# 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 validate 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 (∼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, 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

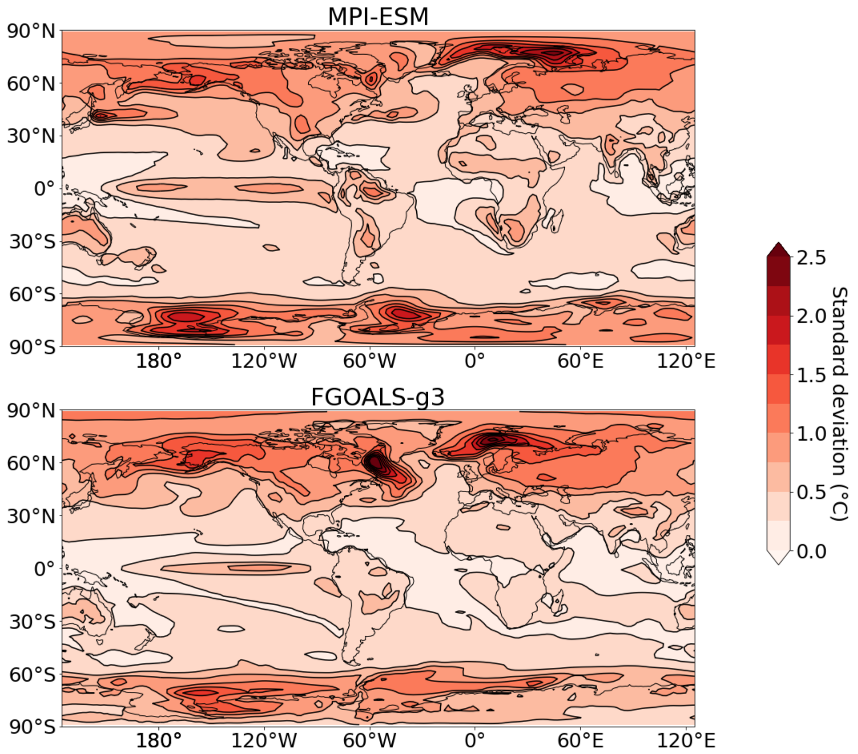


**Figure 1.** Standard deviation of the SAT deviations from the ensemble mean for (top) MPI-ESM and (bottom) FGOALS-g3.

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 ($\sim$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.

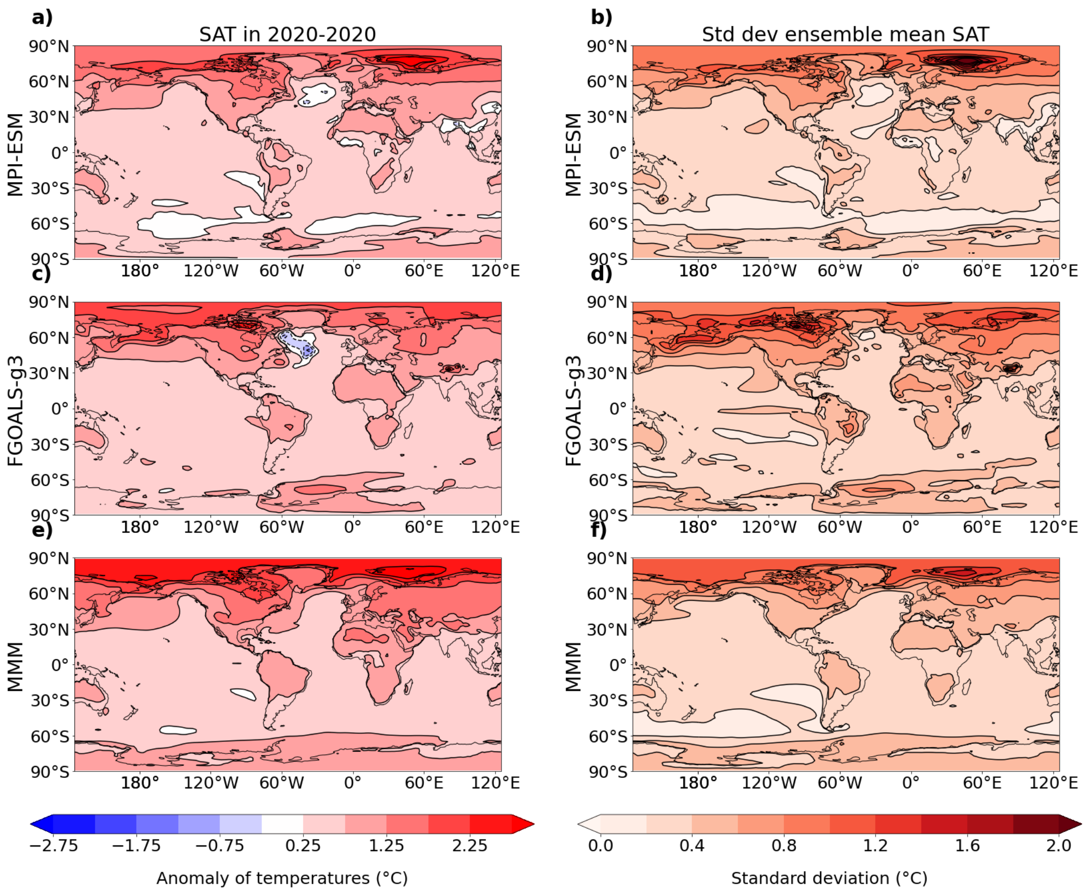


**Figure 2.** a) Ensemble mean of the air surface temperature (°C) in MPI-ESM in 2010-2020. c) Same as a) but for FGOALS-g3. e) Same as a) but for the MMM. b) Standard deviation of the ensemble mean surface air temperature (°C) in 1901-2020 for MPI-ESM. d) Same as b) but for FGOALS-g3. f) Same as b)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

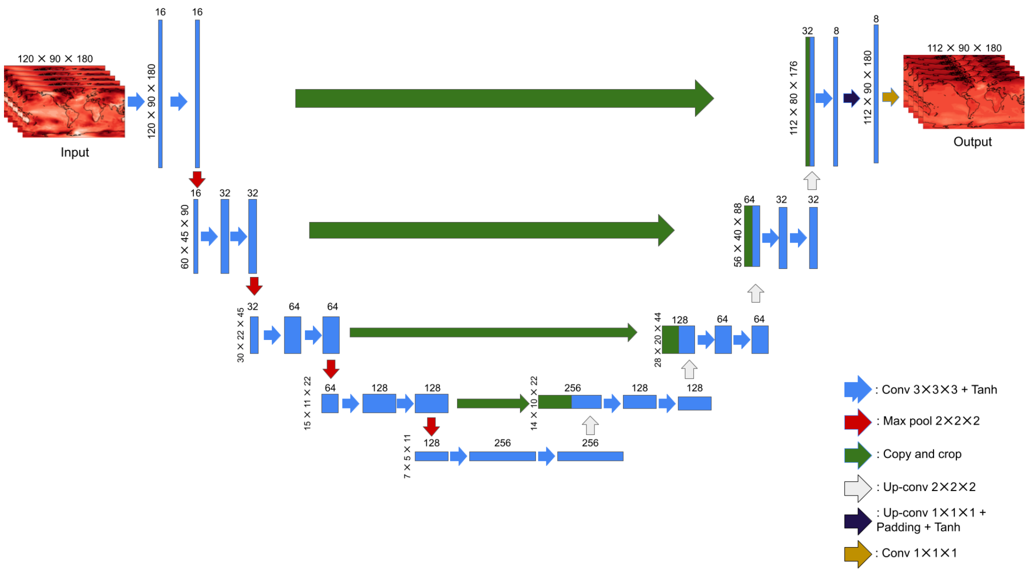


**Figure 3.** Schematic of the U-Net. The arrows represent the operations within the network. The numbers shows the dimension of the data and the number of filters used.

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 comprising 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

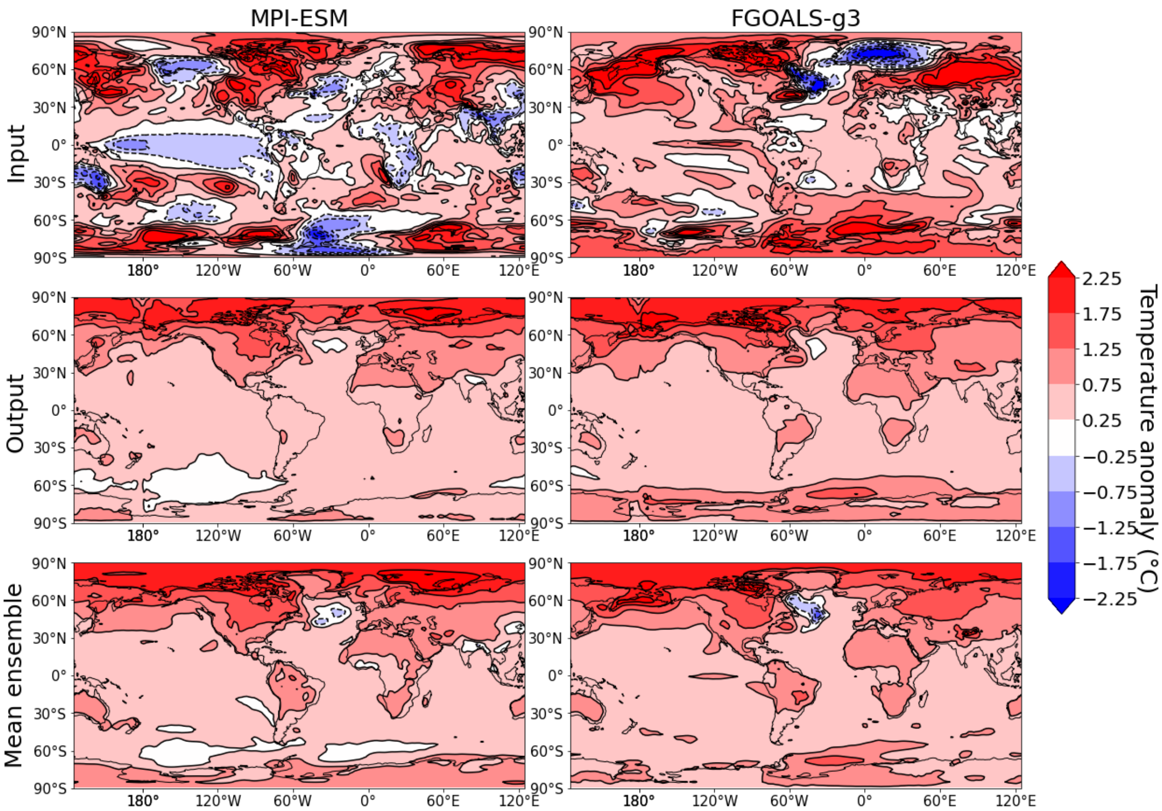


**Figure 4.** (First column) Anomalies of SAT in a randomly chosen member of MPI-ESM, the associated U-Net output and ensemble mean in 2016. (Second column) Same as the first column but for a randomly chosen ensemble member for FGOALS-g3.

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).

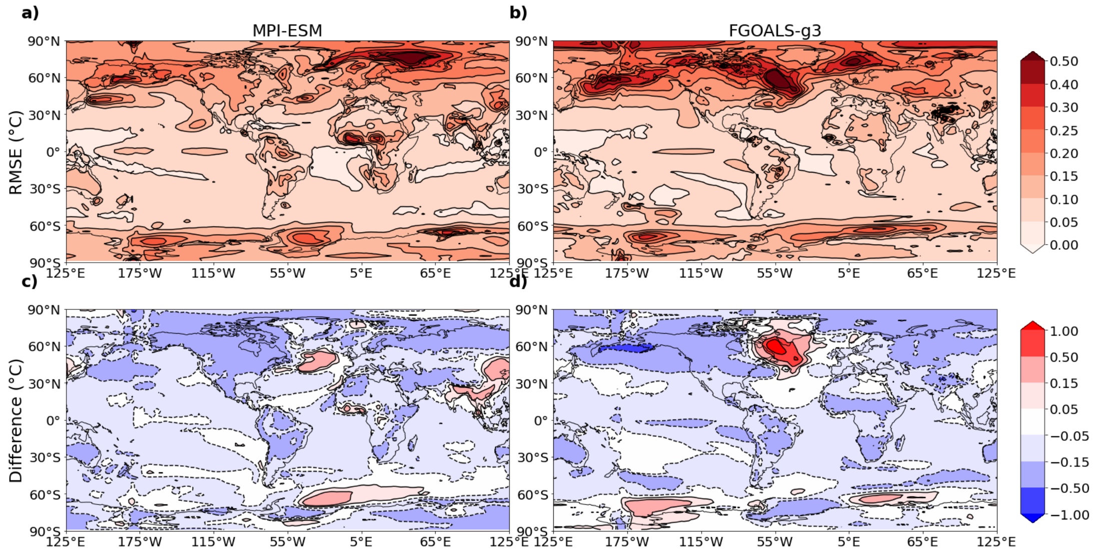


**Figure 5.** a) Root mean square difference of the surface air temperature, in C°, between the outputs of the U-Net and the mean ensemble in MPI-ESM, calculated across the members and all years in 1905-2016 b. b) Same as a) but for FGOALS-g3 c) Difference of the time mean SAT anomaly during 1996-2016,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

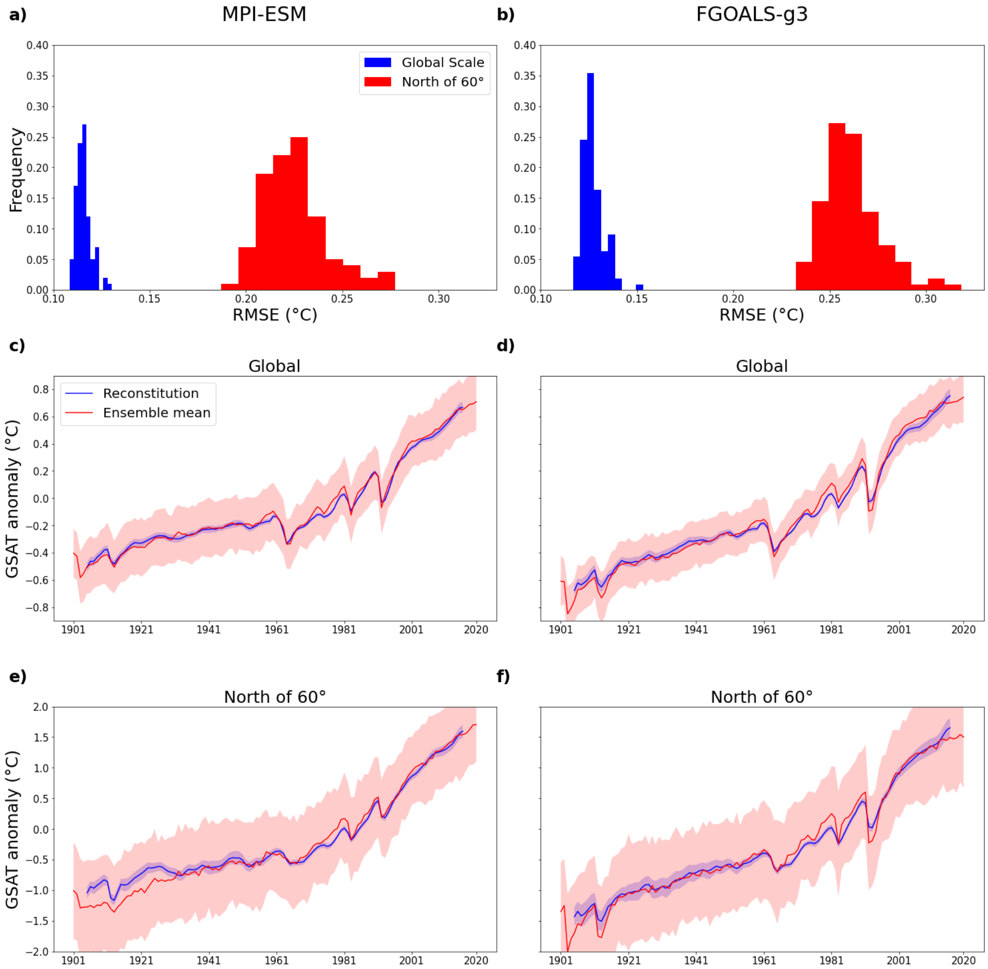


**Figure 6.** a) Histogram showing the distribution of the RMSE between the mean ensemble and the U-Net outputs of MPI-ESM. b) Same as a), but for FGOALS-g3. c) Time evolutions of the global mean surface air temperature, in °C, for the ensemble mean and the mean U-Net outputs for MPI-ESM. Color shade 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 variability, 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 ($\sim$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.

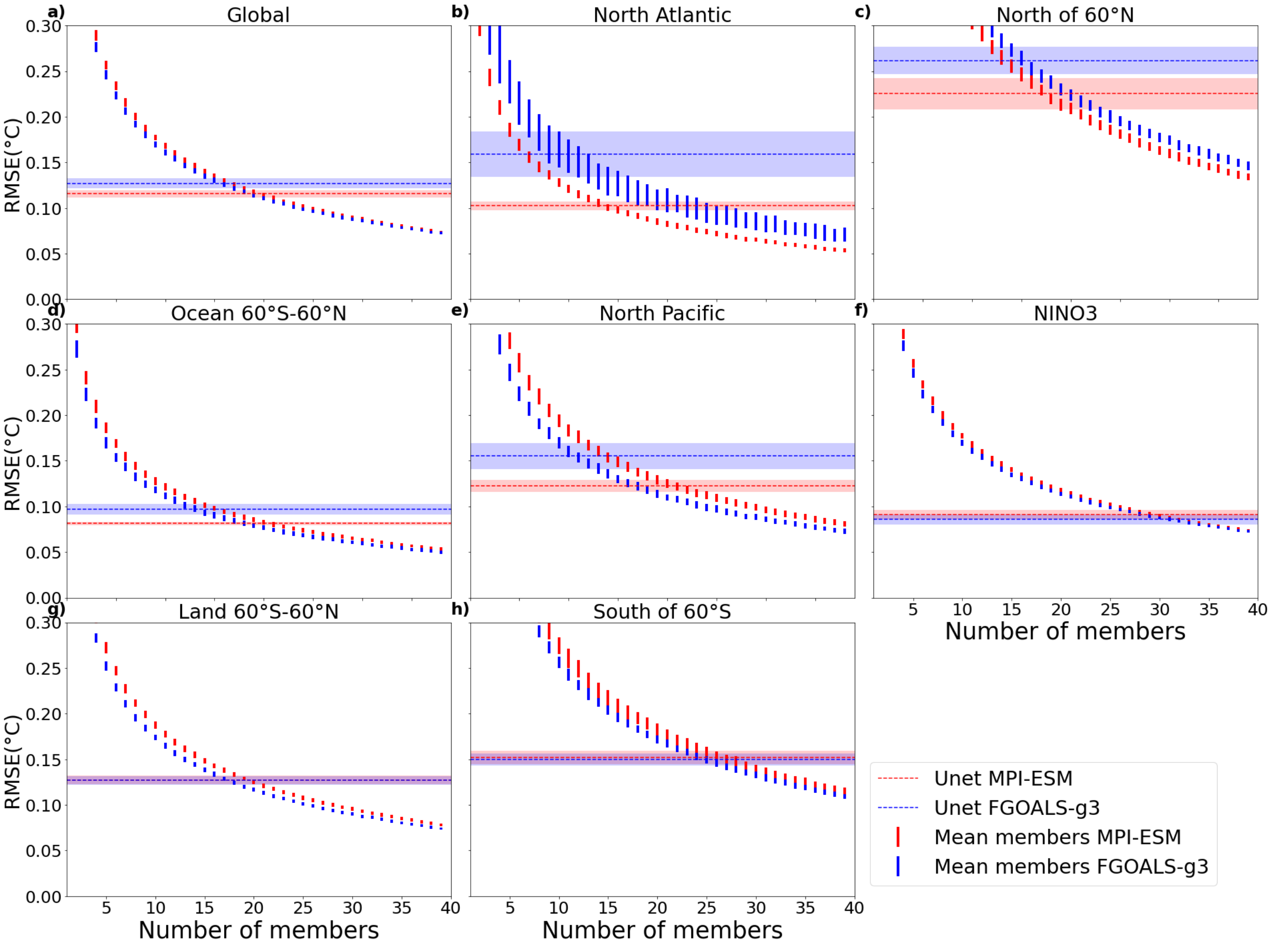


**Figure 7.** Spatial average of the RMSE for the forced variability estimated with the U-Net outputs obtained from each ensemble member, and the forced variability obtained with ensemble averages subsampling ensemble of size 1 to 40; for (red) MPI-ESM and (blue) FGOALS-g3. The RMSE calculated from the 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.

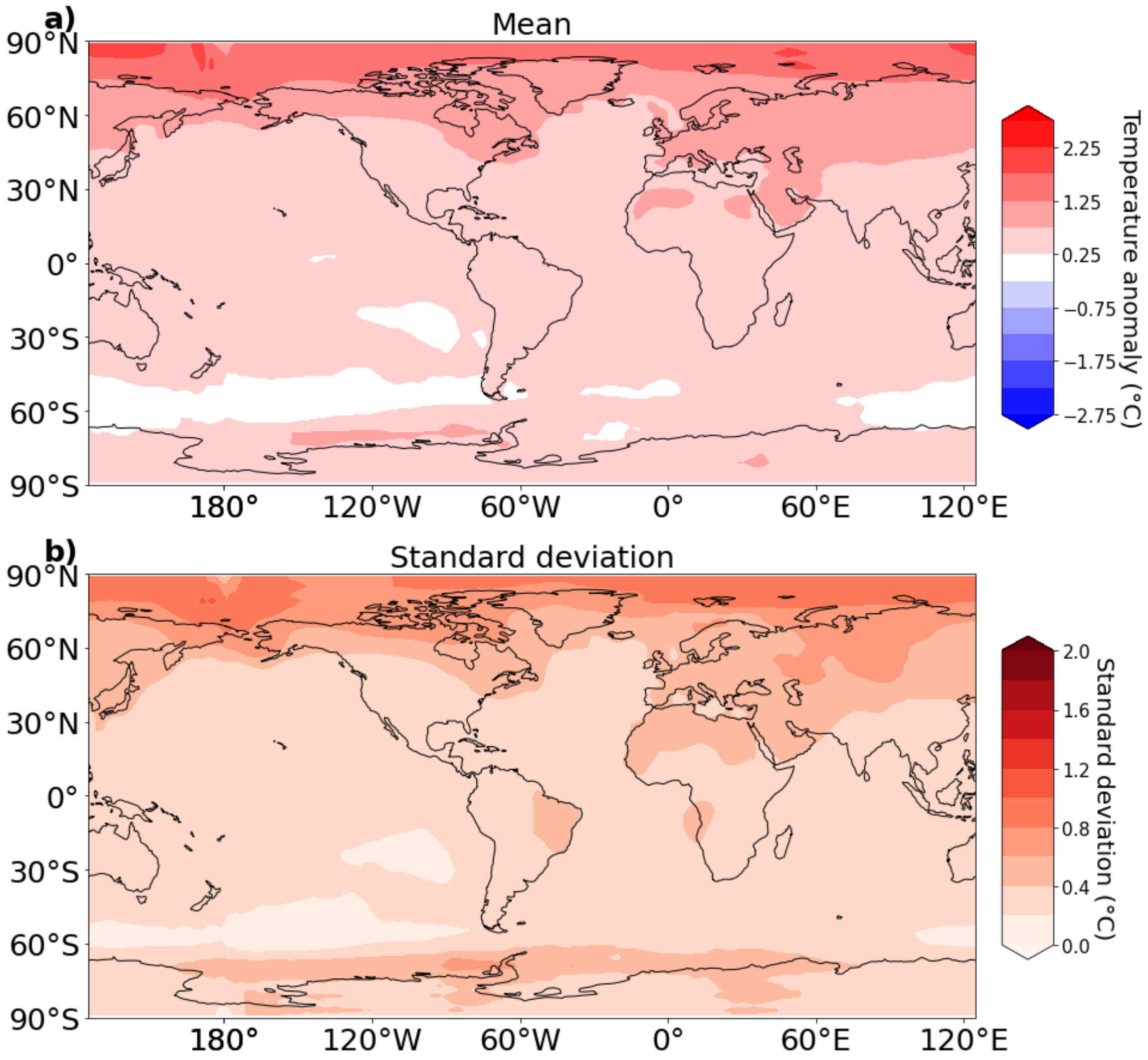


**Figure 8.** Forced surface air temperature (in °C) anomaly when applying the U-Net to GISSTEMP observation : a) time average in 1996-2016; b) standard deviation in 1905-2016.

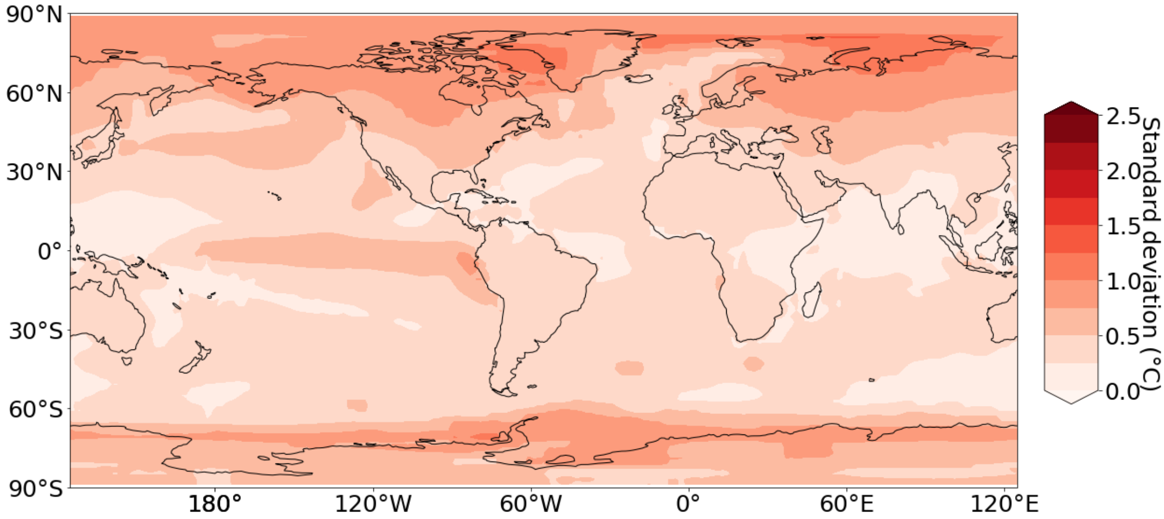


**Figure 9.** Standard deviation of the SAT deviations from the forced SAT, as estimated using the U-Net, in 1905-2016.

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

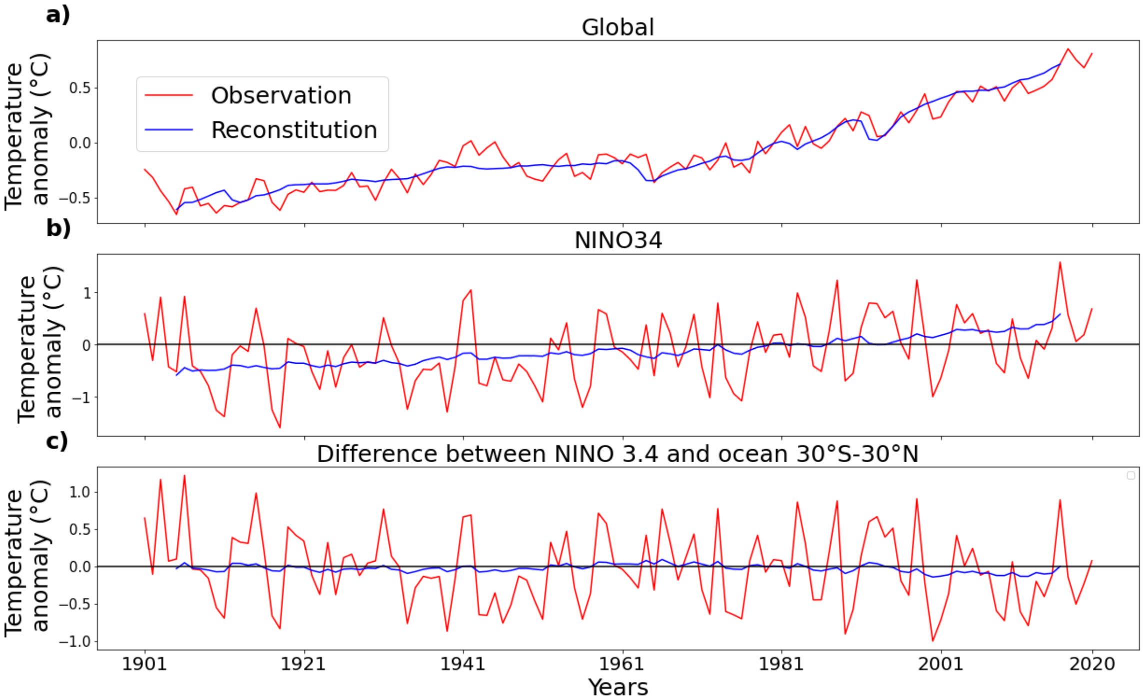


**Figure 10.** Time series of (red) the observed SAT anomaly and (blue) the forced SAT anomaly estimated by the U-Net for a) the global mean b) NINO 3.4 and c) the relative SAT, calculated as the difference between the averaged SAT in Niño 3.4 region and the tropical ocean SAT (30°S-30°N).

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. Improving 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.