Separation of Internal and Forced Variability of Climate Using a U-Net

Constantin Bône, Guillaume Gastineau, Sylvie Thiria, Patrick Gallinari, Carlos Mejia

JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2023MS003964

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, Paris, France, [2]UMR ISIR, Sorbonne Université, CNRS, INSERM, Paris, France, [3]Criteo AI Lab, Paris, France

Key Points:

- We present a new method to separate the forced and internal variability of the surface air temperature
- We utilize 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 utilization of very large ensemble simulations of two distinct climate models

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. Bône,
constantin.bone@sorbonne-universite.fr

Citation:

Bône, C., Gastineau, G., Thiria, S., Gallinari, P., & Mejia, C. (2024). Separation of internal and forced variability of climate using a U-Net. *Journal of Advances in Modeling Earth Systems*, *16*, e2023MS003964. https://doi.org/10.1029/2023MS003964

Received 11 AUG 2023
Accepted 6 MAY 2024

Author Contributions:

Conceptualization: Constantin Bône, Guillaume Gastineau, Sylvie Thiria, Patrick Gallinari, Carlos Mejia
Data curation: Constantin Bône
Methodology: Constantin Bône, Guillaume Gastineau, Sylvie Thiria, Patrick Gallinari, Carlos Mejia
Resources: Guillaume Gastineau
Software: Constantin Bône
Validation: Constantin Bône, Sylvie Thiria, Patrick Gallinari, Carlos Mejia
Writing – original draft: Constantin Bône

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 data set is compiled using surface air temperature data spanning from 1901 to 2020, obtained from an ensemble of Atmosphere-Ocean General Circulation Model 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, 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 capturing the changes in the North Atlantic Ocean. This methodology holds potential for extension to other physical variables, facilitating insights into the climate change triggered by external forcings over the long term.

Plain Language Summary

To anticipate future climate change, it is crucial to detect and understand the impacts of human activities. However, distinguishing the effects of anthropogenic forcing from natural climate variations in observational data is challenging. Natural climate variability, known as internal variability, arises from the chaotic nature of atmospheric and oceanic circulation, and from the interactions among the ocean, atmosphere, and land. Here, a novel approach is introduced to distinguish the changes caused by human activities from internal variability. It is applied to the surface air temperature evolution from 1901 to 2020. This method uses an artificial neural network designed to separate the internal from the human-induced variability. An unprecedented number of climate model simulations are used, enabling precise estimation of human-forced variability in these climate models. The spatio-temporal variations are distinguished by applying a well-known methodology previously used to remove noise from images. The method's performance is evaluated, revealing errors regarding the internal variability that are typically one-fourth of the actual variations. Regions with important internal variability or with low agreement among models exhibit the largest errors. Overall, the skills are comparable to other existing approaches, but improvements are anticipated.

1. Introduction

Global warming is characterized by an elevated surface air temperature, with a notable acceleration during the latter half of the twentieth century (Eyring et al., 2021). Nevertheless, the observed anomalies in surface air temperature can have different drivers. 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 shorter temporal scales and smaller spatial dimensions. For instance, the leading mode of internal variability in global air

Writing – review & editing: Constantin Bône, Guillaume Gastineau, Sylvie Thiria

surface temperature manifests as the El Niño Southern Oscillation (Wang et al., 2017), characterized by significant anomalies in the equatorial Pacific Ocean, accompanied by distant teleconnections, and a prevailing period ranging from two to 7 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. Notable examples involve 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 (England et al., 2014; Kosaka & Xie, 2013; Meehl et al., 2013) or the reduced heat uptake from the Atlantic and Southern oceans (Chen & Tung, 2014, 2018). Lastly, internal variability may also include important centennial and multi-centennial variations (Jiang et al., 2021; S. Li & Huang, 2022) with potential impacts on trends within the 1900–2015 interval (Bonnet et al., 2021; Fan et al., 2023). The distinction between forced variability and internal variability is essential for conducting detection and attribution studies, enabling accurate estimation of the climate's reaction to alterations in radiative forcing. Moreover, this differentiation aids in recognizing and understanding 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. 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).

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 is the ensemble's size (Deser et al., 2014; Frankcombe et al., 2015; Harzallah & Sadourny, 1995; Hawkins & Sutton, 2009; Solomon et al., 2011; Ting et al., 2009). As a result, modeling centers performed numerous ensemble simulations with over 20 or 30 ensemble members (Deser et al., 2020; Jeffrey et al., 2013; Rodgers et al., 2015; Sun et al., 2018). These large ensembles are commonly referred to as single-model initial-condition large ensembles (SMILE; Deser et al., 2020). This offers a valuable data set for designing 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.

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 feedback which significantly impact equilibrium climate sensitivity (ECS) (Zelinka et al., 2016). Additionally, significant uncertainties surround historical aerosol emissions and their radiative forcing (Fyfe et al., 2021; 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 use data from various 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 (Allen & Stott, 2003; Allen & Tett, 1999; Hasselmann, 1993). More recently, the use of scaling factors was also proposed by Frankcombe et al. (2015).

This paper introduces an alternative approach for distinguishing internal and forced variability using climate model data, employing a non-linear method that takes into account the spatio-temporal covariances. This method is rooted in a convolutionnal 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 residual. 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 the best of our knowledge, this represents a 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 utilized data. Section 3 introduces the method and the neural network. In Section 4, the performances are assessed and the neural network is applied to observations. Lastly, Section 5 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; GISTEMP Team, 2023; 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 $2° \times 2°$ grid. The monthly values are aggregated to calculate annual means, and the SAT anomalies are determined using 1950–2014 as a reference period. We chose this reference period to obtain data with an average value close to 0, which helps the training of neural networks.

2.2. Climate Model Simulations

The monthly SAT data is retrieved from the historical simulations of 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). In addition, we use SMILEs from MPI-ESM (Maher et al., 2019) and FGOALS-g3 (Li et al., 2020) using CMIP6 forcings, and CSIRO-Mk3-6-0 (Collier et al., 2011) and EC-Earth (Döscher et al., 2021) using CMIP5 forcings. For these simulations all external forcings are applied from 1850 to 2005 in CMIP5 or from 1850 to 2014 in CMIP6. 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. The details including the climate model names, ensemble sizes, and the names of the employed scenario simulations are provided in Tables S1, S2, and S3 in Supporting Information S1. From 2005 (2014 for CMIP6) to 2020, the outputs of the historical experiment are completed by the high greenhouse gas emission scenario Representation Concentration Pathway 8.5 (RCP8.5) scenario for CMIP5 data (Table S1 in Supporting Information S1; Van Vuuren et al., 2011) and by the intermediate Shared Socio-economic Pathway 2 4.5 (SSP2-4.5) for CMIP6 data (Table S2 in Supporting Information S1; Tebaldi et al., 2020). The SMILEs data incorporated also include extensions using forcing from scenario of CMIP5 or CMIP6 (Table S3 in Supporting Information S1). Differences are anticipated in external forcing between CMIP5 and CMIP6 simulations, with notable uncertainties arising in aerosol emissions (Fyfe et al., 2021; Smith et al., 2020). Modest differences may also emerge between the RCP8.5 (strong) and SSP2-4.5 (moderate) scenarios although these forcing were found to mirror observed forcings until 2020 to a considerable extent (Masson-Delmotte et al., 2021). The use of different scenarios regarding the extension of historical experiment by 2020 was arbitrary. The variety of forcing allows sampling of the forcing uncertainty, while the model uncertainty is sampled using a multi-model ensemble.

The members accessible for scenario simulations is less than that of historical simulations. 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 a scenario ensemble member of the same climate model.

All monthly data are aggregated into annual means starting from 1901. Subsequently, the SAT anomalies are computed for each ensemble member using 1950–2014 as a reference period. We have not explored the sensitivity resulting from another reference period. This furnishes a multi-model ensemble comprising 801 members derived from 47 AOGCMs within the 1901–2020 period. All model data is regridded using bilinear interpolation on the GISTEMP horizontal grid.

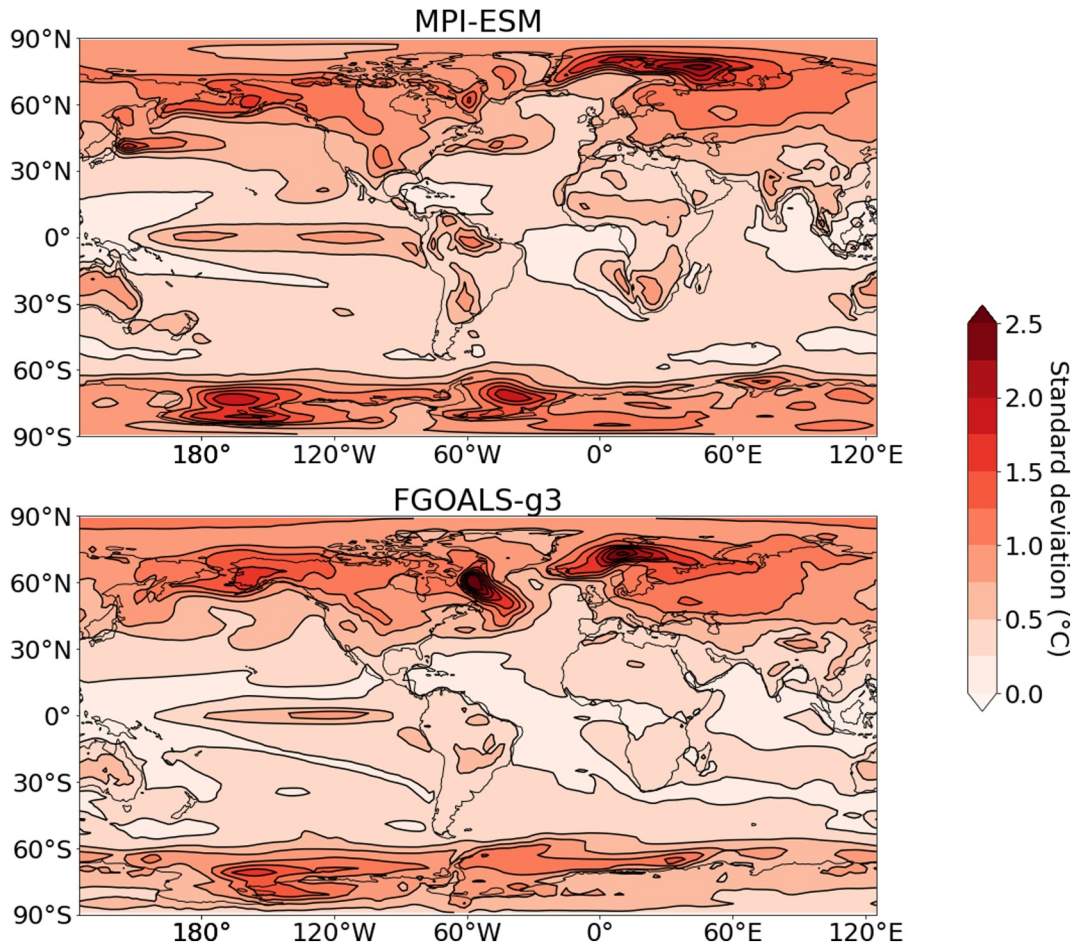

Figure 1. Standard deviation of the SAT deviations from the ensemble mean for (top) MPI-ESM and (bottom) FGOALS-g3.

2.3. Validation of the Data Set

The forced variability simulated within the multi-model ensemble is succinctly examined for the MPI-ESM and FGOALS-g3 climate models from SMILE. These two ensembles are selected as they have a very large size of 100 and 115 members, respectively, which largely exceed the size of other model ensembles. 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 and 40 members. Specifically, at least 6 members are required to detect a change in global SAT between the decades 2005–2014 and 2028–2037 (Deser et al., 2012, 2014). This can exceed 10 members for local analyses such as in North America (Deser et al., 2012, 2014). The data originating from these two models is subsequently employed to evaluate the results of the neural network model in Section 4.1.

We characterize the forced and internal variability from these two ensemble simulations using the ensemble mean and standard deviations, respectively (Deser et al., 2014; Frankcombe et al., 2015). Figure 1 illustrates the standard deviation of the SAT deviation from the ensemble mean. The variability in SAT is more pronounced over land surfaces (~0.3°C) compared to oceans (~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, we observe a marked variability in the equatorial Pacific Ocean linked to the phenomenon of El Niño Southern Oscillation (Neelin et al., 1998), with a standard deviation reaching 1°C (see Figure 1), with a variability more prominent in MPI-ESM than in FGOALS-g3. We found a large peak of variability localized in the subpolar North Atlantic, especially notable for FGOALS-

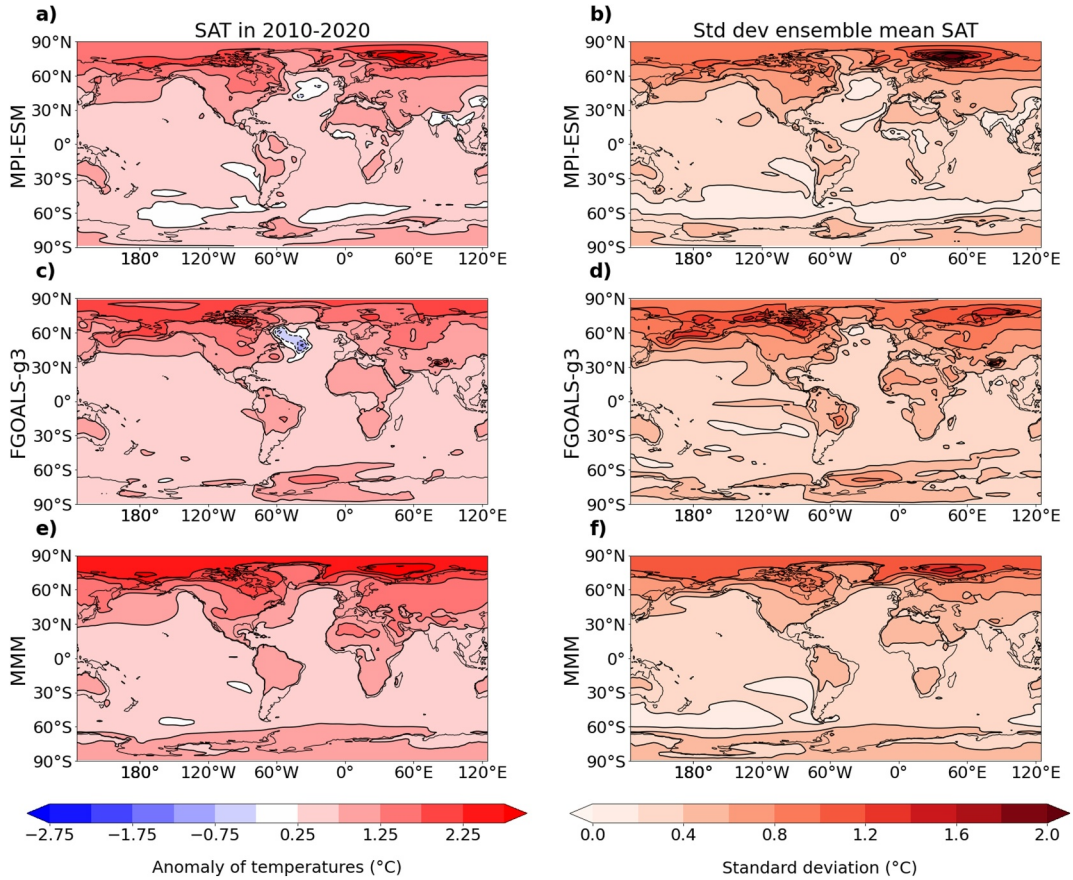

Figure 2. (a) Ensemble mean of the air surface temperature anomaly (°C) in MPI-ESM in 2010–2020 with 1950–2014 as reference period. (c) same as panel (a) but for FGOALS-g3. (e) same as panel (a) but for the multi-model mean (MMM). (b) Standard deviation of the ensemble mean surface air temperature anomaly (°C) in 1901–2020 for MPI-ESM. (d) Same as (b) but for FGOALS-g3. (f) Same as panel (b) but for the MMM.

g3 (reaching up to 2°C), which might reflect an important Atlantic Multidecadal Variability. Significant internal variability is also illustrated over mid-latitude land regions, reflecting the influence of extratropical weather fluctuations. Maxima around regions adjacent to the sea ice edge is also simulated is both models. Such spatial pattern reveals 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 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 (see Tables S1, S2 and S3 in Supporting Information S1), ensuring equal weight for each model. Nonetheless, MPI-ESM and FGOALS-g3 (see Table S3 in Supporting Information S1) 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, Figures 2a, 2c, and 2e illustrates the ensemble mean SAT anomaly for MPI-ESM, FGOALS-g3, and the MMM over 2010–2020 when the forced variability is expected to be dominant. Furthermore, Figures 2b, 2d, and 2f 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 rather low ECS of these two models (3.6°C for MPI-ESM and 2.8°C for FGOALS-g3, Zelinka et al., 2020) when compared to other models employed in this study. 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), was suggested to be associated with a deceleration of the Atlantic meridional overturning circulation (Caesar et al., 2018). It is worth noting that such a minimum is less pronounced in the MMM, presumably due to considerable uncertainties regarding the precise location and intensity of this warming hole. Indeed climate models simulate this warming hole differently, making it less significant in the MMM.

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 a larger greater amplification (refer to Figure 2) in MPI-ESM (1.54°C) than in FGOALS-g3 (1.45°C). 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 the mean sea ice extent. 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 (Goessling et al., 2016). The time standard deviation of the spatially averaged ensemble mean SAT is 0.34°C for MPI-ESM and 0.43°C for FGOALS-g3, while the time standard deviation of the spatially averaged SAT deviations from the ensemble mean is 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 in the 1901–2020 period.

3. Methods

3.1. Neural Network

We use a Convolutional neural network (CNN) to remove the internal variability from the SAT.

A neural network's is shaped by its hyperparameters, which encompass both its architecture and training process. Our approach involves utilizing three distinct data sets, training, validation and testing data set, each composed of input and desired output pairs. The training data set is used to determine the neural network's weights and biases. The validation data set comes into play for estimating the hyperparameters. The test data set is employed to assess the neural network's performance.

3.2. Constitution of the Training, Validation and Test Data Sets

To construct the training data set, 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 consists of all possible 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 (in this case the average image), resulting 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 assimilated to noisy images, where the internal variability introduces the noise component. The ensemble members' mathematical expectation is assimilated to the forced variability, which can be approximated through the ensemble mean.

To create the training data set we consider like Lehtinen et al. (2018) the ensemble mean of each climate model as an additional member. This inclusion only serves to accelerate the training process without introducing other influences. We then constitute for each climate model all possible pairs of different members, with the exception of MPI-ESM, FGOALS-g3 and MIROC6, which are reserved for testing and validation purposes as detailed later. These pairs constitute the desired set of input/output pairs of the neural network. Each member (and the ensemble mean) is therefore used as an input and is associated to all other members (and the ensemble mean) as output in the training process. So if we denote by n the number of members of a model ensemble simulation, this approach produces $n\ (n + 1)$ input/output pairs.

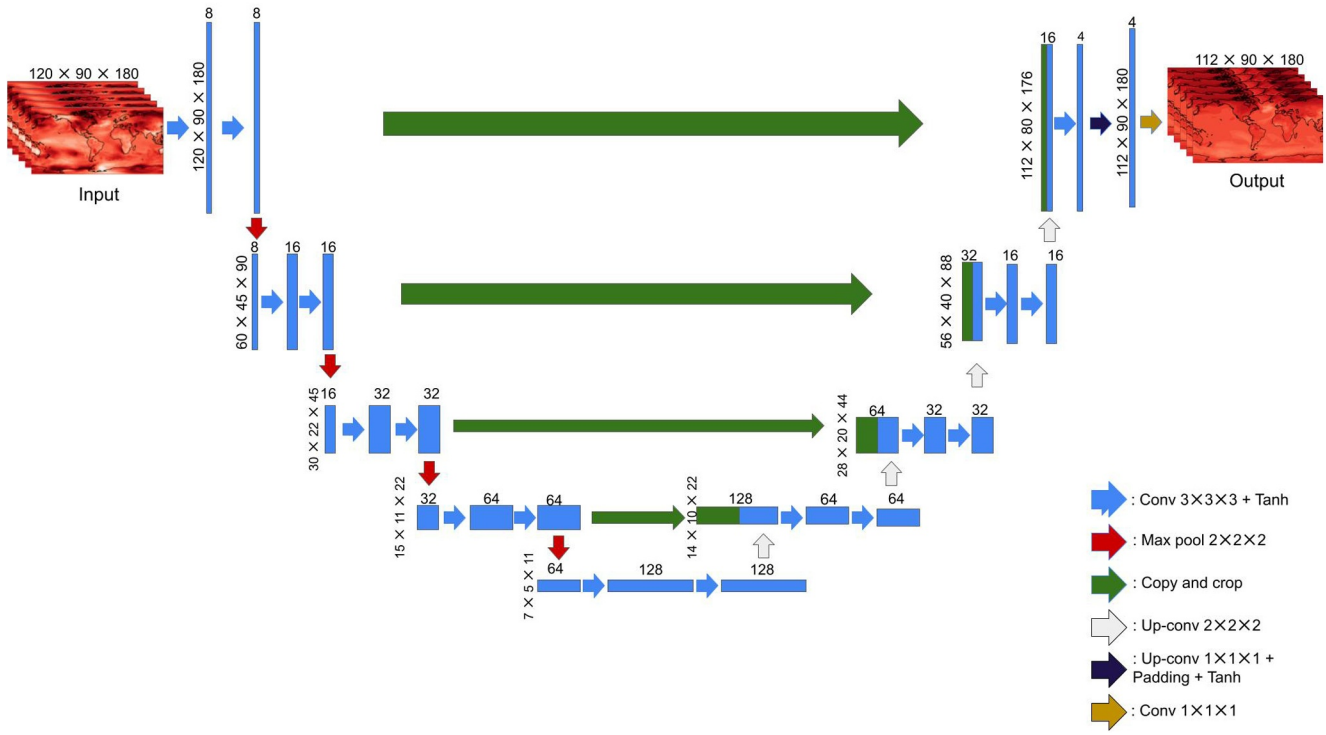

Figure 3. Schematic of the U-Net. The arrows represent the operations within the network. The numbers shows the dimension of the data and the number of filters used.

In total, this provides us with a total of 10,630 input/output pairs. By accumulating such pairs from all models, the resulting training data set primarily comprises ensembles with the largest sizes as seen in Tables S1, S2 and S3 in Supporting Information S1.

To create the validation set, we use the ensemble simulation 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 as the desired output.

To form the test data set, we use the data derived from the FGOALS-g3 and MPI-ESM models, using their extensive ensemble sizes of $n = 110$ and $n = 100$ respectively. Subsequently, we compare 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 show some dependence on the specific model being analyzed, as alternative climate models could have other outcomes. Nevertheless, this approach has been chosen due to its simplicity and its potential to mitigate the impact of any remaining internal variability when using small ensemble size.

3.3. U-Net

Convolutional neural networks (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 the temporal and spatial dimensions of input data.

In this context, a U-Net architecture is employed, which falls within the category 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 Figure 3). The contracting path adheres to a conventional design of a convolutional network, featuring numerous convolutional layers (denoted "Conv" in Figure 3), each followed by an activation function and a max-pooling operation (denoted "Max pool" in Figure 3). The max-pooling operation selects the

maximum element from a kernel that slides across the dimension resulting in a downsampled feature map. As the contracting path advances, spatial and temporal information is diminished while feature information (e.g., the thickness of the layers) is enriched. Conversely, the expansive path amalgamates feature (resulting from downsampling) and information at a smaller scale through a sequence of up-convolutions (denoted "Up-conv" in Figure 3) and concatenations (denoted copy and crop in Figure 3) with high-resolution features derived from the contracting path. These concatenation steps enable the U-Net to jointly study the large- and low-scale information present in the input.

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. This results in the detection of the features in the spatio-temporal domain, without assuming fixed spatial patterns. The last layer is a linear layer, as commonly used in regression. The selected activation function is the hyperbolic tangent (Rasamoelina et al., 2020) (denoted Tanh in Figure 3).

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 1905–2016, 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 4 years. This reduction addresses the substantial uncertainty typically observed at the data set's endpoints, as illustrated later on. The neural network included of a total of 5,659,009 trainable parameters.

For the training process we choose a batch size of 8, and the optimization process employs the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001. The choice of values of these two hyperparameters is arbitrary, and is chosen to limit the use of computational resources. 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. For the time, this results in a distortion of the beginning and end of the time period. Therefore we excluded the initial and final 4 years (1901–1904 and 2017–2020) in the U-Net's outputs. When applying the 1900–2020 period for the output (without excluding the first and last 4 years), the errors actually increase (not shown) which explains why we exclude the endpoints in the present analysis. Circular and periodic padding in the spatial dimension could be implemented to solve these issues. However, that this should only marginally affect the U-Net's results at the end points of the data.

The validation data set 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 "number of filters" designates to the thickness of the convolutional layers. The number of epochs refers to how many times the training data set is processed during the training phase. These hyperparameters are selected to minimize a cost function using the validation data set. The chosen cost function is the root mean squared error (RSME), calculated using an area-weighted mean of the gridded data. The coefficients of the weighted mean are chosen relative to the area surface of each grid point. Examination of the validation RMSE for different values of epochs and layer thickness reveals a consistent pattern (see Figure S1 in Supporting Information S1): a significant reduction in RMSE occurs in the initial epochs, followed by a gradual increase. For all layer thicknesses, the minimum RMSEs are almost identical, suggesting that layer thickness is not an important factor. We select an initial thickness of 8 and 32 iterations. This average choice was made to limit the risk of over- or under-learning.

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 (center 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 exceed 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

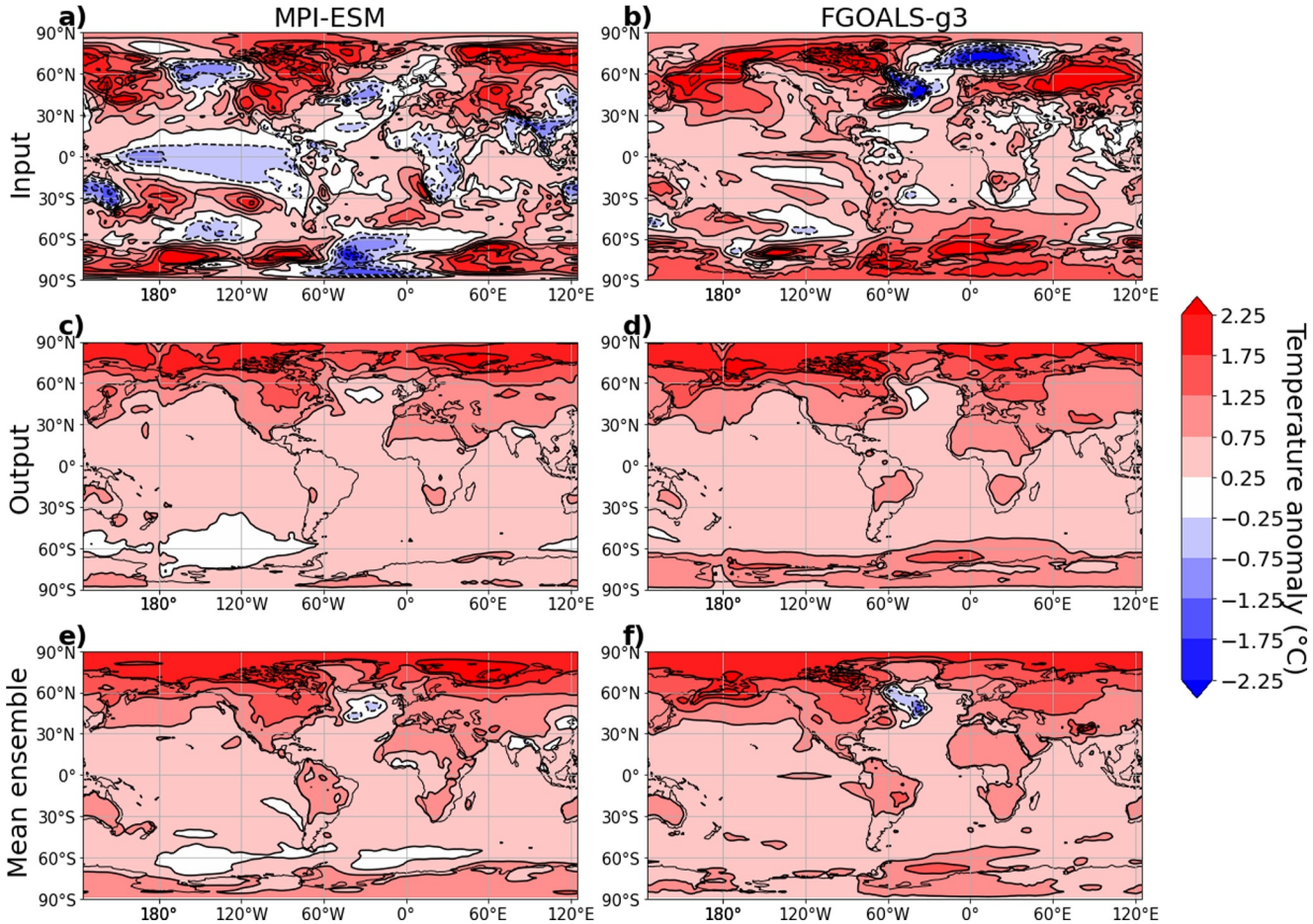

Figure 4. (First column, top to bottom) Anomalies of SAT in a randomly chosen member of MPI-ESM, the associated U-Net output and ensemble mean in 2016. (Second column) Same as the first column but for FGOALS-g3.

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 in the ensemble means of both models, the Arctic and its environs experience strong warming compared to other global regions, due to polar amplification. Conversely, minimal warming is observed in the Southern Ocean, and even a cold anomaly 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 center and bottom panels). In both instances, the pattern is relatively uniform, albeit with amplified warming observed over land areas, coupled with an Arctic Amplification phenomenon. This suggests that the internal variability—such as the influence of El Niño Southern Oscillation events or the effects of persistent weather patterns over continents—has been successfully eliminated. The warming or cooling anomalies are replicated, although the exact position and intensity do not precisely match those of the ensemble mean in certain areas, particularly the Southern Ocean. 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.

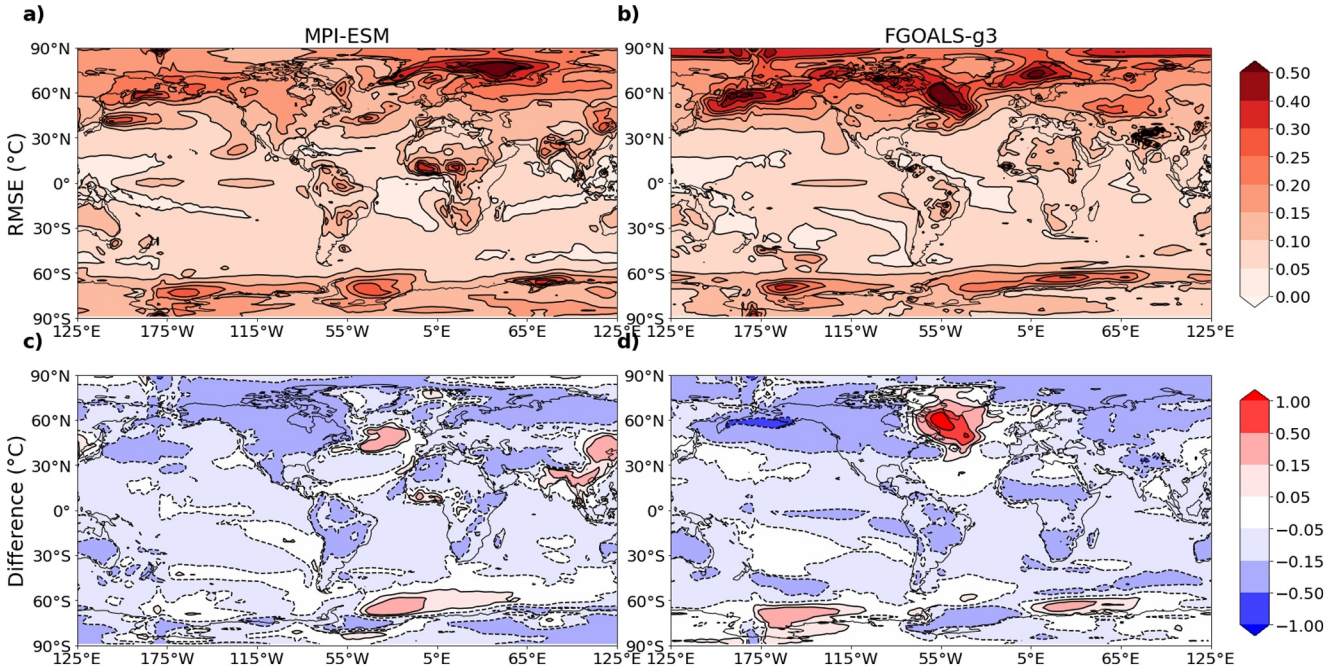

Figure 5. (a) Root mean square difference of the surface air temperature anomaly, in C°, between the outputs of the U-Net and the mean ensemble in MPI-ESM, calculated across the members and all years in 1905–2016 b. (b) Same as (a) but for FGOALS-g3 (c) Difference of the time mean SAT anomaly during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

4. The U-Net as an Internal Variability Filter

4.1. Filtering of the Test Climate Models

Once the U-Net is trained, we apply it to every member of the two test ensemble simulations: 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 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–0.5°C. The discrepancies generally remain below 0.2°C in tropical regions, except for instances over 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 locations 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. In Figure S2 in Supporting Information S1 we show the ratio of the inter-member standard deviation between the U-Net results and the input data. This ratio is calculated using the deviations from the ensemble mean and quantifies the reduction of variability when applying the U-Net. This map shows a uniform reduction of standard deviation, with a ratio of approximately 0.2 with the exception of the North Atlantic in FGOALS-g3 model. The U-Net therefore reduces the effect of internal variability uniformly. The only exception is the North Atlantic Ocean, where the U-Net does not perform similarly in both models.

Between the years 1996 and 2016, both ensemble results exhibit a warming that is roughly 0.1°C lower in the U-Net results when compared to the ensemble mean (as observed in Figures 5c and 5d). We study this period as it represents the last 20 years of the output where climate change is expected to be dominant. This difference is indicated by the nearly consistent negative anomalies situated between latitudes 45°N and 45°S. 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. This underestimation reaches 0.15°C for MPI-ESM and 0.13°C for FGOALS-g3 in these regions. This underestimation of the warming can be explained by the impact of MMM on U-Net's results. The two test models, FGOALS-g3 and MPI-ESM, have relatively high

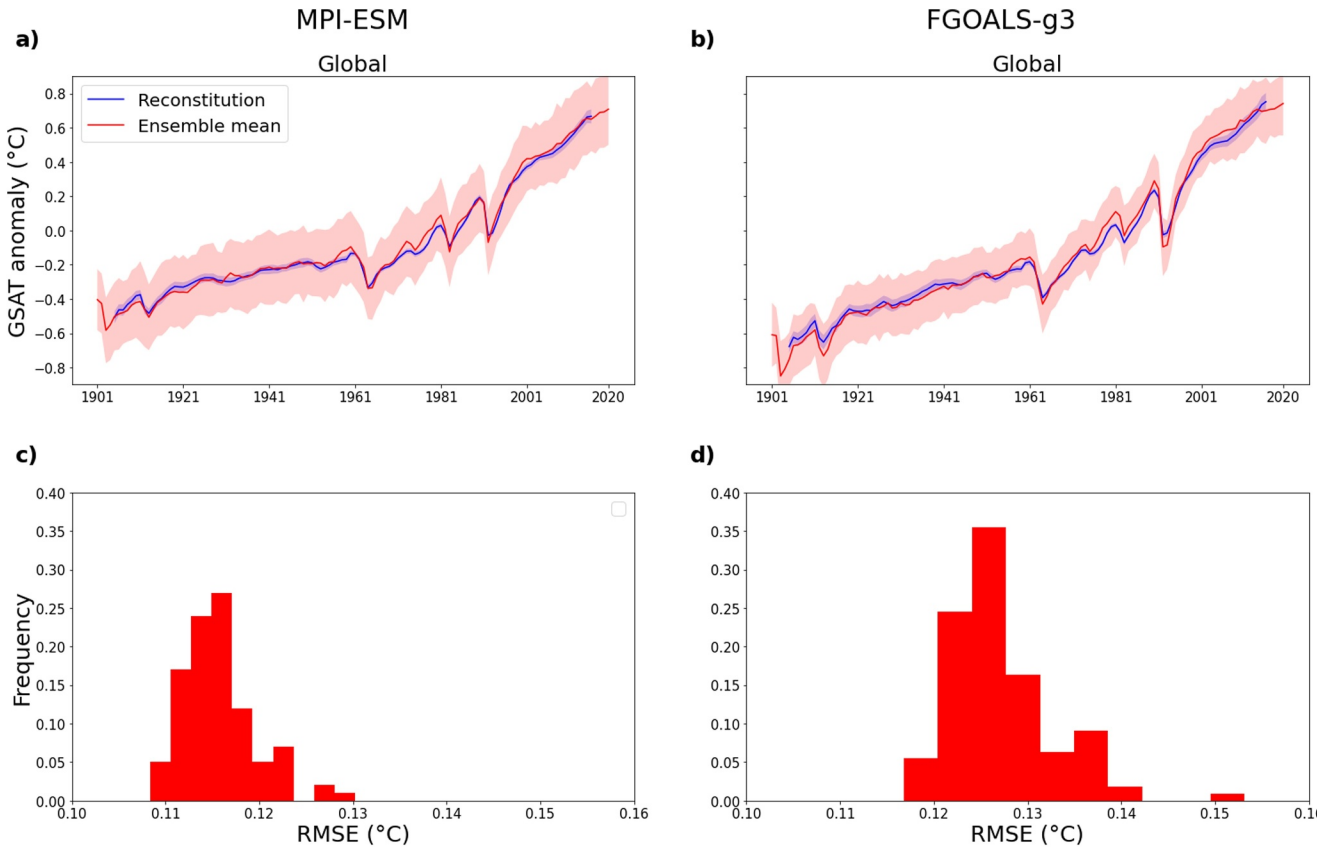

Figure 6. (a) Time evolutions of the global mean surface air temperature, in °C, for the ensemble mean and the mean U-Net outputs for MPI-ESM. Color shade shows the spread of the time series, with the interval encompassing 90% of the distribution assuming a gaussian distribution. (b) Same as panel (a) but for FGOALS-g3. (c) Histogram showing the distribution of the RMSE between the mean ensemble and the U-Net outputs for MPI-ESM over the 1905–2016 period. (d) Same as panel (c), but for FGOALS-g3.

ECS compared with the models used to train the U-Net. However, the U-Net does not simply reproduce the MMM as an output without taking its input into account. In Figure S3 in Supporting Information S1 we show a figure equivalent to Figure 5 using the raw MMM instead of the U-Net to estimate the forced variability. The results obtained show RMSEs and that are much larger than those obtained with U-Net, indicating that U-Net provides an estimated forced variability better than the MMM.

The systematic underestimation does not apply to 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 difference 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 evolution 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 MMM (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).

Figures 6a and 6b 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 a Gaussian distribution. The time evolution of the forced variability 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).

From 1905 to 2016, a GSAT rise is observed, aligning with the observed 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 (see Figures 6a and 6b). 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 Figures 6c and 6d as red histograms), reveals errors of around 0.12°C in MPI-ESM and 0.13°C in FGOALS-g3.

In Figure S4 in Supporting Information S1, the quadratic errors between the ensemble mean 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 are largest during the initial and final years. There is also a local maximum in 1975–1985 in both models. This underscores the presence of substantial uncertainties at the data's edge. 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 in CMIP5 (Taylor et al., 2012) for MPI-ESM and CMIP6 (Eyring et al., 2016) 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 Figure 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 individual 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 (see Figure 7). This RMSE computation is performed across all grid points and is spatially averaged. The intervals that encompass 90% of the distribution from the subsamples, assuming a gaussian distribution 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), Nino3 (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 contrasted forced and internal variability (see Figures 1 and 2). Additionally, the same calculation is applied separately to land and ocean region in the 60°S–60°N.

Figure 7a illustrates the diminution of errors in estimating the forced variability within the subset of members as the size of the subset increases. A larger subset size leads to better estimations of forced variability as the residual internal variability quantified by the standard deviation decreases by a factor of $\sqrt{m}$. 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 oceans and lands regions between 60°N and 60°S, the reduction of the variability remain largely consistent, with a reduction in error magnitude by a factor of approximately four (see Figures 7d and 7g). This reduction corresponds closely to that achieved by using a subset of 15–20 members.

The U-Net's efficacy stands out prominently over the equatorial Pacific region (see Figure 7f) where the U-Net achieves a substantial reduction in variability, amounting to a factor of 5.5. This reduction is similar to that derived from an ensemble mean of 30 members.

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 results

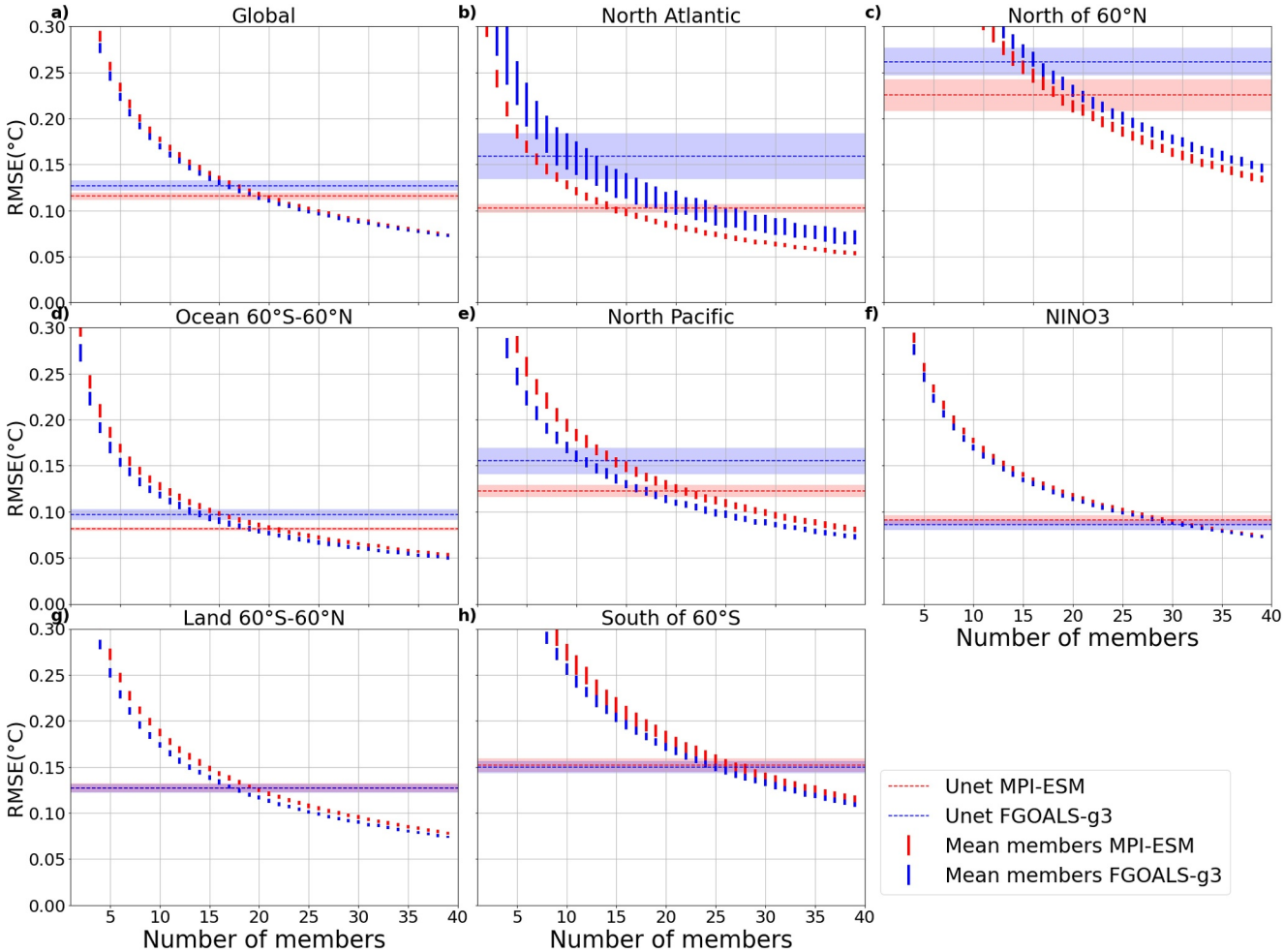

Figure 7. Spatial average of the RMSE in the 1905–2016 period for the forced variability estimated with the U-Net outputs obtained from each ensemble member, and the forced variability obtained with ensemble averages subsampling ensemble of size 1 to 40; for (red) MPI-ESM and (blue) FGOALS-g3. The RMSE 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 and 40 is illustrated with (vertical lines) the intervals including 90% of the ensemble member distribution, also assuming a gaussian distribution.

remains relatively steady. However, in terms of removing internal variability, the U-Net showcases higher efficiency in MPI-ESM for most regions. However, in the North Atlantic (Figure 7b) 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 five 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 observational data.

4.2. Filtering of the Observations

The U-Net is now employed to process SAT observations derived from GISTEMP. 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

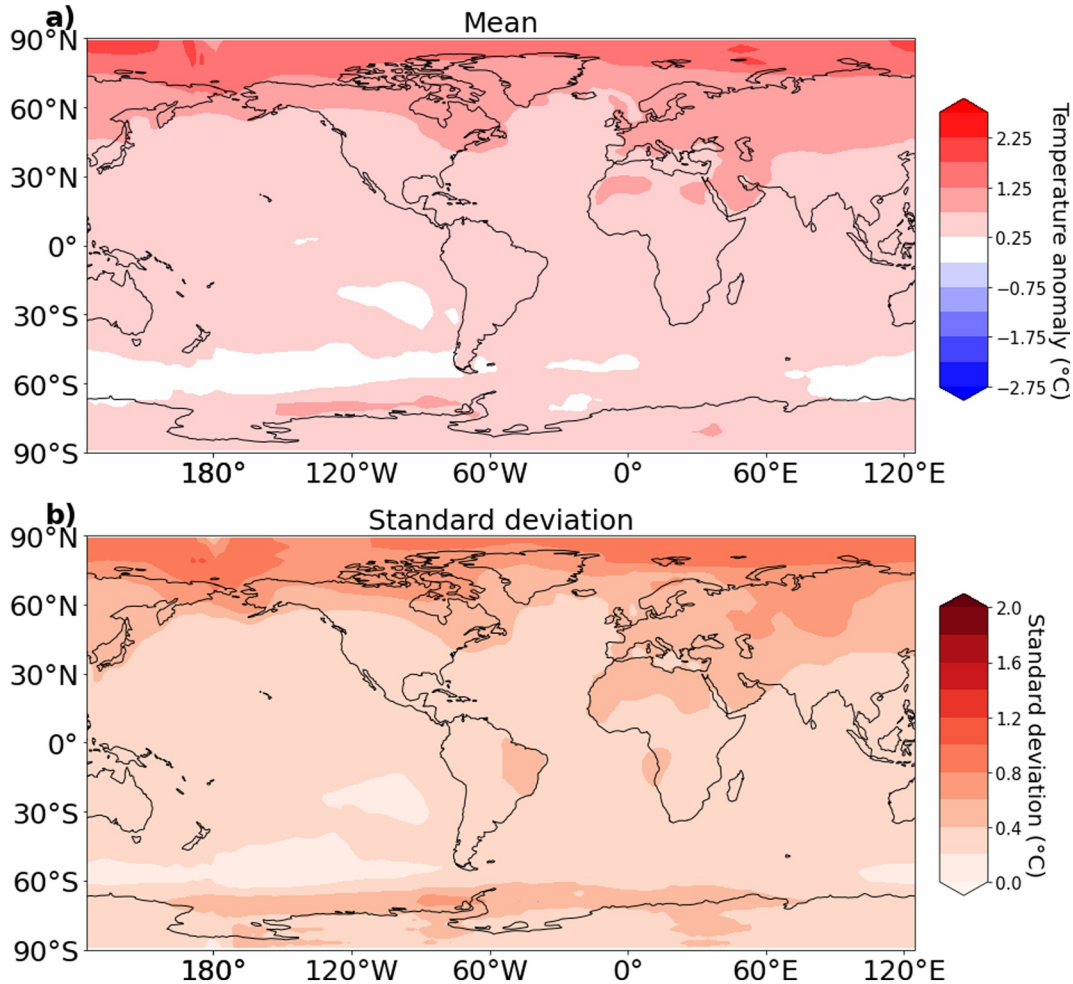

Figure 8. Forced surface air temperature (in °C) anomaly when applying the U-Net to observation: (a) time average in 1996–2016; (b) standard deviation in 1905–2016.

1905 to 2016 using the 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 MMM (as depicted in Figure 2). This highlights the important contribution of the training data set in determining the identified forced changes.

To quantify internal variability within the observations, we compute the deviations of observed SAT anomalies from the forced changes estimated by the U-Net. The resulting internal variability pattern, illustrated by the time standard deviation of these deviations shown in Figure 9, mirrors the model-derived pattern (Figure 1). Important 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 Nino3.4 region (5°N–5°S, 170°W–120°W), with a particular emphasis on Nino3.4 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

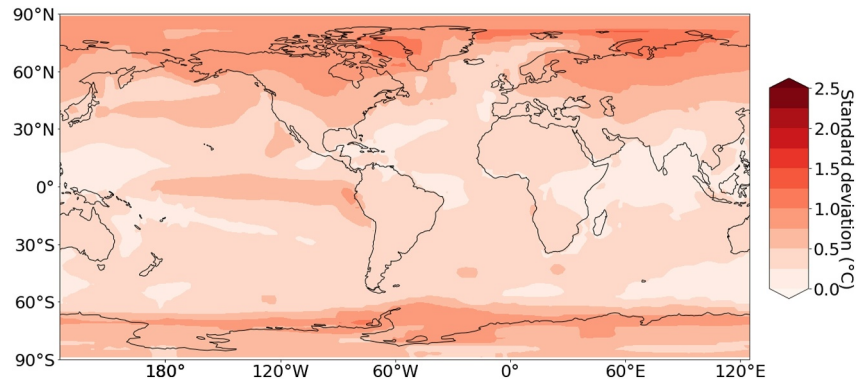

Figure 9. SAT standard deviation of the deviations from the forced signal, as estimated using the U-Net, in 1905–2016.

the Nino3.4 region (Figure 10b), where variability at 2–7 years is almost entirely eliminated. The U-Net estimates the Nino3.4 forced variability, depicting a steady warming trend. To quantify the changes of SAT in Nino3.4 relative to the tropics, we calculate also the relative SAT, defined as the difference between the average SAT on the Nino3.4 region and the average SAT on ocean grid between 30°S and 30°N. The relative SST shows that the warming over the Nino3.4 follows that of the tropics, so that no clear El Niño-like reponse is found, unlike climate models (Figure 2). Some authors (Clement et al., 1996; Heede et al., 2020) have suggested that a forced cooling could exist in the relative SAT, which reflects a process referred to as thermostat effect. Here the relative SAT shows a very small cooling (see Figure 10c). In addition the relative SAT in the Nino3.4 region is not affected by the forcing from the main volcanic eruptions. Therefore, no evidence for a systematic El-Niño response to volcanic eruption (as in Khodri et al., 2017) is found as suggested in McGregor and Timmermann (2011).

5. Conclusion

A novel approach is introduced to effectively eliminate internal variability from a time-evolving two-dimensional data set, specifically focusing on the surface air temperature evolution in 1901– 2020. The method employs a U-

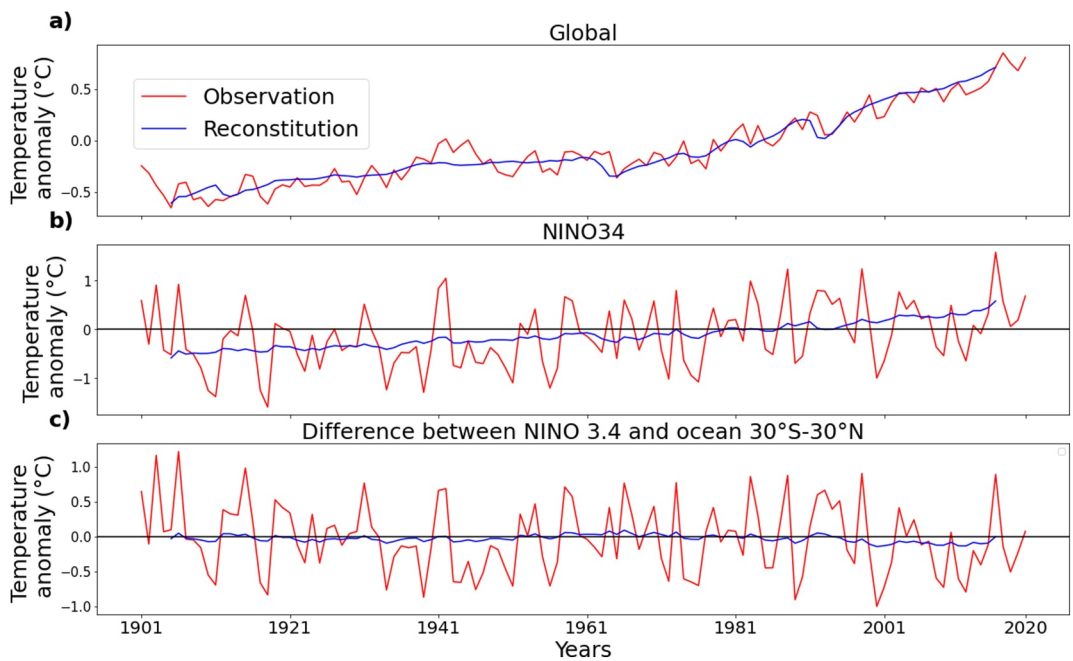

Figure 10. Time series of (red) the observed SAT anomaly and (blue) the forced SAT anomaly estimated by the U-Net for (a) the global mean (b) Nino3.4 and (c) the relative SAT in Nino3.4, calculated as the difference between the averaged SAT in Nino3.4 region and the tropical ocean SAT (30°S–30°N).

Net neural network and a 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 CMIP5 and CMIP6 data sets, to sample uncertainties due to models or forcing. Subsequently, this trained network is applied to observational data to unveil the forced variability signal by attenuating internal variability. The evaluation of this method involves utilizing large ensemble simulations from individual models, specifically the MPI-ESM and FGOALS-g3. The forced variability derived from the ensemble mean is then compared with the outcomes from the U-Net. To quantitatively assess the U-Net's performance in reducing internal variability, an "equivalent ensemble size" is computed (see Figure 7). 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 (e.g., Figure 7a). 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 MMM in terms of spatial patterns (e.g., Figure 8). The U-Net's results do not suggest an El Niño-like response to global warming (e.g., Figure 10). We observe that the U-Net encounters greater challenges in accurately estimating forced variability over the North Atlantic region (see Figure 5). 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 this region (Menary & Wood, 2018; Zhang, 2007) or can be linked to the large uncertainties in models for the evolution of the Atlantic Meridional overturning circulation (Jackson & Petit, 2023). Other choices of data set could have been used such as the National Centers for Environmental Information data set (Smith et al., 2008) or HadCrut5 (Morice et al., 2021). The sensibility of this neural network-based methodology to instrumental uncertainties remains to be established.

A comprehensive comparison with other methods would allow a more complete assessment of the effectiveness of the U-Net. Although such comprehensive assessment is lacking, the results obtained by the U-Net on surface air temperature seems equivalent to those obtained by the low frequency pattern filtering or signal-to-noise pattern filtering (Wills et al., 2020). These methods are much easier to implement involving only two hyperparameters, which limits the risks of overfitting. However, there are some important differences between this method and our approach. The U-Net is a learning method, its performance is related to the quality and size of its training database which is expected to improve over time as climate models data increases in size and improve in quality. The U-Net is also a method with great potential for improvement that focuses on detecting spatiotemporal features in the data and does not assume that the forced variability can be captured by a linear combination of spatial patterns. The linear methods might encounter more difficulties when applied on precipitation which can show non-linear changes. The U-net might be used to study these variables and possibly simultaneously by exploiting the correlations between the different physical variables. Several improvements can be implemented like better padding or a more extensive optimization of hyperparameters. A potential approach is to address the sensitivity of U-Net to the multi-model consensus of future variability by using neural network regularization techniques, such as dropout layers (Wan et al., 2013). In addition, pretreatment methods such as data augmentation could be explored. Improving the evaluation process of the U-Net's performance could involve testing the U-Net on a broader range of climate models to assess its generalizability.

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 data sets. 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.

Data Availability Statement

The CMIP5 and CMIP6 data is available through the Earth System Grid Federation and can be accessed at https://esgf-node.ipsl.upmc.fr/projects/esgf-ipsl/. The data from the Multi-Model Large Ensemble Archive is available at

https://www.cesm.ucar.edu/communityprojects/mmlea. Codes used in this article for the U-Net and the figures are from Bône (2023) software, available at https://zenodo.org/record/8233743.

Acknowledgments
We acknowledge the US CLIVAR Working Group on Large Ensemble for the Multi-Model Large Ensemble Archive (Deser et al., 2020). 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 ANR-19-JPOC-003).

References

Allen, M. R., & Stott, P. A. (2003). Estimating signal amplitudes in optimal fingerprinting, Part I: Theory. *Climate Dynamics*, *21*(5–6), 477–491. https://doi.org/10.1007/s00382-003-0313-9

Allen, M. R., & Tett, S. F. (1999). Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, *15*(6), 419–434. https://doi.org/10.1007/s003820050291

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

Bonnet, R., Swingedouw, D., Gastineau, G., Boucher, O., Deshayes, J., Hourdin, F., et al. (2021). Increased risk of near term global warming due to a recent AMOC weakening. *Nature Communication*, *12*, 6108. https://doi.org/10.1038/s41467-021-26370-0

Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., & Saba, V. (2018). Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature*, *556*(7700), 191–196. https://doi.org/10.1038/s41586-018-0006-5

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. https://doi.org/10.1175/bams-d-16-0286.1

Chen, X., & Tung, K.-K. (2014). Varying planetary heat sink led to global-warming slowdown and acceleration. *Science*, *345*(6199), 897–903. https://doi.org/10.1126/science.1254937

Chen, X., & Tung, K.-K. (2018). Global surface warming enhanced by weak atlantic overturning circulation. *Nature*, *559*(7714), 387–391. https://doi.org/10.1038/s41586-018-0320-y

Clement, A. C., Seager, R., Cane, M. A., & Zebiak, S. E. (1996). An ocean dynamical thermostat. *Journal of Climate*, *9*(9), 2190–2196. https://doi.org/10.1175/1520-0442(1996)009<2190:aodt>2.0.co;2

Collier, M. A., Jeffrey, S. J., Rotstayn, L. D., Wong, K., Dravitzki, S., Moseneder, C., et al. (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).

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. https://doi.org/10.1175/2010jcli3659.1

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*(4), 277–286. https://doi.org/10.1038/s41558-020-0731-2

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: The role of internal variability. *Climate Dynamics*, *38*(3–4), 527–546. https://doi.org/10.1007/s00382-010-0977-x

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. https://doi.org/10.1175/jcli-d-13-00451.1

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arneth, A., Arsouze, T., et al. (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. https://doi.org/10.1016/s0031-3203(01)00178-9

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. https://doi.org/10.1002/joc.1881

England, M. H., McGregor, S., Spence, P., Meehl, G. A., Timmermann, A., Cai, W., et al. (2014). Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nature Climate Change*, *4*(3), 222–227. https://doi.org/10.1038/nclimate2106

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. https://doi.org/10.5194/gmd-9-1937-2016

Eyring, V., Gillett, N., Rao, K. A., Barimalala, R., Parrillo, M. B., Bellouin, N., 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*. Cambridge University Press.

Fan, Y., Liu, W., Zhang, P., Chen, R., & Li, L. (2023). North Atlantic oscillation contributes to the subpolar north Atlantic cooling in the past century. *Climate Dynamics*, *61*(11), 5199–5215. https://doi.org/10.1007/s00382-023-06847-y

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. https://doi.org/10.1175/jcli-d-15-0069.1

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. https://doi.org/10.1073/pnas.2016549118

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., et al. (2016). The detection and attribution model intercomparison project (DAMIP v1. 0) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3685–3697. https://doi.org/10.5194/gmd-9-3685-2016

GISTEMP Team. (2023). GISS Surface Temperature Analysis (GISTEMP), version 4 [Dataset]. *NASA Goddard Institute for Space Studies*. Retrieved from https://data.giss.nasa.gov/gistemp/

Goessling, H. F., Tietsche, S., Day, J. J., Hawkins, E., & Jung, T. (2016). Predictability of the arctic sea ice edge. *Geophysical Research Letters*, *43*(4), 1642–1650. https://doi.org/10.1002/2015gl067232

Gulev, S. K., Thorne, P. W., Ahn, J., Dentener, F. J., Domingues, C. M., Gerland, S., et al. (2021). Changing state of the climate system. https://doi.org/10.1017/9781009157896.004

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, *48*(4). https://doi.org/10.1029/2010rg000345

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. https://doi.org/10.1175/1520-0442(1995)008<0474:ivsfav>2.0.co;2

Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, *6*(10), 1957–1971. https://doi.org/10.1175/1520-0442(1993)006<1957:offtdo>2.0.co;2

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009bams2607.1

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. https://doi.org/10.1175/jcli-d-19-0690.1

Ilesanmi, A. E., & Ilesanmi, T. O. (2021). Methods for image denoising using convolutional neural network: A review. *Complex & Intelligent Systems*, *7*(5), 2179–2198. https://doi.org/10.1007/s40747-021-00428-4

Jackson, L., & Petit, T. (2023). North Atlantic overturning and water mass transformation in CMIP6 models. *Climate Dynamics*, *60*(9–10), 2871–2891. https://doi.org/10.1007/s00382-022-06448-1

Jeffrey, S., Rotstayn, L., Collier, M., Dravitzki, S., Hamalainen, C., Moeseneder, C., et al. (2013). Australia's CMIP5 submission using the CSIRO-Mk3. 6 model. *Australian Meteorological and Oceanographic Journal*, *63*(1), 1–13. https://doi.org/10.22499/2.6301.001

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. https://doi.org/10.1029/2020ms002366

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (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. https://doi.org/10.1175/bams-d-13-00255.1

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. https://doi.org/10.1038/s41558-020-0819-8

Khodri, M., Izumo, T., Vialard, J., Janicot, S., Cassou, C., Lengaigne, M., et al. (2017). Tropical explosive volcanic eruptions can trigger El Niño by cooling tropical Africa. *Nature Communications*, *8*(1), 778. https://doi.org/10.1038/s41467-017-00755-6

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kosaka, Y., & Xie, S.-P. (2013). Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature*, *501*(7467), 403–407. https://doi.org/10.1038/nature12534

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. https://doi.org/10.1029/2018jd029522

Li, L., Yu, Y., Tang, Y., Lin, P., Xie, J., Song, M., et al. (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. https://doi.org/10.1029/2019ms002012

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. https://doi.org/10.1186/s40562-022-00244-9

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., et al. (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. https://doi.org/10.1029/2019ms001639

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., et al. (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.

McGregor, S., & Timmermann, A. (2011). The effect of explosive tropical volcanism on enso. *Journal of Climate*, *24*(8), 2178–2191. https://doi.org/10.1175/2010jcli3990.1

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. https://doi.org/10.1175/jcli-d-12-00548.1

Menary, M. B., Robson, J., Allan, R. P., Booth, B. B., Cassou, C., Gastineau, G., et al. (2020). Aerosol-forced AMOC changes in CMIP6 historical simulations. *Geophysical Research Letters*, *47*(14), e2020GL088166. https://doi.org/10.1029/2020gl088166

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. https://doi.org/10.1007/s00382-017-3793-8

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J., Hogan, E., Killick, R., et al. (2021). An updated assessment of near-surface temperature change from 1850: The HadCRUT5 data set. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. 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*, *103*(C7), 14261–14290. https://doi.org/10.1029/97jc03424

Newman, M., Alexander, M. A., Ault, T. R., Cobb, K. M., Deser, C., Di Lorenzo, E., et al. (2016). The Pacific decadal oscillation, revisited. *Journal of Climate*, *29*(12), 4399–4427. https://doi.org/10.1175/jcli-d-15-0508.1

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

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. https://doi.org/10.1029/2019gl086588

Rasamoelina, A. D., Adjailia, F., & Sinčák, P. (2020). A review of activation function for artificial neural network. In *2020 IEEE 18th world symposium on applied machine intelligence and informatics (SAMI)* (pp. 281–286).

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. https://doi.org/10.5194/bg-12-3301-2015

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* (Vol. 18, pp. 234–241). Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-24574-4_28

Schmidt, A., Mills, M. J., Ghan, S., Gregory, J. M., Allan, R. P., Andrews, T., et al. (2018). Volcanic radiative forcing from 1979 to 2015. *Journal of Geophysical Research: Atmospheres*, *123*(22), 12491–12508. https://doi.org/10.1029/2018jd028776

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., et al. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678. https://doi.org/10.1029/2019rg000678

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. https://doi.org/10.1029/2021gl094948

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., et al. (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16), 9591–9618. https://doi.org/10.5194/acp-20-9591-2020

Smith, D. M., Gillett, N. P., Simpson, I. R., Athanasiadis, P. J., Baehr, J., Bethke, I., et al. (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*. https://doi.org/10.3389/fclim.2022.955414

Smith, T. M., Reynolds, R. W., Peterson, T. C., & Lawrimore, J. (2008). Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *Journal of Climate*, *21*(10), 2283–2296. https://doi.org/10.1175/2007jcli2100.1

Solomon, A., Goddard, L., Kumar, A., Carton, J., Deser, C., Fukumori, I., et al. (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. https://doi.org/10.1175/2010bams2962.1

Steinman, B. A., Mann, M. E., & Miller, S. K. (2015). Atlantic and Pacific multidecadal oscillations and northern Hemisphere temperatures. *Science*, *347*(6225), 988–991. https://doi.org/10.1126/science.1257856

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. https://doi.org/10.1175/jcli-d-18-0134.1

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. https://doi.org/10.1038/nclimate2483

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. https://doi.org/10.1111/nyas.14659

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. https://doi.org/10.1175/bams-d-11-00094.1

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., et al. (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. https://doi.org/10.1016/j.neunet.2020.07.025

Ting, M., Kushnir, Y., Seager, R., & Li, C. (2009). Forced and internal twentieth-century SST trends in the North Atlantic. *Journal of Climate*, *22*(6), 1469–1481. https://doi.org/10.1175/2008jcli2561.1

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., et al. (2011). The representative concentration pathways: An overview. *Climatic Change*, *109*(1–2), 5–31. https://doi.org/10.1007/s10584-011-0148-z

Vincent, L., Zhang, X., Brown, R., Feng, Y., Mekis, E., Milewska, E., et al. (2015). Observed trends in Canada's climate and influence of low-frequency variability modes. *Journal of Climate*, *28*(11), 4545–4560. https://doi.org/10.1175/jcli-d-14-00697.1

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., & Fergus, R. (2013). Regularization of neural networks using dropconnect. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning* (Vol. 28, pp. 1058–1066). PMLR. Retrieved from https://proceedings.mlr.press/v28/wan13.html

Wang, C., Deser, C., Yu, J.-Y., DiNezio, P., & Clement, A. (2017). El Niño and southern oscillation (ENSO): A review. Coral reefs of the eastern tropical Pacific: Persistence and loss. In *A dynamic environment* (pp. 85–106).

Wang, C., & Picaut, J. (2004). Understanding ENSO physics—A review. Earth's climate: The ocean–atmosphere interaction. *Geophysical Monograph Series*, *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. https://doi.org/10.1175/jcli-d-19-0855.1

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. https://doi.org/10.1007/s13244-018-0639-9

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., et al. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782. https://doi.org/10.1029/2019gl085782

Zelinka, M. D., Zhou, C., & Klein, S. A. (2016). Insights from a refined decomposition of cloud feedbacks. *Geophysical Research Letters*, *43*(17), 9259–9269. https://doi.org/10.1002/2016gl069917

Zhang, R. (2007). Anticorrelated multidecadal variations between surface and subsurface tropical north Atlantic. *Geophysical Research Letters*, *34*(12). https://doi.org/10.1029/2007gl030225

Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y.-O., Marsh, R., Yeager, S. G., et al. (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. https://doi.org/10.1029/2019rg000644