT is calculated using 1000 random draws assuming a gaussian distribution for both β_i and X_i. The mean and standard errors of β_i are estimated as in Allen and Stott (2003). The mean and standard deviation of X_i are estimated from the size N of the MMM and the standard deviation of the GSAT obtained from the PI runs. We first calculate the standard deviation for each model (as given in Tab. 1), average the values obtained across models, and then divide by the square root of N. This procedure is valid under the conditions that the uncertainties of β_i and X_i are Gaussian, uncorrelated and small compared to their respective means. The latter hypothesis is not verified for GSAT anomalies close to zero for X_i, such as those obtained in the first decades of our time series (see Fig. 1), or for the GSAT of NAT. Thus the uncertainties for the attributable GSAT are to be taken with caution.

4 Attribution performances

4.1 Performance on synthetic data

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

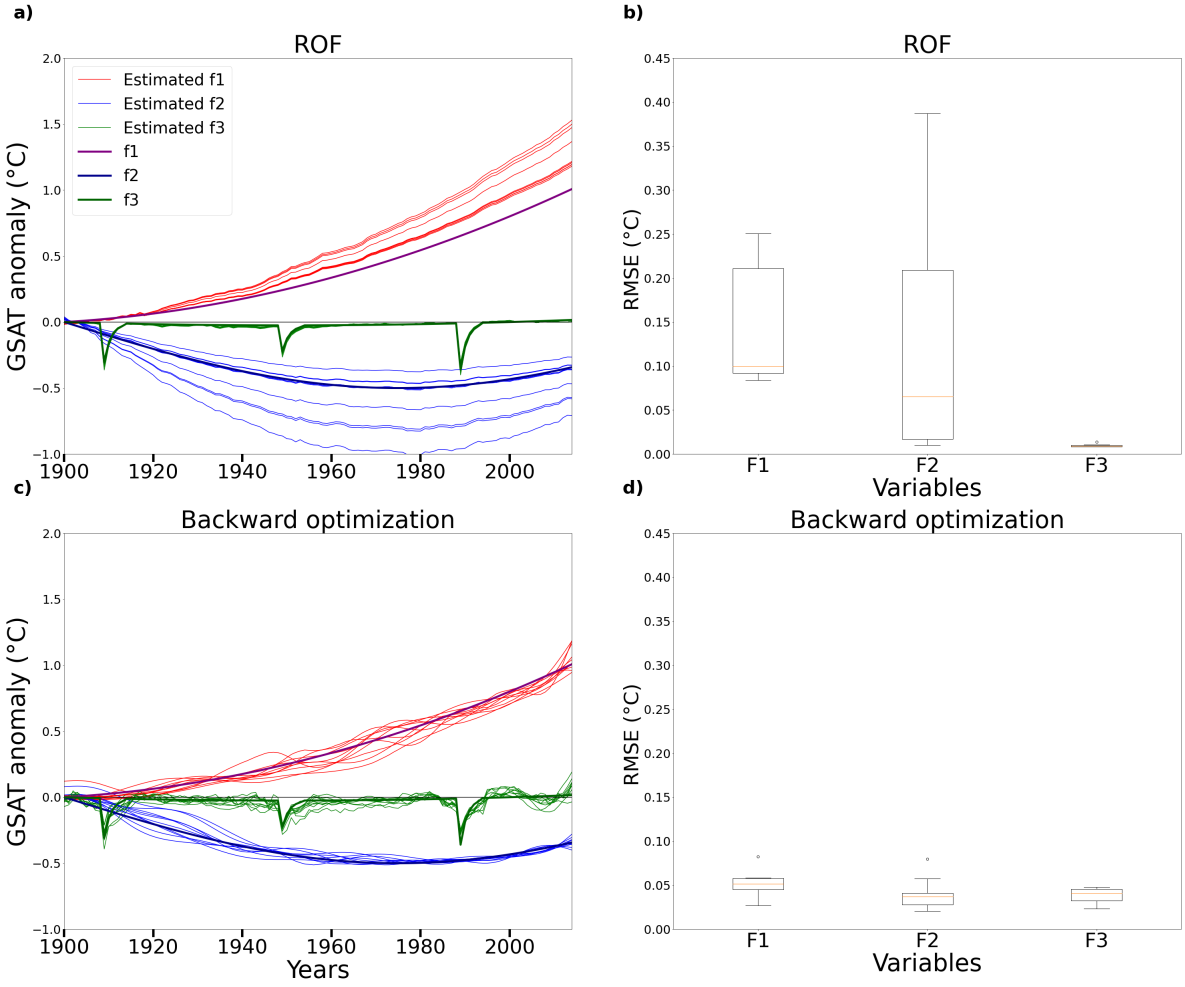

Figure 6. Estimated f_1, f_2 and f_3 given by (a) ROF and (c) backward optimization. The original f_1, f_2 and f_3 ground truth lines are shown in bold. Histograms shows the distribution of the RMSE of the results of b) ROF and d) backward optimization compared to the ground truth.

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

The RMSE between the effect estimated by the different attribution methods and the ground truth are shown in 6bd in the form of boxplot. For ROF the mean RMSE value is 0.14°C for f_1, 0.12°C for f_2 and 0.01°C for f_3. These values are for backward optimization of 0.05°C for f_1, 0.04°C for f_2 and 0.04°C for f_3. Backward optimization therefore provides errors smaller than ROF in case of the non-additive forcing generated, while the use of ROF lead to important errors.

4.2 Evaluation of the performances in attributing climate changes : perfect model approach

To evaluate the performance of the backward optimization and ROF we use a perfect model approach that relies on climate model data only. This approach consists of using the data from all but one of the climate models to perform our two attribution methods. In the case of backward optimization, this implies that we do not use the data from a climate model during the CNN training phase, in the starting points, or in the MMM calculation. For ROF, the data of a model are not used to construct the climate noise estimate or included in the MMM. We use a HIST member of the test dataset (see Section 3.3) from each climate model as the target for the attribution methods. The attributable anomalies associated with each forcing are then compared with the ensemble mean of the GHG, AER and NAT simulations of the removed climate model, even if it includes some residual internal variability, especially when the number of members is small. We use the paradigm that "climate models are statistically indistinguishable from the truth" (Ribes et al., 2017; Hargreaves, 2010; van Oldenborgh et al., 2013), where the difference 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°C, 0.20°C and 0.12°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°C, 0.15°C and 0.12°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°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

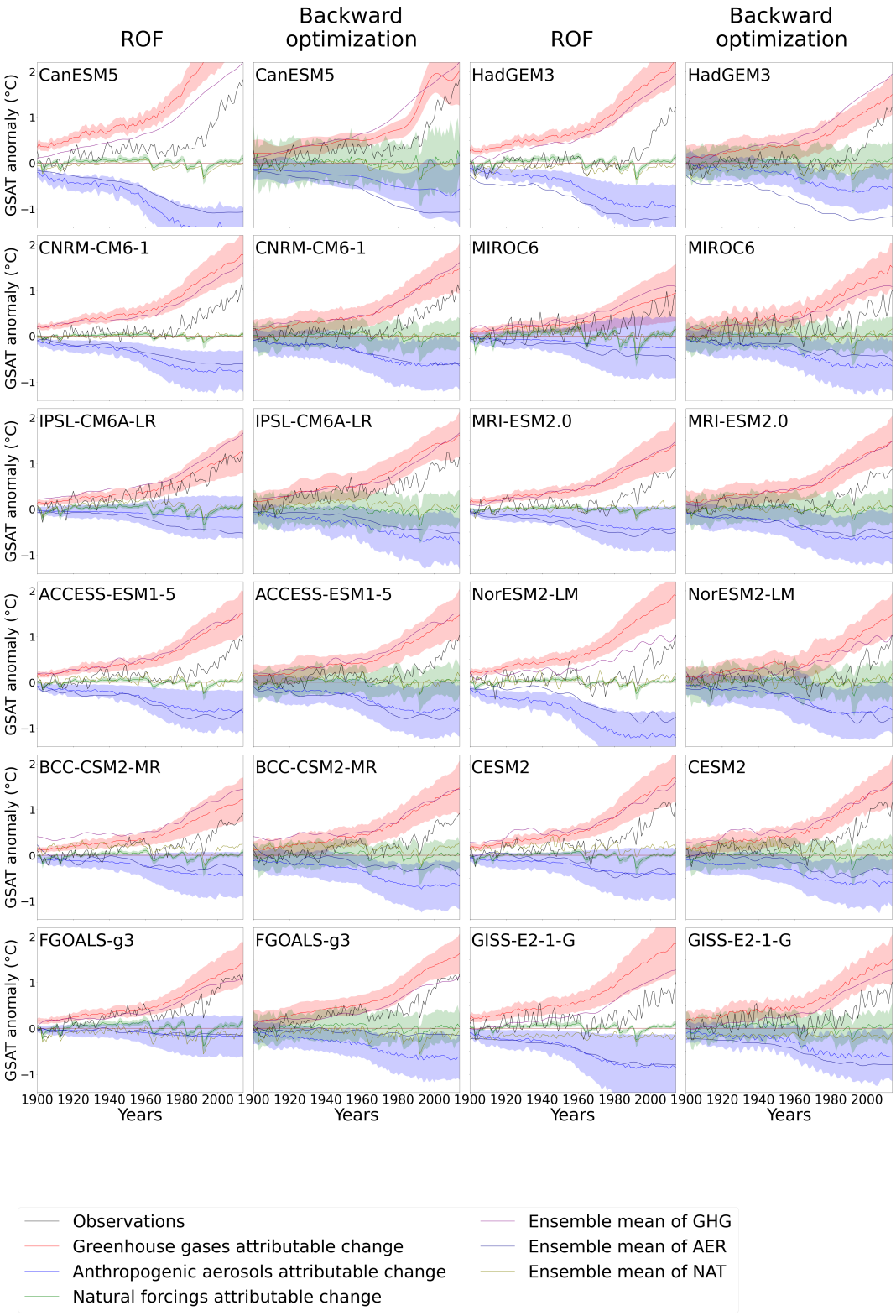

Figure 7. Attributable GSAT in °C calculated for ROF and backward optimization from (black line) a HIST member. The GSAT is decomposed into the attribuable changes due to (red line) greenhouse gases; (blue line) anthropogenic aerosols and (green line) natural forcings. For comparison, the ensemble mean 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.

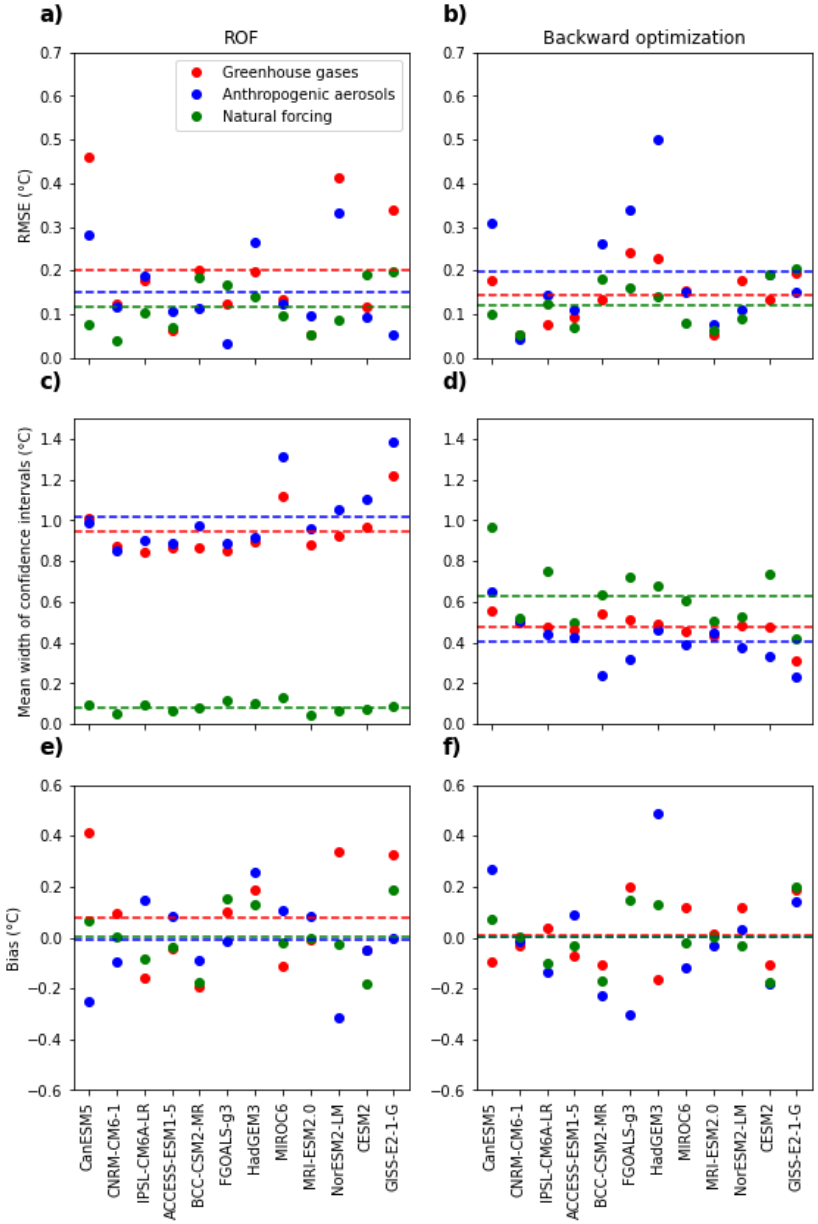

Figure 8. Performances of attribution methods using a perfect model approach. a) RMSE when using ROF for the attributable GSAT anomaly of (red) greenhouse gases, (blue) anthropogenic aerosols, (green) natural forcing. b) Same as a) for the backward optimization. c) Width of the 90% percent confidence 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 attribuable to the forcings when using ROF. f) Same as e) for backward optimization. Dashed lines shows average values across the 12 climate models.

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

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

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

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

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

4.3 Attribution of the observed GSAT

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

The attributable GSAT changes are similar for ROF and backward optimization (see Fig. 9). For example, in 2000-2014, ROF provides a GSAT attributable to greenhouse gases of 1.28°C (90% confidence interval of [0.85°C,1.71°C]), while it is -0.33°C ([-0.80°C,0.12°C]) for anthropogenic aerosols and 0.01°C ([0.0°C,0.02°C]) for natural forcing. In comparison, backward optimization finds attributable changes of 1.42°C ([1.03°C,1.80°C]), -0.61°C ([-1.16°C,-0.06°C]) and 0.02°C ([-0.33°C,0.38°C]), respectively, for these three forcings. Nevertheless, backward optimization provides more noisy time series and more cooling during volcanic eruptions. The similarity of the results between ROF and backward optimization suggests that the GSAT changes are largely additive as found in Marvel et al. (2015) or Shiogama et al. (2013).

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

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

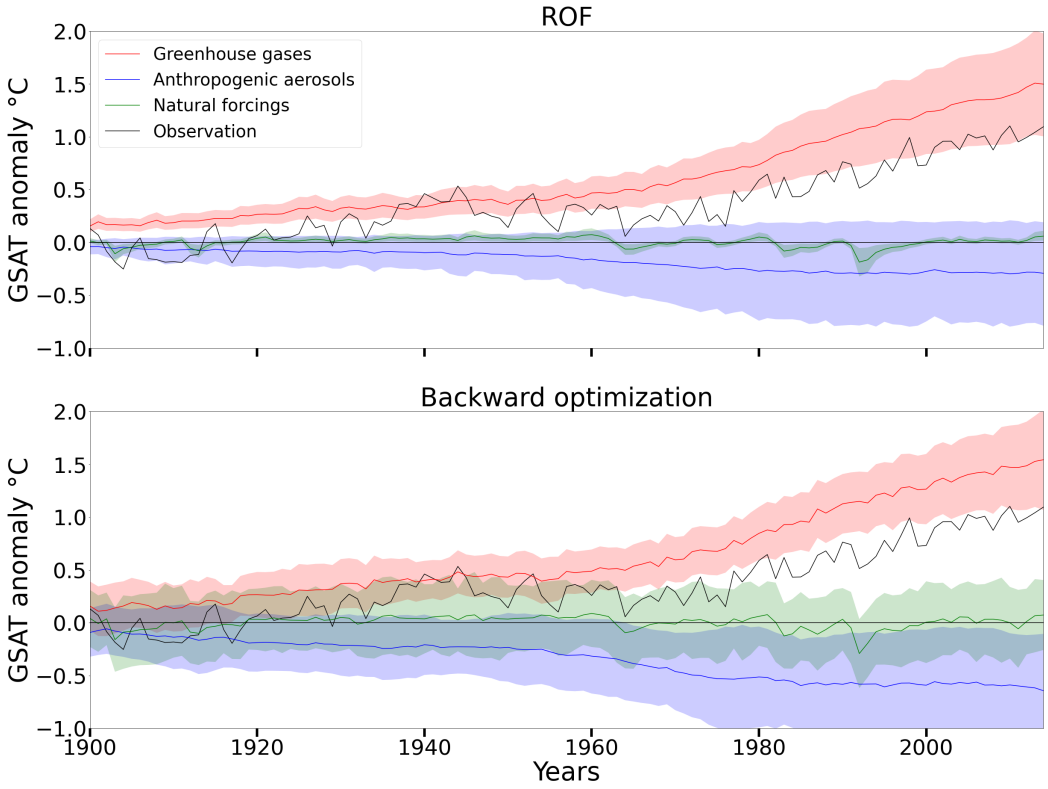

Figure 9. (Top) Attributable GSAT anomaly, in °C, as given by ROF for the effect of the (red) greenhouse gases, (green) natural forcings and (blue) anthropogenic aerosols. The black line shows the observed GSAT. The color shade shows the 90% confidence interval. (Bottom) : same as top, but for backward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

tion very close to the observations (see Fig. S4) and captures most of the internal variability contained within the observations.

4.4 Focus on the main backward optimization results

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

The attributable changes produced by the backward optimization are influenced by the climate model used to generate the initial input. For example, CanESM5 simulates large warming in response to greenhouse gases (see Fig. 1), probably linked to its large equilibrium climate sensitivity. When using the outputs of CanESM5 as initial input of the backward optimization, large attributable changes are obtained for both the greenhouse gases and the anthropogenic aerosols. On the other hand, when using an initial input from FGOALS-g3 the changes due to the greenhouse gases and the anthropogenic aerosols are small. For each forcing, we analyse the dispersion of the GSAT anomalies over the years 2000-2014 by estimating the mean GSAT attributable to the use of all starting points for each of the 12 climate models. The variability explained by the model is calculated by is the standard deviation across these 12 attributable GSAT. The residual variability which accounts for the internal variability of the starting points is estimated by the standard deviation of the 1200 attributable GSAT after subtracting for each time series the average response obtained with their respective climate model. The standard deviation explained by the model of the starting point is 0.22°C for greenhouse gases, 0.29°C for anthropogenic aerosols and 0.14°C for natural forcings. The residual standard deviation is of 0.06°C for the greenhouse gases, 0.09°C for the anthropogenic aerosols and 0.1°C for the natural forcings. The residual variance therefore is smaller than this associated with the climate model for each forcing, especially for the greenhouse gases. The range of attribution results is about 1°C for all forcings, with some particular models providing attribuable anomalies at the head or the tail of the inter-models distribution when used as starting point. Removing or modifying these outliers to improve the

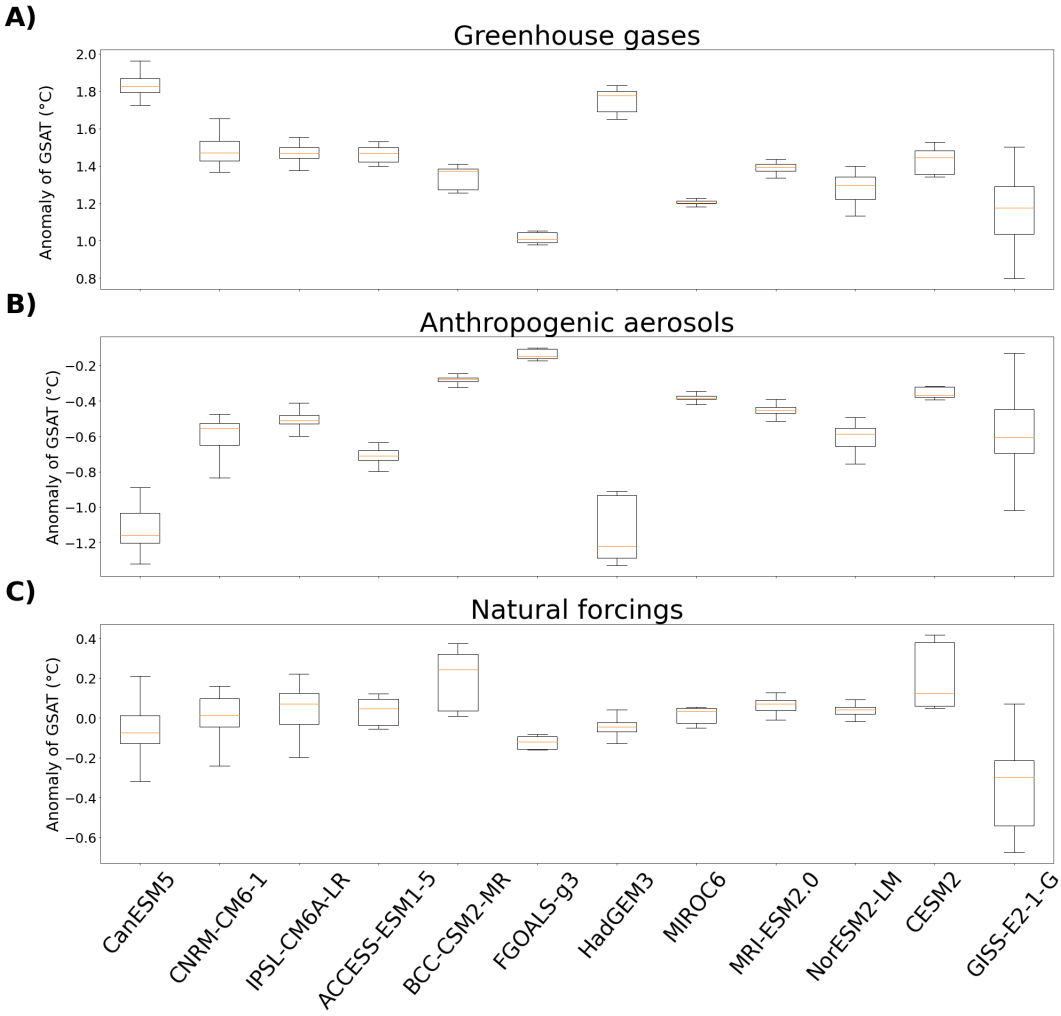

Figure 10. Boxplots of the attributable changes in 2000-2014 when using observation and backward optimization, classified according to the climate model used as initial inputs for (A) the greenhouse gases (B) the anthropogenic aerosols and (C) the natural forcings.

backward optimization results have been considered. However, selecting these initial inputs may imply a selection of climate models which needs to be associated with a careful investigation of the physical mechanisms (Coquard et al., 2004).

5 Discussion and conclusion

We present a method for detection and attribution of climate data based on a backward optimization of a convolutional neural network (CNN). We trained the CNN on the simulated GSATs obtained from outputs of twelve CMIP6 climate models. We then performed a backward optimization to estimate the attributable changes. This methodology does not assume that the effects of the external forcings are additive. Such additivity implies that the total changes simulated by the forcings can be obtained by the sum of the changes due to the individual forcings. The additivity assumption is an important limitation when focusing on precipitation (Marvel et al., 2015) or at regional scale (Pope et al., 2020; Deng et al., 2020). We evaluated the effect of internal variability and model dispersion by using different starting points sampling the simulated distributions. We compared the results of the CNN backward optimization with those obtained using the regularized optimal fingerprinting (ROF) (Allen & Stott, 2003; Ribes et al., 2013). In order to assess the ability of backward optimization to deal with non-additivities in forcing compared to ROF we used synthetic data, which, unlike GSAT, have a strong non-additive behavior. In that case, the backward optimization results are more similar to the true forced effect of the forcings than when using ROF which assumes additivity. To see if this results can be generalised additional investigations need to be conducted using either different synthetic data or real non-additive climate data, as for instance the precipitation field.

We also designed a perfect model approach to evaluate the skill of the two methods. We successively removed the data of each climate model and used an historical member of the removed climate model as pseudo-observation. The attributable changes of each forcing are then compared to their actual effect simulated in the corresponding ensemble mean of single-forcing simulations. Backward optimization is found to provide performances similar to that obtained with ROFs in terms of RMSEs or bias. The confidence intervals of the backward optimization are smaller for greenhouse gases and anthropogenic aerosols in the last years of the studied period and much larger for natural forcings than those obtained by ROF. As the calculation of the uncertainty applied

in ROF has been previously shown to be also underestimated (DelSole et al., 2019), this suggests that backward optimization leads to an even larger underestimation. This might be linked to the internal variability of the target time series, which is not accounted for in the neural network-based method. A solution to solve this issue would be to generate surrogate time series for the backward optimization and repeat the backward optimization. Larger ensemble of single forcing simulations, such as those proposed in the Large Ensemble Single Forcing Model Intercomparison Project (D. M. Smith et al., 2022), would also be required to refine of the estimated errors. In addition, the changes attribuable to natural forcings in the backward optimization have a larger uncertainty than the one of ROF. This is suggested to be an artefact of the estimated uncertainty used, which may be flawed for small changes. Many aspects of the backward optimization can be improved in future works. The backward optimization process can also be improved by giving weights based on the realistic simulation of the interannual to decadal variability. Indeed, the procedure presented here is designed to produce a close agreement between the reconstructed time series and the observations (or pseudo-observations). As shown in Fig. S4, the reconstructed time series, i. e. the image of the CNN using the backward optimization results, closely follow the observations. The CNN might instead be designed to only reproduce the forced component of the anomalies excluding the internal variability unrelated to climate forcings. A better treatment of the initial state could be also investigated, excluding or penalizing the time series used as initial input when inconsistent with observations. In addition, giving different weights to each climate models according to their performance in reproducing observed features could be considered, such as the observed GSAT evolution in Ribes et al. (2021).

Overall, the attributable changes obtained with the backward optimization are consistent with recent attribution results, as reviewed in Eyring et al. (2020a). This confirms the previous detection and attribution results on the GSAT. This study also shows that neural networks can be used to explore the CMIP databases through the backward optimization presented here. Such a method could be deployed on other physical variables, such as precipitation. It could also easily be applied to spatial average instead of global mean where the non-additivities could be an obstacle. Lastly, a similar method applied on gridded data could also be considered without major modifications given that 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.

Brajard, J., Santer, R., Crépon, M., & Thiria, S. (2012). Atmospheric correction of MERIS data for case-2 waters using a neuro-variational inversion. *Remote Sensing of Environment*, *126*, 51–61.

Burger, W., & Burge, M. J. (2009). *Principles of digital image processing: core algorithms*. Springer London.

Bône, C. (2023). *Codes for "Detection and attribution of climate change" [Software].* Retrieved from `https://doi.org/10.5281/zenodo.7248662`

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.-S., & Caltabiano, N. (2018). Decadal climate variability and predictability: Challenges and opportunities. *Bulletin of the American Meteorological Society*, *99*(3), 479–490.

Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., ... others (2022). Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, *8*(1), 1–26.

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., ... others (2014). The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data [Dataset]. *Future Generation Computer Systems*, *36*, 400–417.

Coquard, J., Duffy, P., Taylor, K., & Iorio, J. (2004). Present and future surface climate in the western USA as simulated by 15 global climate models. *Climate Dynamics*, *23*(5), 455–472.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D., DuVivier, A., Edwards, J., ... others (2020). The community earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916.

DelSole, T., Trenary, L., Yan, X., & Tippett, M. K. (2019). Confidence intervals in optimal fingerprinting. *Climate Dynamics*, *52*(7), 4111–4126.

Deng, J., Dai, A., & Xu, H. (2020). Nonlinear climate responses to increasing co2 and anthropogenic aerosols simulated by cesm1. *Journal of Climate*, *33*(1), 281–301.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. (2013). Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(4), 245–268.

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., ... others (2020a). Earth System Model Evaluation Tool (ESMValTool) v2. 0–an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *13*(7), 3383–3438.

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., ... Zimmermann, K. (2020b). *Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP [Software]* (Vol. 13) (No. 7). Retrieved from `https://github.com/ESMValGroup/ESMValTool/releases/tag/v2.9.0`

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958.

Eyring, V., Gillett, N., Rao, K. A., Barimalala, R., Parrillo, M. B., Bellouin, N., ... Zho, B. (2021). Human Influence on the Climate System. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Pres*.

Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., & Rousseau, F. (2021). Learning variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002572.

Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, *147*(8), 2827–2845.

Gillett, N. P., Kirchmeier-Young, M., Ribes, A., Shiogama, H., Hegerl, G. C., Knutti, R., ... others (2021). Constraining human contributions to observed warming since the pre-industrial period. *Nature Climate Change*, *11*(3), 207–212.

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., ... Tebaldi, C. (2016). The detection and attribution model intercomparison project (DAMIP v1. 0) contribution to CMIP6. *Geoscientific Model Develop-*

ment, *9*(10), 3685–3697.

Good, P., Lowe, J. A., Andrews, T., Wiltshire, A., Chadwick, R., Ridley, J. K., ... others (2015). Nonlinear regional warming with increasing CO2 concentrations. *Nature Climate Change*, *5*(2), 138–142.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gulev, S., Thorne, P., Ahn, J., Dentener, F., Domingues, C., Gerland, S., & Vose, R. (2021). Changing state of the climate system. In climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change.

Gupta, A. S., Jourdain, N. C., Brown, J. N., & Monselesan, D. (2013). Climate Drift in the CMIP5 Models. *Journal of Climate*, *26*(21), 8597 - 8615. doi: 10.1175/JCLI-D-12-00521.1

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572.

Hargreaves, J. C. (2010). Skill and uncertainty in climate models. *WIREs Climate Change*, *1*(4), 556-564. Retrieved from `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.58` doi: https://doi.org/10.1002/wcc.58

Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, *6*(10), 1957–1971.

Hawkins, E., Robson, J., Sutton, R., Smith, D., & Keenlyside, N. (2011). Evaluating the potential for statistical decadal predictions of sea surface temperatures with a perfect model approach. *Climate dynamics*, *37*(11), 2495–2509.

Hobbs, W., Palmer, M. D., & Monselesan, D. (2016). An energy conservation analysis of ocean drift in the CMIP5 global coupled models. *Journal of Climate*, *29*(5), 1639–1653.

Irving, D., Hobbs, W., Church, J., & Zika, J. (2021). A Mass and Energy Conservation Analysis of Drift in the CMIP6 Ensemble. *Journal of Climate*, *34*(8), 3157 - 3170. doi: 10.1175/JCLI-D-20-0281.1

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., ... others (2020). GISS-E2. 1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS002025.

Kennedy, J. J., Rayner, N. A., Atkinson, C. P., & Killick, R. E. (2019). An Ensemble Data Set of Sea Surface Temperature Change From 1850: The

Met Office Hadley Centre HadSST.4.0.0.0 Data Set. *Journal of Geophysical Research: Atmospheres*, *124*(14), 7719-7763. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JD029867` doi: https://doi.org/10.1029/2018JD029867

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, *151*, 107398. doi: https://doi.org/10.1016/j.ymssp.2020.107398

Labe, Z. M., & Barnes, E. A. (2021). Detecting climate signals using explainable AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002464.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., ... others (2022). GraphCast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*.

Li, Yu, Y., Tang, Y., Lin, P., Xie, J., Song, M., ... Wang, L. (2020). The flexible global ocean-atmosphere-land system model grid-point version 3 (FGOALS-g3): description and evaluation. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002012.

Li, Zwiers, F., Zhang, X., Li, G., Sun, Y., & Wehner, M. (2021). Changes in Annual Extremes of Daily Temperature and Precipitation in CMIP6 Models. *Journal of Climate*, *34*(9), 3441 - 3460. Retrieved from `https://journals.ametsoc.org/view/journals/clim/34/9/JCLI-D-19-1013.1.xml` doi: https://doi.org/10.1175/JCLI-D-19-1013.1

Li, Z., Zhang, W., Jin, F.-F., Stuecker, M. F., Sun, C., Levine, A. F., ... Liu, C. (2020). A robust relationship between multidecadal global warming rate variations and the Atlantic Multidecadal Variability. *Climate Dynamics*, *55*(7), 1945–1959.

Marvel, K., Schmidt, G. A., Shindell, D., Bonfils, C., LeGrande, A. N., Nazarenko, L., & Tsigaridis, K. (2015). Do responses to different anthropogenic forcings add linearly in climate models? *Environ. Res. Lett.*, *10*(10), 104010. doi: 10.1088/1748-9326/10/10/104010

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., ... Zhou, B. (2021). 2021: Changing State of the Climate System. In Climate

Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press*, 287-422.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199.

Meehl, G. A., Hu, A., Santer, B. D., & Xie, S.-P. (2016). Contribution of the Interdecadal Pacific Oscillation to twentieth-century global surface temperature trends. *Nature Climate Change*, *6*(11), 1005–1008.

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., ... Simpson, I. R. (2021). An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032361` (e2019JD032361 2019JD032361) doi: https://doi.org/10.1029/2019JD032361

Neelin, J. D., Battisti, D. S., Hirst, A. C., Jin, F.-F., Wakata, Y., Yamagata, T., & Zebiak, S. E. (1998). ENSO theory. *Journal of Geophysical Research: Oceans*, *103*(C7), 14261–14290.

Osborn, T. J., Jones, P. D., Lister, D. H., Morice, C. P., Simpson, I. R., Winn, J. P., ... Harris, I. C. (2021). Land Surface Air Temperature Variations Across the Globe Updated to 2019: The CRUTEM5 Data Set. *Journal of Geophysical Research: Atmospheres*, *126*(2), e2019JD032352. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032352` (e2019JD032352 2019JD032352) doi: https://doi.org/10.1029/2019JD032352

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Pope, J. O., Orr, A., Marshall, G. J., & Abraham, N. L. (2020). Non-additive response of the high-latitude Southern Hemisphere climate to aerosol forcing in a climate model with interactive chemistry. *Atmospheric Science Letters*, *21*(12), e1004. doi: https://doi.org/10.1002/asl.1004

Ribes, A., Planton, S., & Terray, L. (2013). Application of regularised optimal fingerprinting to attribution. Part I: method, properties and idealised analysis.

Climate dynamics, *41*(11), 2817–2836.

Ribes, A., Qasmi, S., & Gillett, N. P. (2021). Making climate projections conditional on historical observations. *Science Advances*, *7*(4), eabc0671.

Ribes, A., Zwiers, F. W., Azais, J.-M., & Naveau, P. (2017). A new statistical approach to climate change detection and attribution. *Climate Dynamics*, *48*(1), 367–386.

Richardson, M., Cowtan, K., & Millar, R. J. (2018). Global temperature definition affects achievement of long-term climate goals. *Environmental Research Letters*, *13*(5), 054004.

Roberts, M. J., Baker, A., Blockley, E. W., Calvert, D., Coward, A., Hewitt, H. T., ... others (2019). Description of the resolution hierarchy of the global coupled HadGEM3-GC3. 1 model as used in CMIP6 HighResMIP experiments. *Geoscientific Model Development*, *12*(12), 4999–5028.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., ... others (2020). Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations. *Geoscientific Model Development*, *13*(12), 6165–6200.

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., ... others (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678.

Shiogama, H., Stone, D. A., Nagashima, T., Nozawa, T., & Emori, S. (2013). On the linear additivity of climate forcing-response relationships at global and continental scales. *International Journal of Climatology*, *33*(11), 2542-2550. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3607` doi: https://doi.org/10.1002/joc.3607

Smith, Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., ... Michou, M. (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16), 9591–9618.

Smith, D. M., Gillett, N. P., Simpson, I. R., Athanasiadis, P. J., Baehr, J., Bethke, I., ... Ziehn, T. (2022). Attribution of multi-annual to decadal changes in the climate system: The Large Ensemble Single Forcing Model Intercomparison Project (LESFMIP). *Front. Clim.*, *4*, 955414. doi: 10.3389/fclim.2022.955414

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., ... others (2019). The Canadian earth system model version 5 (CanESM5. 0.3). *Geoscientific Model Development*, *12*(11), 4823–4873.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., ... others (2019). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, *12*(7), 2727–2765.

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002.

van Oldenborgh, G. J., Reyes, F. J. D., Drijfhout, S. S., & Hawkins, E. (2013, mar). Reliability of regional climate model trends. *Environmental Research Letters*, *8*(1), 014055. Retrieved from `https://dx.doi.org/10.1088/1748-9326/8/1/014055` doi: 10.1088/1748-9326/8/1/014055

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., ... others (2019). Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2177–2213.

Wild, M. (2009). Global dimming and brightening: A review. *Journal of Geophysical Research: Atmospheres*, *114*(D10). doi: https://doi.org/10.1029/2008JD011470

Wu, T., Lu, Y., Fang, Y., Xin, X., Li, L., Li, W., ... others (2019). The Beijing Climate Center climate system model (BCC-CSM): the main progress from CMIP5 to CMIP6. *Geoscientific Model Development*, *12*(4), 1573–1600.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, *9*(4), 611–629.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., ... others (2019). The Meteorological Research Institute Earth System Model version 2.0, MRI-ESM2. 0: Description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japan. Ser. II*.

Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M., ... Srbinovsky, J. (2020). The Australian earth system model: ACCESS-ESM1. 5. *Journal of Southern Hemisphere Earth Systems Science*, *70*(1),

889 193–214.

Supporting Information for "Detection and attribution of climate change using a neural network"

Constantin Bône[12], Guillaume Gastineau[1], Sylvie Thiria[1], Patrick Gallinari[23]and Carlos Mejia[1]

[1]UMR LOCEAN, IPSL, Sorbonne Université, IRD, CNRS, MNHN

[2]UMR ISIR, Sorbonne Université, CNRS, INSERM

[3]Criteo AI Lab

Contents of this file

1. Text S1 to S4
2. Figures S1 to S3
3. Tables S1 to S2

Introduction

This supporting information gives some details on the construction of the synthetic data. We present how the f_1, f_2 and f_3 time series are constructed. The methodology adopted for the choice of the hyperparameters for the neural network and the backward optimization is alsopresented.Then, the effect of the internal variability is investigated by repeating ROF and backward optimization using the HIST member from IPSL-CM6-LR the changes of the attributable anomalies are illustrated when accounting land use and ozone forcing in ROF. Lastly, we illustrate the reconstitution of the observation by the CNN.

Text S1. Synthetic dataset

We define three time series, f_1, f_2 and f_3 as $t \in \{1, 2, 3...115\}$:

$f_1 = 6.10^{-5}t^2 + 2.10^{-3}t$

$f_2 = -0.5sin(\frac{t\pi}{150})$

$f_3 = 1.10^{-5}t^2 - 1.10^{-3}t + f_{add}(t)$

f_{add} is a term added to represent the effect of three pseudo-volcanic eruptions for $t \in \{9, 49, 89\}$. This term is an additional anomaly that last for five years and is defined as : $f_{add} = e^{\frac{2}{3}(t-t_j)}$ if $t \in [t_j, t_j + 4]$ and $t_j \in \{9,49,89\}$ and 0 otherwise

Text S2. Choice of hyper-parameters of the neural network

The hyperparameters of the CNN are the number of hidden layers, the cost function, the non-linear activation function, the size of the kernel, the length of the hidden layers, the learning rate, the type of padding used, and the batch size. The effects of the type of padding, the activation function, the batch size and the learning rate have not been investigated. We use the RMSE cost function and zero-values padding. A non-linear activation function is used between the hidden layers of the neural network in our case the hyperbolic tangent function. To determine the other hyper-parameters we use a cross validation. We considered the data from the 12 models but leaving out the data of one climate model. We train a CNN using the remaining models. The process was repeated by excluding successively each climate model. For each CNN built we also select randomly a historical member of the climate model left out as pseudo-observations, and perform the backward optimization. We compare the results to the ensemble mean of the simulations for this climate model. The mean value of the 12 backward optimization RMSE, is illustrated in Fig. S1 for different sets of hyperparameters.

The backward optimization RMSE are between 0.18°C and 0.41°C. The number of filters of the layer shows the largest influence, with a reduction of the RMSE for increasing length of the hidden layers. The number of hidden layers and the kernel sizes does not affect the RMSE.

We choose the architecture that gives the lowest backward optimization RMSE while keeping a small number of weights and biases with three hidden layers, a kernel sizes of 5 and number of filters of 32.

Text S3. Choice of the hyper-parameter of the backward optimization

Tables S1 and S2 shows the mean RMSE of the backward optimization described, for differents values of A, B, and C. The difference of performance is small in all experiments. We noted that large values of A and B reduce dramatically the variability of results of the backward optimization (not shown) and select A=0.05, B=0.01 and C=0.1. We choose a non-zero value for B to keep a background term although it only has a marginal effect on the RMSE.

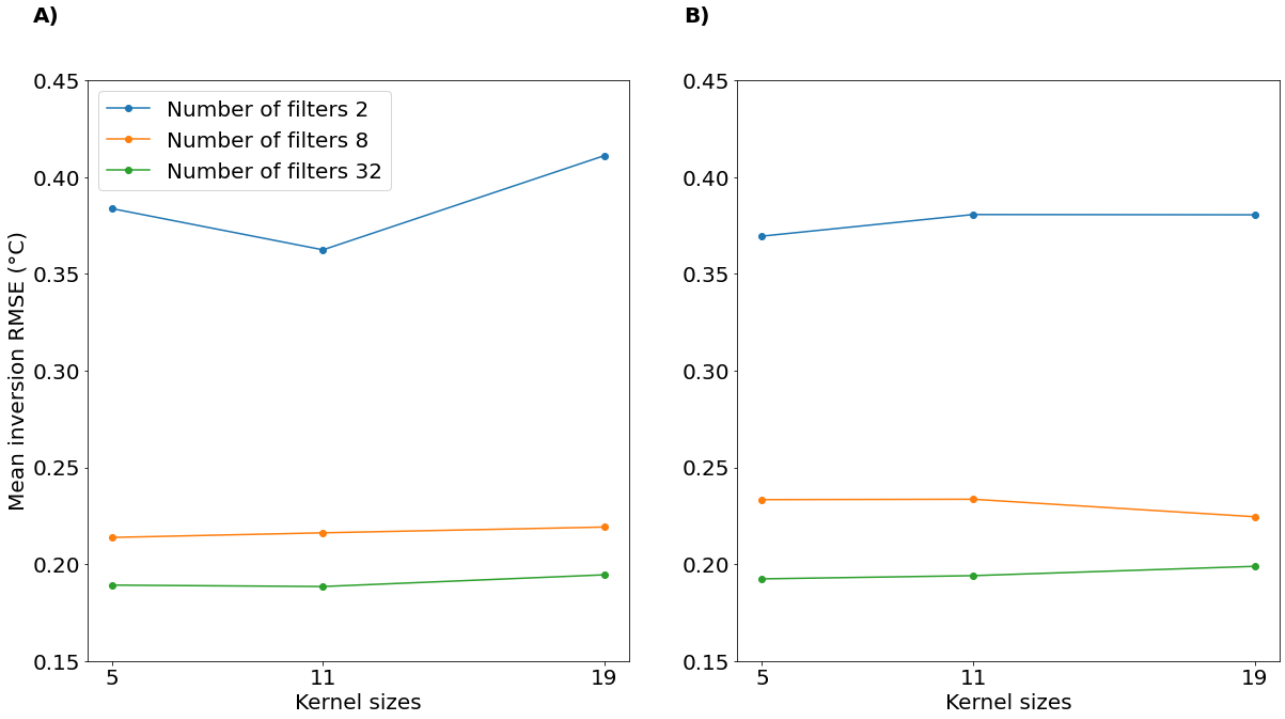

Figure S1. A) Mean cross-validation RMSE (in °C) for differents kernel sizes and number of filters while using three hidden layers. B) same as A) but with 5 hidden layers.

Table S1. Mean cross-validation RMSE (in °C) of the backward optimization for different values of A and B, while C is fixed to 0.1.

	A=0.01	A=0.05	A=0.1
B=0	0.205	0.190	0.189
B=0.01	0.199	0.189	0.190
B=0.1	0.191	0.191	0.192

Table S2. Mean cross-validation RMSE (in °C) of the backward optimization for different values of B and C, while A is fixed to 0.05°C.

	C=0	C=0.01	C=0.1
B=0	0.188	0.187	0.188
B=0.01	0.190	0.188	0.189
B=0.1	0.191	0.191	0.191

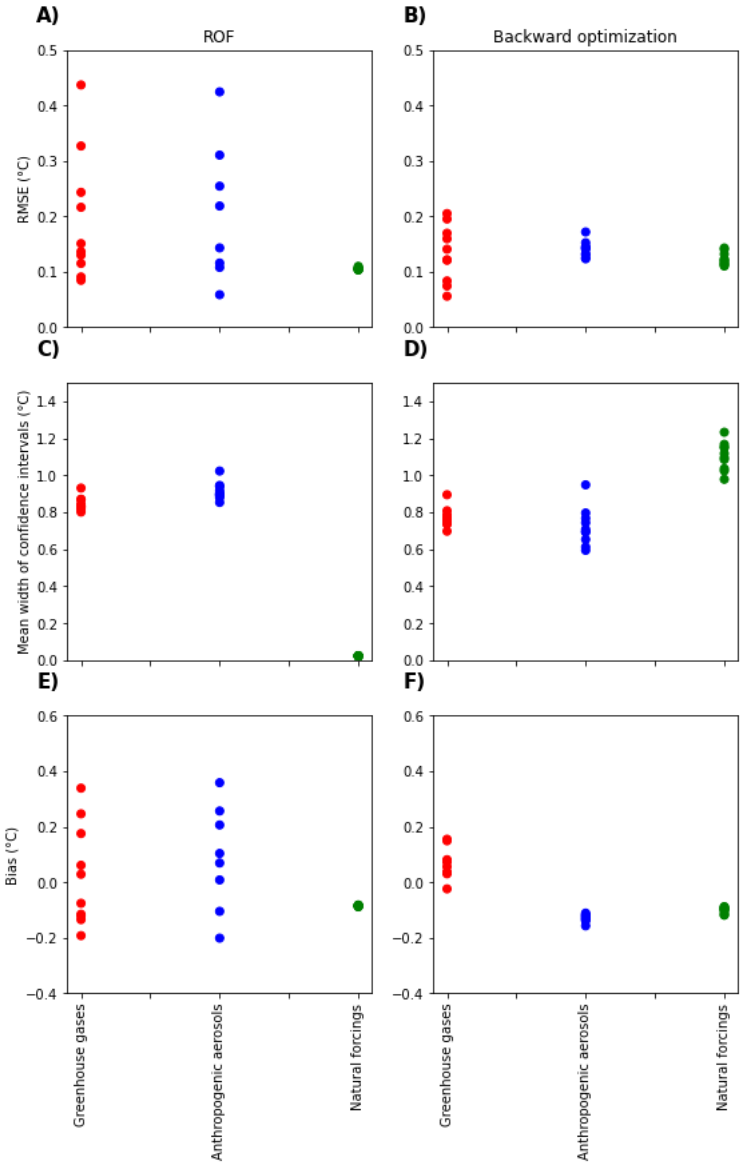

Figure S2. Performances of attribution methodologies on the 10 removed IPSL-CM6-LR members A) RMSE distribution when using ROF and all 10 removed members as pseudo-observation for the attributable GSAT anomaly of (red) greenhouse gases, (blue) anthropogenic aerosols, (green) natural forcing. B) Same as A) for the backward optimization. C) Distribution of the widths of the 90 % percent confidence intervals in 2000-2014 when using ROF. D) same as C) but for backward optimization E) Distribution of the time mean differences between the estimated and ensemble mean GSAT attribuable to the forcings when using ROF. F) Same as E) for backward optimization.

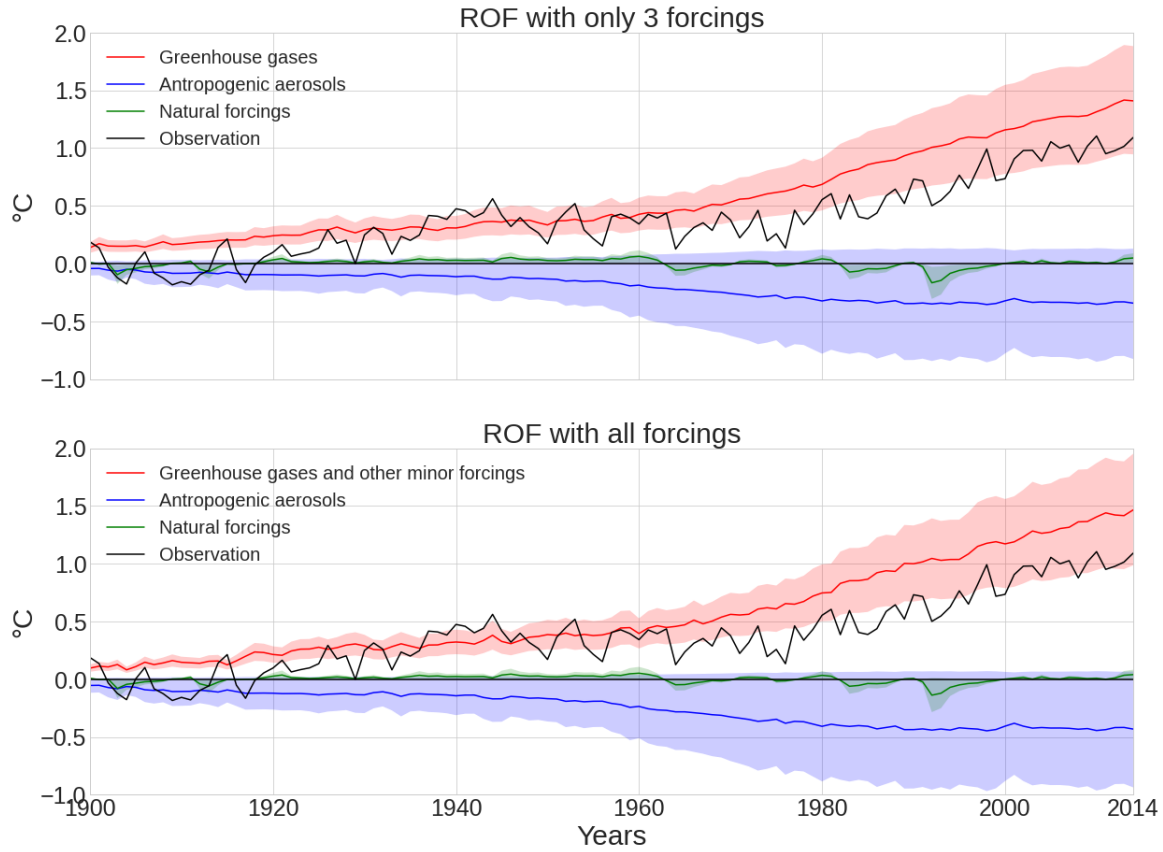

Figure S3. Attributable GSAT anomalies calculated from ROF with observations when using anthropogenic aerosols, natural forcing and greenhouse gases as forcings (top) anthropogenic aerosols, natural forcing and greenhouse gases and other anthropogenic effect combined (bottom).

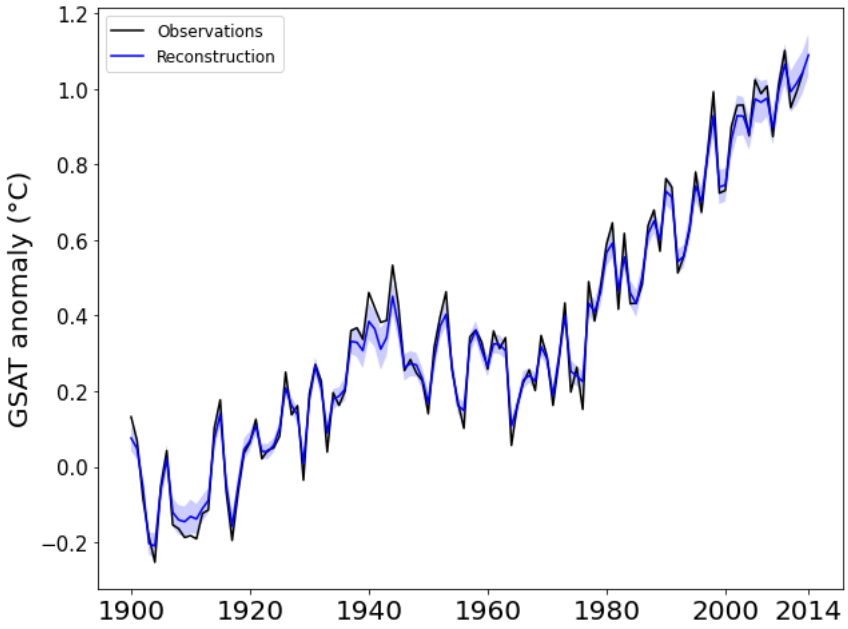

Figure S4. (Black) Observed GSAT anomalies, in °C, and (blue) the mean reconstruction of the observation by the CNN. Color shade shows the 90% percents confidence intervals of the mean reconstruction obtained across the 1200 backward optimization results available.

Chapitre 5

Conclusion et perspectives

5.1 Conclusion

La détection et l'attribution du changement climatique est un problème de grande importance pour la compréhension du fonctionnement du système climatique et l'élaboration de politiques efficaces d'adaptation au changement climatique (Eyring et al., 2021). Cependant, de grandes difficultés existent pour résoudre ce problème. La première est la période d'observation du climat qui est relativement limité : pour la température de surface, nous n'avons des observations fiables qu'à partir de 1850. De plus, ces observations ne sont pas homogènes sur le globe. Cela représente un problème, car l'effet de certains modes de variabilité interne peuvent être de très basse fréquence avec des échelles de temps centennal ou multicentennal (Jiang et al., 2021; Li and Huang, 2022). Pour les forçages, et la variabilité forcée associée, les observations sont également de courte durée. Les forçages anthropiques de gaz à effet de serre n'ont en effet, commencé à avoir un effet marqué que dans la seconde moitié du vingtième siècle.

Un second verrou est l'incertitude structurelle des modèles climatiques. C'est particulièrement le cas concernant leur sensibilité climatique d'équilibre (Sherwood et al., 2020) dû à des phénomènes très incertains comme la rétroaction nuageuse (Zelinka et al., 2016). De plus, des incertitudes signifiantes existent dans les estimations des forçages externes, en particulier pour le forçage

radiatif relié aux aérosols (Menary et al., 2020; Smith and Forster, 2021). La variabilité interne varie également entre les modèles climatiques.

Un dernier verrou repose sur les hypothèses faites dans certaines méthodes d'attribution. En effet, une hypothèse commune des méthodologies d'attribution est l'additivité des effets des différents forçages externes. Cette hypothèse peut être validée à l'échelle du globe pour la température (Marvel et al., 2015; Shiogama et al., 2013) mais se révèle invalide pour d'autres variables physiques comme les précipitations (Marvel et al., 2015) ou la glace de mer (Kim et al., 2023). Elle se retrouve également invalide pour des régions extra-tropical comme l'Arctique (Deng et al., 2020) ou l'hémisphère Sud (Pope et al., 2020).

Nous avons, lors de cette thèse, abordé deux problèmes liés à la détection et attribution du changement climatique que sont la séparation de la variabilité interne et forcé et l'attribution du changement climatique à trois groupes de forçages : les gaz à effet de serre, les forçages anthropiques et les forçages naturels. Nous avons développé pour chacun de ces problèmes une méthode statistique se reposant sur des réseaux de neurones. La méthode de séparation de la variabilité interne et forcé est appelé *Noise to Noise* alors que celle d'attribution est appelée optimisation inverse. Cette dernière a la particularité de prendre en compte les non-additivités dans les effets des forçages. Ces deux méthodes ont utilisé des sorties de simulations de modèles climatiques et des observations et nous les avons toutes deux appliqués à la température de surface de l'air. Cependant, alors que la méthode *Noise to Noise* utilise des données tridimensionnelles (temps, latitude, longitude), la méthode d'optimisation inverse utilise des données globales, et donc monodimensionnelle.

Les résultats obtenus pour ces deux méthodes sont comparés à celles d'autres méthodes comme les empreintes optimisées pour l'optimisation inverse ou

les moyennes d'ensemble pour *Noise to Noise* et sont assez similaires. En effet, la méthode d'optimisation inverse a donné des résultats similaires à une méthode d'empreintes optimisés régularisés comme celle utilisée dans Gillett et al., 2021 en utilisant les observations ou des pseudo-observations issues de modèles climatiques à l'exception de leurs intervalles de confiance. Cependant, l'optimisation inverse obtient de meilleurs résultats que les empreintes optimisées sur un jeu de données synthétique exhibant une forte non-additivité.

Noise to Noise montre globalement une réduction de la variabilité d'un facteur quatre bien qu'existent de fortes disparités régionales : ENSO est très facilement éliminé du signal forcé, là où la méthode rencontre plus de difficultés dans d'autres régions comme les marges de la banquise ou sur les continents sur lesquels les changements de température de surface sont plus intenses.

Ces deux nouvelles méthodes ont toutefois des limites. Tout d'abord, une évaluation plus rigoureuse de ces méthodes est toujours manquante. Cela est le cas en particulier pour la méthode *Noise to Noise* qui a été comparé à la méthode des moyennes d'ensemble que sur deux modèles climatiques. Une comparaison plus exhaustive de cette méthode aux autres méthodes de séparation de la variabilité interne et forcé a cependant été commencé lors d'un hackathon nommé ForceSMIP s'étant déroulé en Suisse.

Une seconde limite de ce travail concernant la méthode d'optimisation inverse est l'absence de plus-value claire par rapport aux autres méthodes d'attribution : la méthode d'optimisation inverse n'a été testé que dans un cas additif, celui de la GSAT ou sur des données synthétiques.

Une dernière limite de ce travail est le manque de maîtrise de tous les paramètres des méthodes proposé. En effet, pour les deux méthodes, un grand nombre d'architectures de réseaux ou d'hyper-paramètres d'entraînement auraient pu être testés. La nature même du *machine learning* rend ce processus très long de par le très grand nombre d'architectures existantes. Nous avons

choisi dans cette thèse, pour des raisons de clareté, de reprendre des outils et des méthodes connus et relativement simples de l'IA et de les appliquer à ces problèmes climatiques. L'utilisation de réseaux de neurones ou de méthodologies plus de pointe pourrait grandement changer les résultats.

Néanmoins, le travail accompli illustre le fait que le *machine learning* est un outil plein de potentiel pour la climatologie, comme d'autres travaux le laissait déjà présager (Ham et al., 2019; Gagne II et al., 2019; Labe and Barnes, 2021). Les méthodes proposées (l'optimisation inverse et le *Noise to noise*) sont toutes deux des premières applications du *machine learning* à leurs problèmes respectifs et peuvent facilement être poursuivies et perfectionnées. Elles apportent une nouvelle confirmation des résultats des méthodes de référence tout en permettant d'aborder les verrous scientifiques entourant la question de l'attribution du changement climatique et de la séparation de la variabilité interne et forcée. En effet, la méthode d'optimisation inverse ne fait pas l'hypothèse d'additivité de l'effet des forçages. De plus, le *machine learning* permet de se servir et d'exploiter au maximum les données à disposition, qu'elles soient issues de modèles climatiques ou d'observations. La jointure de ces deux disciplines, la climatologie et le *machine learning*, fut également une constante du contexte de la thèse. Comme doctorant, j'étais en effet complètement novice en matière de climatologie et spécialiste d'IA, là où mon encadrement direct était au contraire spécialiste de climatologie et novice en matière d'IA.

5.2 Perspectives

Les méthodes d'optimisation inverse et de *Noise to Noise* pourraient aisément être développées et perfectionnées.

La méthode de séparation de la variabilité interne et forcé *Noise to Noise* pourrait être amélioré avec des techniques de régularisation ou en testant d'autres types d'architectures. Des évaluations plus systématiques comme une approche en modèle parfait pourraient être employés pour évaluer cette méthode. Un travail de comparaison des performances de la méthode *Noise to Noise* avec d'autres méthodes de pointe de la littérature a été commencé lors d'un hackathon nommé "ForceSIMP" se déroulant en Suisse. La méthode peut s'appliquer à d'autres variables physiques comme la précipitation ou la pression à la surface de la mer. Cependant, l'effet de la variabilité forcé est plus difficile à discerner de la variabilité interne dans ces variables comparés à la température de surface (Deser et al., 2012; Deser et al., 2014). Les performances obtenues devraient donc être bien différentes de celles obtenues avec la température de l'air. L'utilisation de différentes variables physiques pourrait se faire de manière conjointe dans l'entraînement du réseau de neurones afin d'exploiter les covariances temporelles et spatiales qui peuvent exister entre les différents champs physiques.

Une autre perspective pour ce travail est d'utiliser cette méthode non pas sur les observations, mais pour d'autres types de simulations comme les simulations DAMIP. En effet, ces simulations sont utilisées pour le problème de la détection et attribution du changement climatique et le traitement de la variabilité interne est un défi majeur de ce problème. Filtrer la variabilité interne de ces simulations représenterait une étape préalable facilitant toute étude de détection et attribution. Cependant, les simulations de DAMIP ont été uniquement réalisés par certains modèles CMIP6 et avec des tailles d'ensembles bien inférieurs à ceux disponibles pour les simulations historiques. Entraîner un réseau de neurones sur une base de données aussi restreinte pourrait se révéler problématique. Une méthode

pour contourner cette difficulté pourrait être d'utiliser comme point de départ de l'entraînement de l'U-Net le réseau déjà entraîné avec les simulations historiques. Cette approche dite d'« apprentissage transféré » permet d'entrainer un réseau de neurones de taille importante en disposant uniquement d'une base de données limitée.

La méthode créée pour la détection et attribution du changement climatique, l'optimisation inverse, a comme perspective d'être utilisé sur d'autres variables physiques et à l'échelle locale afin de mieux exploiter sa capacité à gérer les non-additivités. En effet, il a été montré que l'hypothèse d'additivité utilisée dans plusieurs techniques de détection et attribution comme les empreintes optimisées est une importante limitation quand on s'intéresse aux précipitations (Marvel et al., 2015) à la glace de mer (Deng et al., 2020) ou à l'échelle régionale (Pope et al., 2020). Pour l'échelle régionale, plusieurs pistes de travail sont possibles. Une première méthode directe serait de faire une moyenne régionale de la variable physique utilisé. L'étude serait alors assez semblable à celle déjà effectuée avec un CNN monodimensionnelle. Un travail dans ce sens a été commencé lors d'un stage de master. Une seconde option serait d'étudier toute la variabilité spatiale conjointement à la variabilité temporelle contenue dans les observations et les sorties des modèles. Cette approche nécessite de changer l'architecture utilisée avec, par exemple, un U-Net tridimensionnel. La méthode d'optimisation inverse peut également être amélioré en étudiant mieux l'impact des hyper-paramètres, pas assez exhaustivement testé. Une meilleure prise en compte de l'effet de la variabilité interne dans les résultats et la méthode pourrait également être nécessaire pour affiner les résultats.

Les deux travaux effectués lors de la thèse sont liés et n'ont donc pas forcément pour vocation à rester séparés l'un de l'autre. Le problème de la

détection et attribution du changement climatique et celui de la séparation de la variabilité sont des problèmes très proches. Comme se fut argumenté plus haut, une méthode de filtrage de la variabilité interne peut être utilisée comme étape préalable aux méthodes de détection et attribution. Une perspective de plus long terme serait de créer une méthode statistique capable de résoudre ces deux problèmes de manière conjointe. Le but de cette méthode serait de réussir à caractériser dans les observations de n'importe quel variable physique et à toute échelle spatiale le rôle exact de la variabilité interne, forcé et de chaque forçage. Pour ce faire, une possibilité serait de reprendre l'idée développée plus haut d'apprentissage transféré appliqué au réseau *Noise to Noise*. Une fois le réseau entraîné à retrouver le signal forcé dans les observations, il devrait pouvoir être possible par apprentissage transféré de le faire efficacement apprendre à retrouver un signal forcé d'un forçage particulier depuis les observations directement. La méthode pourrait être répliquée pour chaque forçage en faisant un apprentissage transféré pour chacun d'entre eux avec le même réseau de départ entraîné sur des simulations historiques. Cette méthode, si elle fonctionnait, permettrait de retirer la variabilité interne et de caractériser la variabilité forcée de chaque forçages sans réduction spatiale ou temporelle.

Bibliographie

Abramowitz, Gab et al. (2019). “ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing”. In: *Earth System Dynamics* 10.1, pp. 91–105.

Allan, Richard P et al. (2020). “Advances in understanding large-scale responses of the water cycle to climate change”. In: *Annals of the New York Academy of Sciences* 1472.1, pp. 49–75.

Allen, Myles R and Peter A Stott (2003). “Estimating signal amplitudes in optimal fingerprinting, Part I: Theory”. In: *Climate Dynamics* 21, pp. 477–491.

Allen, Myles R and Simon FB Tett (1999). “Checking for model consistency in optimal fingerprinting”. In: *Climate Dynamics* 15, pp. 419–434.

André, Jean-Claude et al. (2014). “High-Performance Computing for Climate Modeling”. In: *Bulletin of the American Meteorological Society* 95.5, ES97–ES100.

Bach, Sebastian et al. (2015). “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”. In: *PloS one* 10.7, e0130140.

Balaji, Venkatramani et al. (2017). “CPMIP: measurements of real computational performance of Earth system models in CMIP6”. In: *Geoscientific Model Development* 10.1, pp. 19–34.

Banerjee, Antara et al. (2020). “A pause in Southern Hemisphere circulation trends due to the Montreal Protocol”. In: *Nature* 579.7800, pp. 544–548.

Barnes, Elizabeth A et al. (2019). "Viewing forced climate patterns through an AI lens". In: *Geophysical Research Letters* 46.22, pp. 13389–13398.

Bellard, Céline et al. (2012). "Impacts of climate change on the future of biodiversity". In: *Ecology letters* 15.4, pp. 365–377.

Bergen, Karianne J et al. (2019). "Machine learning for data-driven discovery in solid Earth geoscience". In: *Science* 363.6433, eaau0323.

Bishop, Christopher M (1994). "Novelty detection and neural network validation". In: *IEE Proceedings-Vision, Image and Signal processing* 141.4, pp. 217–222.

Boé, Julien (2018). "Interdependency in multimodel climate projections: Component replication and result similarity". In: *Geophysical Research Letters* 45.6, pp. 2771–2779.

Boer, George J et al. (2016). "The decadal climate prediction project (DCPP) contribution to CMIP6". In: *Geoscientific Model Development* 9.10, pp. 3751–3777.

Bonnet, Remy et al. (2020). "Underestimated Global Warming Hidden by Internal AMOC Weakening". In: *AGU Fall Meeting Abstracts*. Vol. 2020, GC117–0017.

Boukabara, Sid-Ahmed et al. (2019). "Leveraging modern artificial intelligence for remote sensing and NWP: Benefits and challenges". In: *Bulletin of the American Meteorological Society* 100.12, ES473–ES491.

Brajard, Julien et al. (2012). "Atmospheric correction of MERIS data for case-2 waters using a neuro-variational inversion". In: *Remote Sensing of Environment* 126, pp. 51–61.

Cassou, Christophe et al. (2018). "Decadal climate variability and predictability: Challenges and opportunities". In: *Bulletin of the American Meteorological Society* 99.3, pp. 479–490.

Chen D., M. Rojas B.H. Samset K. Cobb A. Diongue Niang P. Edwards S. Emori S.H. Faria E. Hawkins P. Hope P. Huybrechts M. Meinshausen S.K.

Mustafa G.-K. Plattner and A.-M. Tréguier (2021). "Framing, Context, and Methods. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change". In: *Cambridge University Press*.

Choudhary, Kamal et al. (2022). "Recent advances and applications of deep learning methods in materials science". In: *npj Computational Materials* 8.1, p. 59.

Cowtan, Kevin and Robert G Way (2014). "Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends". In: *Quarterly Journal of the Royal Meteorological Society* 140.683, pp. 1935–1944.

Dai, Aiguo and Christine E Bloecker (2019). "Impacts of internal variability on temperature and precipitation trends in large ensemble simulations by two climate models". In: *Climate dynamics* 52.1-2, pp. 289–306.

DelSole, Timothy et al. (2011). "A significant component of unforced multidecadal variability in the recent acceleration of global warming". In: *Journal of Climate* 24.3, pp. 909–926.

DelSole, Timothy et al. (2019). "Confidence intervals in optimal fingerprinting". In: *Climate Dynamics* 52.7, pp. 4111–4126.

Deng, Jiechun et al. (2020). "Nonlinear climate responses to increasing CO2 and anthropogenic aerosols simulated by CESM1". In: *Journal of Climate* 33.1, pp. 281–301.

Deser, Clara et al. (2012). "Uncertainty in climate change projections: the role of internal variability". In: *Climate dynamics* 38, pp. 527–546.

Deser, Clara et al. (2014). "Projecting North American climate over the next 50 years: Uncertainty due to internal variability". In: *Journal of Climate* 27.6, pp. 2271–2296.

Deser, Clara et al. (2016). "Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications". In: *Journal of Climate* 29.6, pp. 2237–2258.

Edwards, Paul N (2011). "History of climate modeling". In: *Wiley Interdisciplinary Reviews: Climate Change* 2.1, pp. 128–139.

Egmont-Petersen, Michael et al. (2002). "Image processing with neural networks—a review". In: *Pattern recognition* 35.10, pp. 2279–2301.

Enfield, David B et al. (2001). "The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US". In: *Geophysical Research Letters* 28.10, pp. 2077–2080.

Eyring, V. et al. (2021). "Human Influence on the Climate System. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change". In: *Cambridge University Pres*.

Eyring, Veronika et al. (2016). "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". In: *Geoscientific Model Development* 9.5, pp. 1937–1958.

Fan, Feng-Lei et al. (2021). "On interpretability of artificial neural networks: A survey". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.6, pp. 741–760.

Feldl, Nicole et al. (2020). "Sea ice and atmospheric circulation shape the high-latitude lapse rate feedback". In: *NPJ climate and atmospheric science* 3.1, p. 41.

Frajka-Williams, Eleanor et al. (2017). "Emerging negative Atlantic Multidecadal Oscillation index in spite of warm subtropics". In: *Scientific Reports* 7.1, p. 11224.

Frame, David J et al. (2020). "Climate change attribution and the economic costs of extreme weather events: a study on damages from extreme rainfall and drought". In: *Climatic Change* 162, pp. 781–797.

Frankcombe, Leela M et al. (2015). "Separating internal variability from the externally forced climate response". In: *Journal of Climate* 28.20, pp. 8184–8202.

Frankignoul, Claude et al. (2017). "Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic multidecadal oscillation and the Pacific decadal oscillation". In: *Journal of Climate* 30.24, pp. 9871–9895.

Gagne II, David John et al. (2019). "Interpretable deep learning for spatial analysis of severe hailstorms". In: *Monthly Weather Review* 147.8, pp. 2827–2845.

Gillett, Nathan P et al. (2016). "The detection and attribution model intercomparison project (DAMIP v1. 0) contribution to CMIP6". In: *Geoscientific Model Development* 9.10, pp. 3685–3697.

Gillett, Nathan P et al. (2021). "Constraining human contributions to observed warming since the pre-industrial period". In: *Nature Climate Change* 11.3, pp. 207–212.

Good, Peter et al. (2015). "Nonlinear regional warming with increasing CO2 concentrations". In: *Nature Climate Change* 5.2, pp. 138–142.

Goodfellow, Ian et al. (2016). *Deep learning*. MIT press.

Goosse, Hugues et al. (2018). "Quantifying climate feedbacks in polar regions". In: *Nature communications* 9.1, p. 1919.

Gulev, SK et al. (2021). "Changing state of the climate system. In climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change". In.

Ham, Yoo-Geun et al. (2019). "Deep learning for multi-year ENSO forecasts". In: *Nature* 573.7775, pp. 568–572.

Hannart, Alexis and Philippe Naveau (2018). "Probabilities of causation of climate changes". In: *Journal of Climate* 31.14, pp. 5507–5524.

Hannart, Alexis et al. (2014). "Optimal fingerprinting under multiple sources of uncertainty". In: *Geophysical Research Letters* 41.4, pp. 1261–1268.

Hannart, Alexis et al. (2016). “Causal counterfactual theory for the attribution of weather and climate-related events”. In: *Bulletin of the American Meteorological Society* 97.1, pp. 99–110.

Hansen, Gerrit and Dáithí Stone (2016). “Assessing the observed impact of anthropogenic climate change”. In: *Nature Climate Change* 6.5, pp. 532–537.

Hansen, James et al. (2005). “Efficacy of climate forcings”. In: *Journal of geophysical research: atmospheres* 110.D18.

Hansen, James et al. (2006). “Global temperature change”. In: *Proceedings of the National Academy of Sciences* 103.39, pp. 14288–14293.

Harzallah, Ali and Robert Sadourny (1995). “Internal versus SST-forced atmospheric variability as simulated by an atmospheric general circulation model”. In: *Journal of Climate* 8.3, pp. 474–495.

Hasselmann, Klaus (1993). “Optimal fingerprints for the detection of time-dependent climate change”. In: *Journal of Climate* 6.10, pp. 1957–1971.

Hawkins, Ed and Rowan Sutton (2009). “The potential to narrow uncertainty in regional climate predictions”. In: *Bulletin of the American Meteorological Society* 90.8, pp. 1095–1108.

Hirabayashi, Yukiko et al. (2013). “Global flood risk under climate change”. In: *Nature climate change* 3.9, pp. 816–821.

Holland, Marika M and Cecilia M Bitz (2003). “Polar amplification of climate change in coupled models”. In: *Climate dynamics* 21.3-4, pp. 221–232.

Hourdin, Frédéric et al. (2017). “The art and science of climate model tuning”. In: *Bulletin of the American Meteorological Society* 98.3, pp. 589–602.

Huggel, Christian et al. (2015). “Potential and limitations of the attribution of climate change impacts for informing loss and damage discussions and policies”. In: *Climatic Change* 133, pp. 453–467.

Huntingford, Chris et al. (2006). "Incorporating model uncertainty into attribution of observed temperature change". In: *Geophysical Research Letters* 33.5.

Ilesanmi, Ademola E and Taiwo O Ilesanmi (2021). "Methods for image denoising using convolutional neural network: a review". In: *Complex & Intelligent Systems* 7.5, pp. 2179–2198.

Imbers, Jara et al. (2014). "Sensitivity of climate change detection and attribution to the characterization of internal climate variability". In: *Journal of Climate* 27.10, pp. 3477–3491.

Jia, Liwei and Timothy DelSole (2012). "Multi-year predictability of temperature and precipitation in multiple climate models". In: *Geophysical research letters* 39.17.

Jiang, Weimin et al. (2021). "Multicentennial variability driven by salinity exchanges between the Atlantic and the Arctic Ocean in a coupled climate model". In: *Journal of Advances in Modeling Earth Systems* 13.3, e2020MS002366.

Kadow, Christopher et al. (2020). "Artificial intelligence reconstructs missing climate information". In: *Nature Geoscience* 13.6, pp. 408–413.

Katzfuss, Matthias et al. (2017). "A Bayesian hierarchical model for climate change detection and attribution". In: *Geophysical Research Letters* 44.11, pp. 5720–5728.

Keil, Paul et al. (2020). "Multiple drivers of the North Atlantic warming hole". In: *Nature Climate Change* 10.7, pp. 667–671.

Kim, Yeon-Hee et al. (2023). "Observationally-constrained projections of an ice-free Arctic even under a low emission scenario". In: *Nature Communications* 14.1, p. 3139.

Kirchmeier-Young, Megan C et al. (2017). "Attribution of extreme events in Arctic sea ice extent". In: *Journal of Climate* 30.2, pp. 553–571.

Knutti, Reto et al. (2010). "Challenges in combining projections from multiple climate models". In: *Journal of Climate* 23.10, pp. 2739–2758.

Knutti, Reto et al. (2013). "Climate model genealogy: Generation CMIP5 and how we got there". In: *Geophysical Research Letters* 40.6, pp. 1194–1199.

Labe, Zachary M and Elizabeth A Barnes (2021). "Detecting climate signals using explainable AI with single-forcing large ensembles". In: *Journal of Advances in Modeling Earth Systems* 13.6, e2021MS002464.

LeCun, Yann et al. (1989). "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4, pp. 541–551.

Lehner, F. et al. (2020). "Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6". In: *Earth System Dynamics* 11.2, pp. 491–508. DOI: 10.5194/esd-11-491-2020. URL: https://esd.copernicus.org/articles/11/491/2020/.

Lehtinen, Jaakko et al. (2018). "Noise2Noise: Learning image restoration without clean data". In: *arXiv preprint arXiv:1803.04189*.

Li, Baosheng et al. (Apr. 2018). "Asymmetric Response of Predictability of East Asian Summer Monsoon to ENSO". In: *SOLA* 14, pp. 52–56. DOI: 10.2151/sola.2018-009.

Li, Shufan and Ping Huang (2022). "An exponential-interval sampling method for evaluating equilibrium climate sensitivity via reducing internal variability noise". In: *Geoscience Letters* 9.1, pp. 1–10.

Luber, George and Michael McGeehin (2008). "Climate change and extreme heat events". In: *American journal of preventive medicine* 35.5, pp. 429–435.

Maher, Nicola et al. (2021). "Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble". In: *Earth System Dynamics* 12, pp. 401–418.

Malde, Ketil et al. (2020). "Machine intelligence and the data-driven future of marine science". In: *ICES Journal of Marine Science* 77.4, pp. 1274–1285.

Marjanac, Sophie et al. (2017). "Acts of God, human influence and litigation". In: *Nature Geoscience* 10.9, pp. 616–619.

Marvel, K. et al. (2015). "Do responses to different anthropogenic forcings add linearly in climate models?" In: *Environ. Res. Lett.* 10.10, p. 104010. DOI: 10.1088/1748-9326/10/10/104010.

Masson-Delmotte et al. (2021a). "Annex IV: Modes of Variability. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change". In: *Cambridge University Press*.

Masson-Delmotte, V. et al. (2021b). "IPCC, 2021: Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change". In: *Cambridge University Press*.

Meehl, Gerald A et al. (2000). "The coupled model intercomparison project (CMIP)". In: *Bulletin of the American Meteorological Society* 81.2, pp. 313–318.

Meehl, Gerald A et al. (2007). "The WCRP CMIP3 multimodel dataset: A new era in climate change research". In: *Bulletin of the American meteorological society* 88.9, pp. 1383–1394.

Meehl, Gerald A et al. (2016). "Antarctic sea-ice expansion between 2000 and 2014 driven by tropical Pacific decadal climate variability". In: *Nature Geoscience* 9.8, pp. 590–595.

Menary, Matthew B et al. (2020). "Aerosol-forced AMOC changes in CMIP6 historical simulations". In: *Geophysical Research Letters* 47.14, e2020GL088166.

Mimura, Nobuo (2013). "Sea-level rise caused by climate change and its implications for society". In: *Proceedings of the Japan Academy, Series B* 89.7, pp. 281–301.

Monerie, Paul-Arthur et al. (2019). "Effect of the Atlantic multidecadal variability on the global monsoon". In: *Geophysical Research Letters* 46.3, pp. 1765–1775.

Montavon, Grégoire et al. (2017). "Explaining nonlinear classification decisions with deep taylor decomposition". In: *Pattern recognition* 65, pp. 211–222.

Mukherjee, Sourav et al. (2018). "Climate change and drought: a perspective on drought indices". In: *Current climate change reports* 4, pp. 145–163.

Nauels, Alexander et al. (2019). "Attributing long-term sea-level rise to Paris Agreement emission pledges". In: *Proceedings of the National Academy of Sciences* 116.47, pp. 23487–23492.

Nebeker, Frederik (1995). *Calculating the weather: Meteorology in the 20th century*. Elsevier.

Neelin, J David et al. (1998). "ENSO theory". In: *Journal of Geophysical Research: Oceans* 103.C7, pp. 14261–14290.

Notz, Dirk (2015). "How well must climate models agree with observations?" In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2052, p. 20140164.

O'Neill, Brian C et al. (2014). "A new scenario framework for climate change research: the concept of shared socioeconomic pathways". In: *Climatic change* 122, pp. 387–400.

Paeth, Heiko et al. (2017). "Quantifying the evidence of climate change in the light of uncertainty exemplified by the Mediterranean hot spot region". In: *Global and Planetary Change* 151, pp. 144–151.

Pan, Yi Hong and Abraham H Oort (1983). "Global climate variations connected with sea surface temperature anomalies in the eastern equatorial Pacific Ocean for the 1958–73 period". In: *Monthly Weather Review* 111.6, pp. 1244–1258.

Parker, Wendy S (2009). "II—Confirmation and adequacy-for-purpose in climate modelling". In: *Aristotelian Society Supplementary Volume*. Vol. 83. 1. Wiley Online Library, pp. 233–249.

Pearl, Judea (2009). "Causal inference in statistics: An overview". In.

Philander, SG (1990). *El Niño, La Niña, and the southern oscillation, 293 pp*.

Pithan, Felix and Thorsten Mauritsen (2014). "Arctic amplification dominated by temperature feedbacks in contemporary climate models". In: *Nature geoscience* 7.3, pp. 181–184.

Pope, James O. et al. (2020). "Non-additive response of the high-latitude Southern Hemisphere climate to aerosol forcing in a climate model with interactive chemistry". In: *Atmospheric Science Letters* 21.12, e1004. DOI: https://doi.org/10.1002/asl.1004.

Reichstein, Markus et al. (2019). "Deep learning and process understanding for data-driven Earth system science". In: *Nature* 566.7743, pp. 195–204.

Ribes, Aurélien et al. (2009). "Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate". In: *Climate Dynamics* 33, pp. 707–722.

Ribes, Aurélien et al. (2013). "Application of regularised optimal fingerprinting to attribution. Part I: method, properties and idealised analysis". In: *Climate dynamics* 41.11, pp. 2817–2836.

Ribes, Aurélien et al. (2021). "Making climate projections conditional on historical observations". In: *Science Advances* 7.4, eabc0671.

Roe, Gerard H and Michael A O'Neal (2009). "The response of glaciers to intrinsic climate variability: observations and models of late-Holocene variations in the Pacific Northwest". In: *Journal of Glaciology* 55.193, pp. 839–854.

Ronneberger, Olaf et al. (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, pp. 234–241.

Rounsevell, Mark DA and Marc J Metzger (2010). "Developing qualitative scenario storylines for environmental change assessment". In: *Wiley interdisciplinary reviews: climate change* 1.4, pp. 606–619.

Schneider, Tapio and Isaac M Held (2001). "Discriminants of twentieth-century changes in Earth surface temperatures". In: *Journal of Climate* 14.3, pp. 249–254.

Schurer, Andrew et al. (2018). "Estimating the transient climate response from observed warming". In: *Journal of Climate* 31.20, pp. 8645–8663.

Sherwood, SC et al. (2020). "An assessment of Earth's climate sensitivity using multiple lines of evidence". In: *Reviews of Geophysics* 58.4, e2019RG000678.

Shin, Sang-Ik and Prashant D Sardeshmukh (2011). "Critical influence of the pattern of tropical ocean warming on remote climate trends". In: *Climate Dynamics* 36, pp. 1577–1591.

Shiogama, Hideo et al. (2013). "On the linear additivity of climate forcing-response relationships at global and continental scales". In: *International Journal of Climatology* 33.11, pp. 2542–2550. DOI: https://doi.org/10.1002/joc.3607. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.3607. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3607.

Shrikumar, Avanti et al. (2017). "Learning important features through propagating activation differences". In: *International conference on machine learning*. PMLR, pp. 3145–3153.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Sippel, Sebastian et al. (2019). "Uncovering the forced climate response from a single ensemble member using statistical learning". In: *Journal of Climate* 32.17, pp. 5677–5699.

Smith, Christopher J and Piers M Forster (2021). "Suppressed late-20th century warming in CMIP6 models explained by forcing and feedbacks". In: *Geophysical Research Letters* 48.19, e2021GL094948.

Smith, Doug M et al. (2012). "What is the current state of scientific knowledge with regard to seasonal and decadal forecasting?" In: *Environmental Research Letters* 7.1, p. 015602.

Smoliak, Brian V et al. (2015). "Dynamical adjustment of the Northern Hemisphere surface air temperature field: Methodology and application to observations". In: *Journal of Climate* 28.4, pp. 1613–1629.

Solomon, Amy et al. (2011). "Distinguishing the roles of natural and anthropogenically forced decadal climate variability: implications for prediction". In: *Bulletin of the American Meteorological Society* 92.2, pp. 141–156.

Springenberg, Jost Tobias et al. (2014). "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806*.

Srivastava, Abhishekh K and Timothy DelSole (2014). "Robust forced response in South Asian summer monsoon in a future climate". In: *Journal of Climate* 27.20, pp. 7849–7860.

Staniforth, Andrew and John Thuburn (2012). "Horizontal grids for global weather and climate prediction models: a review". In: *Quarterly Journal of the Royal Meteorological Society* 138.662, pp. 1–26.

Steinman, Byron A et al. (2015). "Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures". In: *Science* 347.6225, pp. 988–991.

Sundararajan, Mukund et al. (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR, pp. 3319–3328.

Sutton, Rowan T and Daniel LR Hodson (2005). "Atlantic Ocean forcing of North American and European summer climate". In: *science* 309.5731, pp. 115–118.

Sutton, RT et al. (2018). "Atlantic multidecadal variability and the UK ACSIS program". In: *Bulletin of the American Meteorological Society* 99.2, pp. 415–425.

Tandon, Neil F and Paul J Kushner (2015). "Does external forcing interfere with the AMOC's influence on North Atlantic sea surface temperature?" In: *Journal of Climate* 28.16, pp. 6309–6323.

Taylor, Karl E et al. (2012). "An overview of CMIP5 and the experiment design". In: *Bulletin of the American meteorological Society* 93.4, pp. 485–498.

Terray, Laurent (2012). "Evidence for multiple drivers of North Atlantic multi-decadal climate variability". In: *Geophysical Research Letters* 39.19.

Tian, Chunwei et al. (2020). "Deep learning on image denoising: An overview". In: *Neural Networks* 131, pp. 251–275.

Ting, Mingfang et al. (2009). "Forced and internal twentieth-century SST trends in the North Atlantic". In: *Journal of Climate* 22.6, pp. 1469–1481.

Toms, Benjamin A et al. (2020). "Physically interpretable neural networks for the geosciences: Applications to earth system variability". In: *Journal of Advances in Modeling Earth Systems* 12.9, e2019MS002002.

Trenberth, Kevin E and Dennis J Shea (2006). "Atlantic hurricanes and natural variability in 2005". In: *Geophysical research letters* 33.12.

Trenberth, Kevin E et al. (2002). "Evolution of El Niño–Southern Oscillation and global atmospheric surface temperatures". In: *Journal of Geophysical Research: Atmospheres* 107.D8, AAC–5.

Trenberth, Kevin E et al. (2009). "Earth's global energy budget". In: *Bulletin of the American Meteorological Society* 90.3, pp. 311–324.

Trenberth, Kevin E et al. (2014). "Earth's energy imbalance". In: *Journal of Climate* 27.9, pp. 3129–3144.

Trenberth, Kevin E et al. (2015). "Attribution of climate extreme events". In: *Nature Climate Change* 5.8, pp. 725–730.

Von Schuckmann, K et al. (2016). "An imperative to monitor Earth's energy imbalance". In: *Nature Climate Change* 6.2, pp. 138–144.

Von Schuckmann, Karina et al. (2020). "Heat stored in the earth system: Where does the energy go? The GCOS earth heat inventory team". In: *Earth System Science Data Discussions* 2020, pp. 1–45.

Wallace, John M et al. (2012). "Simulated versus observed patterns of warming over the extratropical Northern Hemisphere continents during the cold season". In: *Proceedings of the National Academy of Sciences* 109.36, pp. 14337–14342.

Walsh, Kevin JE et al. (2016). "Tropical cyclones and climate change". In: *Wiley Interdisciplinary Reviews: Climate Change* 7.1, pp. 65–89.

Walther, Gian-Reto (2010). "Community and ecosystem responses to recent climate change". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1549, pp. 2019–2024.

Wang, Chunzai (2018). "A review of ENSO theories". In: *National Science Review* 5.6, pp. 813–825.

Wills, Robert C et al. (2018). "Disentangling global warming, multidecadal variability, and El Niño in Pacific temperatures". In: *Geophysical Research Letters* 45.5, pp. 2487–2496.

Wills, Robert CJ et al. (2020). "Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations". In: *Journal of Climate* 33.20, pp. 8693–8719.

Winsberg, Eric (2018). *Philosophy and climate science*. Cambridge University Press.

Wu, Zhaohua et al. (2009). "The multi-dimensional ensemble empirical mode decomposition method". In: *Advances in Adaptive Data Analysis* 1.03, pp. 339–372.

Wu, Zhaohua et al. (2011). "On the time-varying trend in global-mean surface temperature". In: *Climate dynamics* 37, pp. 759–773.

Yan, Xiaoqin et al. (2019). "A multivariate AMV index and associated discrepancies between observed and CMIP5 externally forced AMV". In: *Geophysical Research Letters* 46.8, pp. 4421–4431.

Yan, Xing et al. (2016). "A new method of satellite-based haze aerosol monitoring over the North China Plain and a comparison with MODIS Collection 6 aerosol products". In: *Atmospheric research* 171, pp. 31–40.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, pp. 818–833.

Zelinka, Mark D et al. (2016). "Insights from a refined decomposition of cloud feedbacks". In: *Geophysical Research Letters* 43.17, pp. 9259–9269.

Zhang, Rong (2017). "On the persistence and coherence of subpolar sea surface temperature and salinity anomalies associated with the Atlantic multidecadal variability". In: *Geophysical Research Letters* 44.15, pp. 7865–7875.

Zhisheng, An et al. (2015). "Global monsoon dynamics and climate change". In: *Annual review of earth and planetary sciences* 43, pp. 29–77.

Zhou, Yi-Tong and Rama Chellappa (1988). "Computation of optical flow using a neural network". In: *IEEE 1988 International Conference on Neural Networks*, 71–78 vol.2.

Zintgraf, Luisa M et al. (2017). "Visualizing deep neural network decisions: Prediction difference analysis". In: *arXiv preprint arXiv:1702.04595*.