JGR Atmospheres

RESEARCH ARTICLE
10.1029/2023JD039202

Causally-Informed Deep Learning to Improve Climate Models and Projections

Fernando Iglesias-Suarez[1], Pierre Gentine[2,3], Breixo Solino-Fernandez[1], Tom Beucler[4], Michael Pritchard[5,6], Jakob Runge[7,8], and Veronika Eyring[1,9]

[1]Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institute of Atmospheric Physics, Oberpfaffenhofen, Germany, [2]Department of Earth and Environmental Engineering, Center for Learning the Earth with Artificial intelligence and Physics (LEAP), Columbia University, New York, NY, USA, [3]Earth and Environmental Engineering, Earth and Environmental Sciences, Learning the Earth with Artificial intelligence and Physics (LEAP) Science and Technology Center, Columbia University, New York, NY, USA, [4]University of Lausanne, Institute of Earth Surface Dynamics, Lausanne, Switzerland, [5]Department of Earth System Science, University of California, Irvine, CA, USA, [6]NVIDIA Corporation, Santa Clara, CA, USA, [7]Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institute of Data Science, Jena, Germany, [8]Technische Universität Berlin, Institute of Computer Engineering and Microelectronics, Berlin, Germany, [9]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

Abstract Climate models are essential to understand and project climate change, yet long-standing biases and uncertainties in their projections remain. This is largely associated with the representation of subgrid-scale processes, particularly clouds and convection. Deep learning can learn these subgrid-scale processes from computationally expensive storm-resolving models while retaining many features at a fraction of computational cost. Yet, climate simulations with embedded neural network parameterizations are still challenging and highly depend on the deep learning solution. This is likely associated with spurious non-physical correlations learned by the neural networks due to the complexity of the physical dynamical system. Here, we show that the combination of causality with deep learning helps removing spurious correlations and optimizing the neural network algorithm. To resolve this, we apply a causal discovery method to unveil causal drivers in the set of input predictors of atmospheric subgrid-scale processes of a superparameterized climate model in which deep convection is explicitly resolved. The resulting causally-informed neural networks are coupled to the climate model, hence, replacing the superparameterization and radiation scheme. We show that the climate simulations with causally-informed neural network parameterizations retain many convection-related properties and accurately generate the climate of the original high-resolution climate model, while retaining similar generalization capabilities to unseen climates compared to the non-causal approach. The combination of causal discovery and deep learning is a new and promising approach that leads to stable and more trustworthy climate simulations and paves the way toward more physically-based causal deep learning approaches also in other scientific disciplines.

Plain Language Summary Climate models have biases compared to observations that have been present for a long time because certain processes, like convection, are only approximated using simplified methods. Neural networks can better represent these processes, but often learn incorrect connections leading to unreliable results and climate model crashes. To solve this, we used a new method that informs neural networks with causal drivers, therefore, respecting the underlying physical processes. By doing so, we developed more reliable and trustworthy neural networks, allowing us to accurately represent the climate of the original high-resolution simulation on which these neural networks were trained.

Key Points:

- Causal discovery unveils causal drivers of subgrid-scale processes across climates
- The causally-informed hybrid model runs stably and generates a climate close to the original high-resolution simulation
- Spurious correlations are evident in the non-causal parameterization, leading to underestimate feature importance of physical drivers

Supporting Information:

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

Correspondence to:

F. Iglesias-Suarez,
fernando.iglesias-suarez@dlr.de

Citation:

Iglesias-Suarez, F., Gentine, P., Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J., & Eyring, V. (2024). Causally-informed deep learning to improve climate models and projections. *Journal of Geophysical Research: Atmospheres*, *129*, e2023JD039202. https://doi.org/10.1029/2023JD039202

Received 3 MAY 2023
Accepted 26 JAN 2024

Author Contributions:

Conceptualization: Fernando Iglesias-Suarez, Pierre Gentine, Breixo Solino-Fernandez, Tom Beucler, Michael Pritchard, Jakob Runge, Veronika Eyring
Data curation: Tom Beucler
Formal analysis: Fernando Iglesias-Suarez, Breixo Solino-Fernandez
Funding acquisition: Pierre Gentine, Veronika Eyring
Investigation: Fernando Iglesias-Suarez
Methodology: Fernando Iglesias-Suarez, Pierre Gentine, Breixo Solino-Fernandez, Tom Beucler, Michael Pritchard, Jakob Runge, Veronika Eyring

1. Introduction

Our understanding of the climate system and how it may change in the future under different scenarios has greatly improved thanks to climate models (IPCC, 2021a). Yet, systematic biases still plague current climate models compared to observations (Eyring, Gillett, et al., 2021; Flato et al., 2013) and limit their ability to accurately project climate change at global and regional scales (Lee et al., 2021; Tebaldi et al., 2021). Many important processes determining the Earth's climate occur at scales smaller than current climate models grid size, typically ranging 50–100 km horizontally (IPCC, 2021b). The effect of these subgrid-scale or unresolved processes, such as clouds and convection, on the system are approximated via physical parameterizations in current models, and are a key source of the uncertainty in climate projections (Gentine et al., 2021; Schneider et al., 2017).

Resources: Michael Pritchard, Veronika Eyring
Validation: Fernando Iglesias-Suarez
Visualization: Fernando Iglesias-Suarez
Writing – original draft: Fernando Iglesias-Suarez
Writing – review & editing: Fernando Iglesias-Suarez, Pierre Gentine, Breixo Solino-Fernandez, Tom Beucler, Michael Pritchard, Jakob Runge, Veronika Eyring

High-resolution storm-resolving models (SRMs), run at a horizontal scale of few kilometers, explicitly represent deep convection and dynamics of convective storms, and alleviate a number of biases present in coarser climate models (Bock et al., 2020; Sherwood et al., 2014). For instance, they show improvements in representing the Intertropical Convergence Zone (ITCZ) (Klocke et al., 2017), storm-tracks and precipitation (Stevens et al., 2019), as well as subseasonal variability (Rasp et al., 2018). Accurate representation of convective processes is also essential to capture Earth's system feedbacks in a changing climate, like cloud-radiation feedbacks (Bony et al., 2015). Yet, global SRMs simulations are only possible over short periods of time (months) due to their staggering computational costs, whereas climate research requires a number of realizations over hundreds of years (Schneider et al., 2017).

Machine learning (ML) approaches, and in particular deep learning (DL) methods, have shown great potential in learning explicitly resolved small-scale processes such as deep convection from SRM simulations (Eyring, Mishra, et al., 2021; Gentine et al., 2018, 2021; Grundner et al., 2022) and represent them in coarser resolution models. Hybrid models, that is, ML-based parameterizations coupled to a host climate model, have shown great performance simulating the climate of the original SRM in terms of mean state and its variability (e.g., tropical waves) (Bretherton et al., 2022; Rasp et al., 2018; Wang et al., 2022; Watt-Meyer et al., 2021; Yuval & O'Gorman, 2020). ML methods for Earth system modeling is an active area of research. Particularly challenging to address are the poor representation of unseen climates and regimes (i.e., generalization capabilities) (Grundner et al., 2022; O'Gorman & Dwyer, 2018; Scher & Messori, 2019), and hybrid modeling instabilities associated with the interaction between the ML-based algorithm and the dynamical core of the host climate model (Brenowitz, Beucler, et al., 2020). Training the ML algorithm directly online, coupled to the climate model (Frezat et al., 2022; Lopez-Gomez et al., 2022), crude ablation (Brenowitz & Bretherton, 2019), and deeper DL algorithms (Rasp et al., 2018) have been proposed to overcome hybrid ML modeling instabilities. Yet, the causes of such instabilities are not fully understood. Our working hypothesis is that ML-based parameterizations can accurately reproduce subgrid-scale processes using non-causal relationships (i.e., mere correlations), and these correlations might be overfitting the training dataset (Brenowitz, Henn, et al., 2020). In a nutshell, ML algorithms can skillfully learn a given task for the wrong reasons using spurious non-physical relationships, but may struggle in conditions deviating from the training data in which causes and effects might differ from the initial correlations present in the training data (Brenowitz, Henn, et al., 2020).

Integrating domain knowledge in the form of causal relationships (i.e., inductive bias) is a recent and emerging theme in ML research (Pearl & Mackenzie, 2018; Runge, Bathiany, et al., 2019; Schölkopf et al., 2021) to overcome shortcomings of standard ML methods, which predictions are mostly based on correlations between predictors and predictands. While correlation is a statistical relationship between two variables (i.e., where a change in one variable is associated with a change in the other variable), causality is the relationship between an action (the cause) and its outcome (the effect). As an example, for data coming from a causal model $X^1 \leftarrow X^2 \rightarrow Y$, an ML algorithm may learn to predict Y from both X^1 and X^2. However, such a prediction would fail if the ML method is employed under changing environments where the confounder X^1 is in a different state. This can easily be the case when an ML algorithm learns an atmospheric physical process. Owing to the strong correlation induced by convective processes in the atmospheric profile environment, surface precipitation, for example, can be influenced not only by conditions in the troposphere but also by upper tropospheric moisture (confounder) (Brenowitz & Bretherton, 2019). Causal discovery methods aim to discover such causal relationships from data (Runge, Bathiany, et al., 2019; Runge, Nowack, et al., 2019), going beyond simple correlations. The goal of our study is to merge the power of causal discovery with the data exploitation capacity of neural networks, and investigate whether such a causally-informed neural network can help better understand and predict (subgrid) physical mechanisms in the atmosphere.

We build on existing data-driven subgrid parameterization work (Rasp et al., 2018) and combine causal discovery and deep learning, using the same high-resolution modeling data, to improve DL-based parameterizations. The task of the causal discovery algorithm is to unveil the causal drivers of the subgrid-scale processes respecting the underlying physical mechanisms. The causally-informed neural network algorithms have two steps. We first identify causal drivers of subgrid-scale processes using a causal discovery method based on conditional independence tests. Then we build novel causally-informed neural network algorithms, in which subgrid-scale processes are learned from the causal drivers. In other words, we build sparser (lower dimensional) neural networks in which non-causal connections are dropped. We demonstrate several key aspects of this novel method: (a) causal discovery removes spurious links; (b) causal drivers are "climate invariant" (i.e., robust across colder and

Table 1
Summary of the Model Simulations

Climate models	Parameterizations	Causal-threshold
SPCAM	SP component (2–D SRM)	—
Non-causalNNCAM	Non-causalNN	—
$Causal_{0.59}$NNCAM	Causally-informed$_{0.59}$NN	quantile 0.59[a]
CausalNNCAM	Causally-informedNN	quantile optimized[b]

[a]Single optimized causal-threshold for all outputs. [b]Varying optimized causal-threshold (see Supporting Information S1).

warmer climates (Beucler et al., 2021)); and (c) the causally-informed hybrid model accurately represents the climate of the original high-resolution climate model, retaining many convection-related properties. We finish by discussing a potential broader role of causal discovery in the context of ML for physical sciences, and key remaining challenges for future work.

2. Causally-Informed Hybrid Modeling

We extend previous work (Gentine et al., 2018; Rasp et al., 2018) to build a causally-informed hybrid climate model (see Table 1). Figure 1 shows a schematic overview of the causally-informed neural network approach. We use an aquaplanet (i.e., ocean only without topography) setup of the Super-parameterized Community Atmosphere Model v3.0 (SPCAM) (Collins et al., 2006). The model extends from the surface to the upper stratosphere (3.5 hPa) with 30 vertical levels and includes a horizontal resolution of 2.8° × 2.8° (latitude by longitude). Stationary—no seasonality but with diurnal cycle—zonal mean sea surface temperatures are imposed using a realistic equator-to-pole gradient (Andersen & Kuang, 2012). The time step of the climate model component (CAM) is 30 min. The superparameterization component (SP) is a 2–D SRM embedded in each grid column, explicitly resolving most of deep convection but parameterizing turbulence and cloud microphysics (M. F. Khairoutdinov & Randall, 2001; Pritchard et al., 2014; Pritchard & Bretherton, 2014). For consistency with the former study (Rasp et al., 2018), the SP component uses eight 4–km-wide meridionally oriented columns (west to east), and time steps of 20 s. SPCAM alleviates a number of climate model biases (Oueslati & Bellon, 2015), including a more realistic Madden-Julian oscillation and a single ITCZ, as well as better representation of precipitation extremes (Arnold & Randall, 2015; Benedict & Randall, 2009; Kooperman et al., 2016a, Kooperman et al., 2016b, 2018).

The task of the neural networks (NNs) is to learn subgrid-scale processes (output predictands) as represented by the SP component given the environmental conditions (input predictors) of the climate model, CAM. The training data are column-based values of the model's subgrid physics package, which includes the SP subgrid resolution of

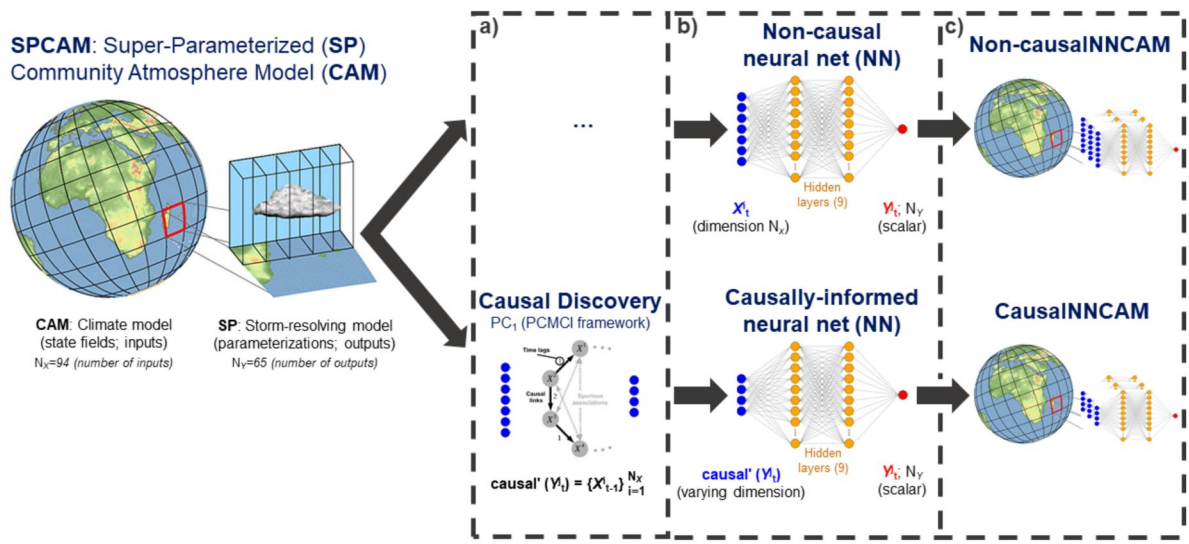

Figure 1. Schematic overview of the causally-informed neural network approach. The Superparameterized Community Atmosphere Model (SPCAM) is used to learn subgrid-scale processes (Y) as represented by the SP component given the environmental conditions (X) of the climate model, CAM. Two data-driven parameterizations are considered: (top) non-causal; and (bottom) causally-informed. For the causally-informed approach, (a) we use a causal discovery algorithm, the PC_1 phase of the PCMCI method (Runge, Nowack, et al., 2019), to prune the fully connected input vector and eliminate non-physical spurious inputs and connections. Therefore, (b) we develop 65 separate single-output NNs, each one having a specific subset of causal inputs, $causal'\left(Y_t^j\right)$, with a varying input vector length. The main difference between the analogous Causally-informedNN and Non-causalNN parameterizations is that the latter always includes an input vector of length 94 (i.e., full set of environmental conditions). Finally, (c) we couple both the Non-causalNN parameterization and the new Causally-informedNN parameterization of SPCAM physics within CAM, resulting in the Non-causalNNCAM and CausalNNCAM models respectively (Table 1). The causal discovery diagram was adapted from Runge, Nowack, et al. (2019) and the sketch of SPCAM from Prein et al. (2015).

convective transport, turbulence and radiation scheme, with a few omissions (condensed water species and ozone). The inputs (X) are column-wise values of temperature $T(p)$ [K], specific humidity $q(p)$ [kg kg^{-1}], and meridional wind $V(p)$ [m s^{-1}] at each column level, surface pressure P_{srf} [Pa], incoming solar radiation Q_{sol} at the top of the atmosphere, as well as sensible- and latent-heat fluxes at the surface, Q_{sen} and Q_{lat} in [W m^{-2}], respectively. The outputs (Y) comprise: heating tendencies $\Delta T_{phy}(p)$ (including convection and radiative heating rates) [K s^{-1}]; and moistening tendencies $\Delta q_{phy}(p)$ [kg kg^{-1} s^{-1}] at each model level; net shortwave and longwave radiative heat fluxes at the model top and at the surface (Q_{sw}^{top}, Q_{lw}^{top}, Q_{sw}^{srf} and Q_{lw}^{srf} respectively) [Wm^{-2}], and surface precipitation P [kg m^{-2} d^{-1}]. Only the heating and moistening tendencies are coupled to the climate model's dynamical core, with the other outputs being diagnostics. Two data-driven DL parameterizations are considered: non-causal; and causally-informed (see Parameterizations column in Table 1). In both cases, and based on previous work (Rasp et al., 2018), we use fully connected feedforward NNs, with nine hidden layers and 256 units per layer (around 0.5 million parameters). The NNs are optimized by minimizing the loss defined as the mean squared error of the prediction compared to the SPCAM "truth" value. We use 3 months of an SPCAM simulation for training, validation and test sets (each set approximately including 45 million samples using every model time step and grid column). Note the training set is shuffled in time and space (grid columns). See Supporting Information S1 for additional details.

We develop 65 separate single-output NNs for both cases, causal and non-causal parameterizations (corresponding to 30 vertical levels for heating and moistening rates, plus net shortwave and longwave radiative heat fluxes at the model top and at the surface, and surface precipitation), rather than a single NN for the entire column (Han et al., 2020). In this way, each causally-informed NN has a specific subset of causal inputs (drivers) obtained during the causal discovery phase, and therefore a varying input vector length (see below). The main difference between the analogous causally-informed NNs and non-causal NNs is that the latter always include an input vector of length 94 (i.e., 3 times the number of levels for temperature, humidity, and wind, as well as 2–D fields of surface pressure, sensible and latent heat fluxes, and top of the atmosphere incoming solar radiation) (see Supporting Information S1). For the causally-informed NNs, our goal is to prune the fully connected input vector to eliminate non-physical spurious inputs and connections. From a causal perspective, the setup here is simplified because the inputs and outputs are known to be separated in time based on the SP climate model's structure. Hence, we can utilize a causal discovery selection algorithm that removes those inputs that are conditionally independent of the output, thus, providing no additional information to predict the output.

While no causal discovery method is infallible, we can gain a deeper understanding of how a dynamical system works and develop more informed algorithms based on causal evidence rather than simply relying on correlations or associations between variables (i.e., environmental conditions driving subgrid-scale processes). Our selection algorithm is the PC$_1$ phase of the PCMCI method (Runge, Nowack, et al., 2019), which is based on the PC algorithm (Spirtes & Glymour, 1991), and the Momentary Conditional Independence (MCI) test. PCMCI, and its different flavors, have been widely used in recent years in climate sciences, such as for better understanding teleconnections in the Earth system (Kretschmer et al., 2016, 2018; Runge et al., 2014; Siew et al., 2020) and their pathways (Galytska et al., 2023; Karmouche et al., 2023; Kretschmer et al., 2021; Runge et al., 2015) or to investigate land-atmosphere interactions (Krich et al., 2020).

PC$_1$ starts by a fully connected matrix between all inputs and outputs. This initializes the causal drivers to $causal_g\left(Y_t^j\right) = \{X_{t-1}^i\}_{i=1}^{N_X}$, where N_X is the number of all potential drivers across the different vertical levels and g refers to each column of the model grid. Then PC$_1$ tests whether each input $\left(X_{t-1}^i\right)$ is conditionally independent of an output $\left(Y_t^j\right)$ given selected subsets of the other correlated inputs in the dataset $\left(causal_g\left(Y_t^j\right)\right)$. If two variables are found to be conditionally independent, it is inferred that there is no direct causal relationship between them. Specifically, it removes drivers X_{t-1}^i from $causal_g\left(Y_t^j\right)$ if they are conditionally independent (irrelevant or redundant) of Y_t^j given subsets $S_k \subset causal_g\left(Y_t^j\right)$ whose cardinality k iteratively increases. For $k = 0$, all X_{t-1}^i with $X_{t-1}^i \perp Y_t^j$, where $\perp$ refers to conditional independence, are removed. For $k = 1$, those with $X_{t-1}^i \perp Y_t^j | S_1$ are removed, where S_1 is the strongest driver (as measured by their absolute partial correlation value) from the previous step. For $k = 2$, those with $X_{t-1}^i \perp Y_t^j | S_2$ are removed, where S_2 are the two strongest drivers (regarding the individual absolute partial correlation values), among the remaining drivers, excluding X_{t-1}^i, from the previous

step. A simple forward-selection method would always keep the strongest driver at each iteration step, while our approach re-tests them conditional on the remaining k strongest drivers. This procedure continues until the algorithm converges and stops when there are no more possible combinations S_k, that is, if the cardinality of S_k is equal to the number of remaining drivers. We note that the iterative process of repeatedly testing each potential causal driver of an output against efficiently chosen subsets of the other potential inputs is what establishes PC_1 as causal under the assumptions discussed below. Here conditional independence is tested using partial correlation because we may reasonably assume that at a short 30 min time-scale, even non-linear relations are sufficiently well captured by a linear model. The independence test is based on a standard significance level $\alpha_{pc} = 0.01$.

Our goal here is primarily to remove spurious inputs. A causal interpretation of $causal_g\left(Y_t^j\right)$ using the above algorithm rests on the following assumptions: causal sufficiency (all common causes are observed), the Markov condition (dependence must be due to causal connectedness), faithfulness (independencies are not by coincidence but structural, therefore, follow the Markov condition). This approach yields a set of causal drivers $causal_g\left(Y_t^j\right)$ for each variable Y^j at every column g and every vertical level. Because we want these drivers to generalize across all columns (across space and time), we define "robust" causal drivers as $causal'\left(Y_t^j\right) = \left\{X_{t-1}^i : P\left(X_{t-1}^i \rightarrow Y_t^j \in causal_g\left(Y_t^j\right)\right) > q\right\}$. Specifically, we only consider causal drivers X_{t-1}^i for a given output Y_t^j, those whose probability P of being causally-linked to the output across the 8,192 latitude and longitude columns is at least quantile q. This is also the main idea behind the invariant causal prediction approach (Peters et al., 2016). We explore two cases to choose q: a single optimized quantile-threshold of $q = 0.59$ for all outputs; and a varying quantile-threshold optimized for each output separately (quantile-threshold optimization is based on causally-informed NNs offline performance; see Supporting Information S1). q is the primary hyperparameter of the algorithm. A too loose threshold would lead to keeping non-physical spurious inputs similar to a fully connected (non-causal) feedforward NN. A too strict threshold would lead to neglecting some important input features (key drivers) and thus having poor predictive skills. The results presented here are based on the varying quantile-threshold (see Supporting Information S1 for the single optimized quantile-threshold results).

Finally, to test the value of our causal discovery approach (PC_1) in removing potential spurious inputs-to-outputs links, we explore two additional feature selection approaches. First, we compare PC_1 with a baseline (non-causal) correlation method, which follows the column-wise approach. Second, linear Lasso regression is applied to the entire training set without separating each location (Tibshirani, 1996), which under the above causal assumptions may also work as a complementary method for selecting causal features. Although we choose in this work the PC_1 algorithm and explore linear Lasso regression, we note there are a number of causal discovery methods that may well be suitable (Runge et al., 2023).

3. Results

3.1. Causal Discovery

Based on the causal discovery algorithm, we can investigate the main drivers of the subgrid-scale processes as represented by the SPCAM model. Causal drivers of SPCAM's parameterizations inferred by our causal discovery algorithm are in agreement with current physics understanding (Figure 2; Figure S1 in Supporting Information S1). Causal matrices, similar to transilient matrices (Romps & Kuang, 2011; Stull, 1993), reveal largest coefficients—ratio of the causal drivers appearance across the model's grid—on the diagonal, meaning that key direct drivers are primarily local in the vertical. This is especially strong for humidity, which is known to be a key regulator of convective mixing and updraft buoyancy through lateral entrainment (de Rooy et al., 2013; Stommel, 1951; Warner, 1970). Nonetheless, the causal matrices reveal the influence of the lower troposphere ($p > 600$ hPa) on the profile. This is expected since convective processes and buoyant plumes originate from the boundary layer, affecting the troposphere above. In particular, boundary-layer and lower troposphere temperature is causally-linked (driver) to moistening and heating rates throughout the troposphere. Several studies have shown that cold pools induced by unsaturated downdrafts organize the boundary layer (Del Genio & Wu, 2010; Kuang & Bretherton, 2006; M. Khairoutdinov & Randall, 2006; Mapes & Neale, 2011), and organizing convection leads to changes in atmospheric heating and moistening tendencies and precipitation (Muller & Bony, 2015; Tompkins, 2001). Furthermore, heating rates in the upper-troposphere and lower-stratosphere (~100–300 hPa) are

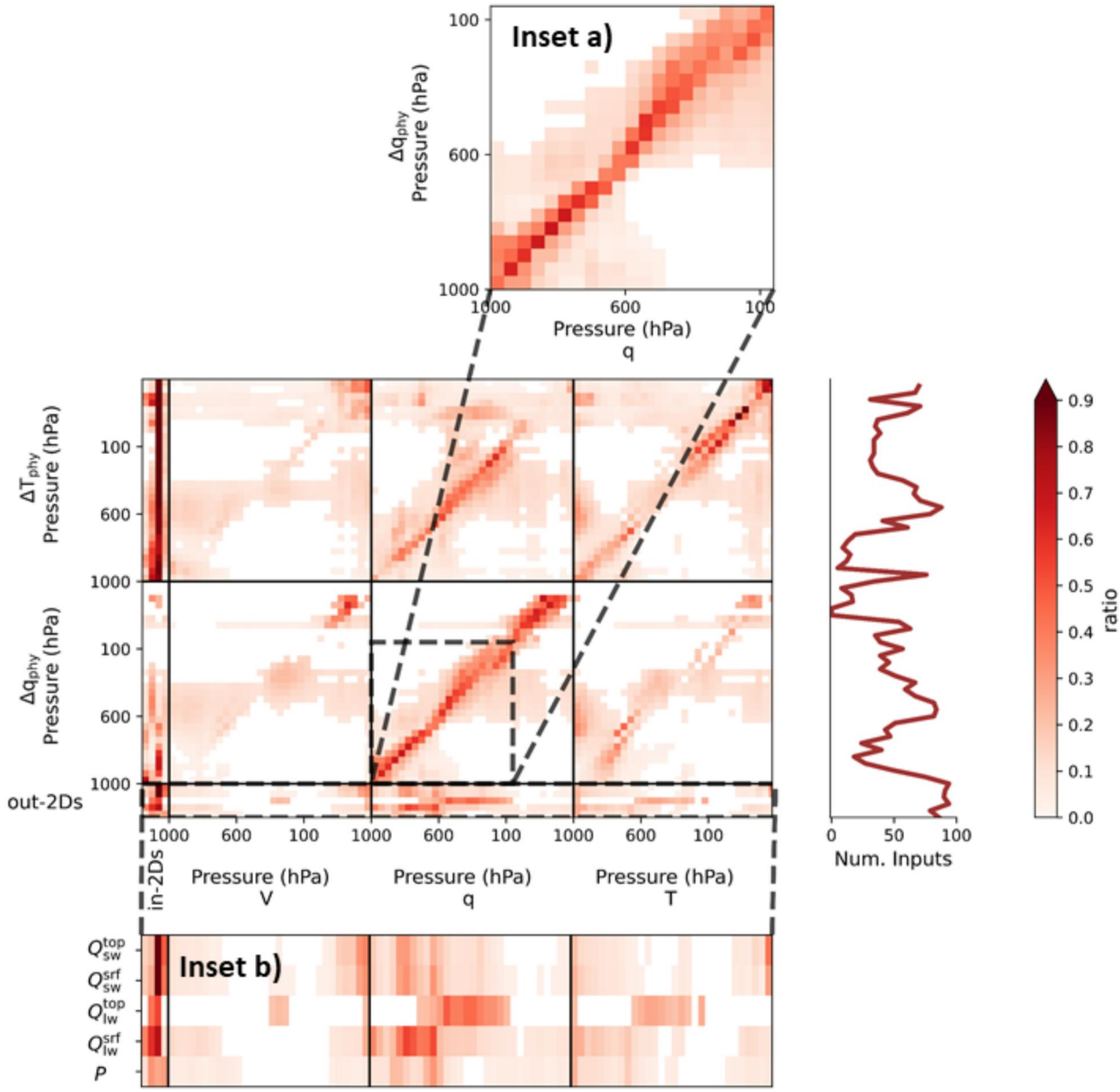

Figure 2. Causal discovery feature selection matrix of subgrid-scale processes in SPCAM for the varying optimized causal-threshold per output (see Supporting Information S1). The inputs of the neural networks are given on the *x*-axes: 2–D variables (Q_{lat}, Q_{sen}, Q_{sol}, P_{srf}); and 3–D variables (V, q, T) from the surface (10^3 hPa) to the model's top (3 hPa), respectively. The outputs (subgrid-scale processes) are represented on the *y*-axes: 2–D variables (P, Q_{lw}^{top}, Q_{sw}^{top}, Q_{lw}^{srf}, and Q_{sw}^{srf}); and 3–D variables (Δq_{phy} and ΔT_{phy}) from the surface to the model's top, respectively. Insets (a) and (b) zoom-in into Δq_{phy} and 2–D output variables, respectively. Contour colors represent the coefficient (absolute ratio) of the causal drivers appearance throughout the model's grid. Right panel shows the number of causal drivers for each output, with a mean number of inputs of 48 (51% of the total).

associated with mid-tropospheric moistening, where deep convection can have a substantial radiative effect due to cirrus clouds. Incoming solar radiation is the most important driver of heating rates because of its regulation of the diurnal cycle. Interestingly, precipitation at the surface is strongly causally-linked by environmental conditions from the lower to the middle troposphere, and therefore associated with convective processes and rain re-evaporation, consistent with the strong relationship between precipitation and the bulk temperature and moisture in the lower troposphere (Ahmed & Neelin, 2018; Del Genio, 2012; D'Andrea et al., 2014).

Our causal feature selection methodologies outperform a more naïve baseline feature selection approach removing potential spurious inputs-to-outputs links. Using a simple correlation method, we find that correlations

across the atmospheric profile are largely non-local and redundant among state fields (i.e., less physical; see Figure S2 in Supporting Information S1). This is due to the large inter-correlation in the atmospheric profile associated with convective processes. These strong correlations across levels would nonetheless include potential spurious links and primarily define the strength of the neural network connections. Moreover, using simple correlations to optimize the connectivity matrices is challenging, since either a number of outputs lack input links (e.g., upper tropospheric moisture; Figure S2 in Supporting Information S1), or the system is quasi-fully connected (not shown). Furthermore, we use linear Lasso regression (Tibshirani, 1996), which under the above causal assumptions (sufficiency, Markov, and faithfulness) may work as a causal feature selection method. While selected features show largest Lasso coefficients on the diagonal (Figure S3 in Supporting Information S1), meaning it captures that key direct physical drivers are primarily local in the vertical, there are clear spurious features (e.g., moistening and heating tendencies in the lower troposphere are associated with environmental conditions in the stratosphere; see also Section 3.3 and Figure S9 in Supporting Information S1). While the correlationally-informed parameterization performs sub-optimally compared to the reference non-causal case (particularly in the lower troposphere), lasso-informed parameterization shows in general similar skills (Figure S7 in Supporting Information S1). Causal discovery is arguably a more complex method that rests on some expert knowledge of the physical problem (allowing us to choose a suitable causal algorithm and its setup), and on a number of mathematical assumptions. However, it goes beyond standard feature selection approaches and helps further remove spurious links, as demonstrated by PC_1 and linear Lasso regression.

As a further test of the credibility of our causal feature selection methodology and its stability with a changing input distribution, we also explore its sensitivity to climate change (Galytska et al., 2023; Karmouche et al., 2023). Thermodynamic features driving different atmospheric processes are "climate invariant", that is, they govern the same processes regardless of the climate state of the system, as physics does not change with climate change. For example, whatever the state of the climate system, we expect the key direct drivers of heating and moistening tendencies to remain local, though deep convection affects them non-locally throughout the troposphere. Reassuringly, we find that causal drivers of subgrid-scale processes as inferred by our causal discovery algorithm in SPCAM are consistent across climates for both, global 4 K sea surface temperatures cooling and warming (only around 5% inconsistent non-local causal drivers in the vertical, Figure 3). This result suggests that causal discovery helps unveil the most direct key drivers of smaller-scale processes represented by SPCAM and remove some of the confounding effects present when using neural networks, that would otherwise affect their performance due to spurious inter-correlations across the vertical profile (shown below).

3.2. Mean Climate and Variability

We here couple both the standard NN emulation (Non-causalNN) and the new causally-informed NN emulation (Causally-informedNN) of SPCAM physics within CAM, resulting in the Non-causalNNCAM and CausalNNCAM models respectively (Table 1). We evaluate their response when run online (i.e., coupled to the coarse-resolution model), using zonal-mean daily averaged output from 1 year prognostic runs with a 1 month spin-up (after reaching climate equilibrium; not shown). Note that instabilities generally would develop within the first 6 months (Ott et al., 2020).

The CausalNNCAM model accurately represents the mean tropospheric temperature and variability of the original model (Figure 4) while the Non-causalNNCAM model shows significant cold biases in the tropics. Key features of the SPCAM simulations used here are the representation of a single ITCZ associated with a primary tropospheric tower of heating rates (warming; Figure 4) and moistening rates (drying; Figure S4 in Supporting Information S1) due to subgrid-scale processes, as well as secondary free-troposphere maxima at midlatitudes storm-tracks. While these features are well represented in CausalNNCAM, a spurious double-ITCZ is clearly represented in the Non-causalNNCAM. We note these biases are not present in the former non-causal hybrid model upon which this work builds, NNCAM (Rasp et al., 2018), which represents very well the mean climate of the original SPCAM simulation (see Section 4.2 and Figure S11 in Supporting Information S1). It is also worth noting that the $\text{Causal}_{0.59}$NNCAM simulation shows similar deficiencies in the troposphere as in the non-causal realization. This suggests that a single optimized causal-threshold may well be too strict for a number of subgrid-scale processes (output predictands), which results in neglecting key physical drivers (input predictors) (see Figures S5–S6 in Supporting Information S1). Stratospheric temperature biases are evident in all prognostic simulations with DL-based parameterization, and are very likely associated with the important role of ozone

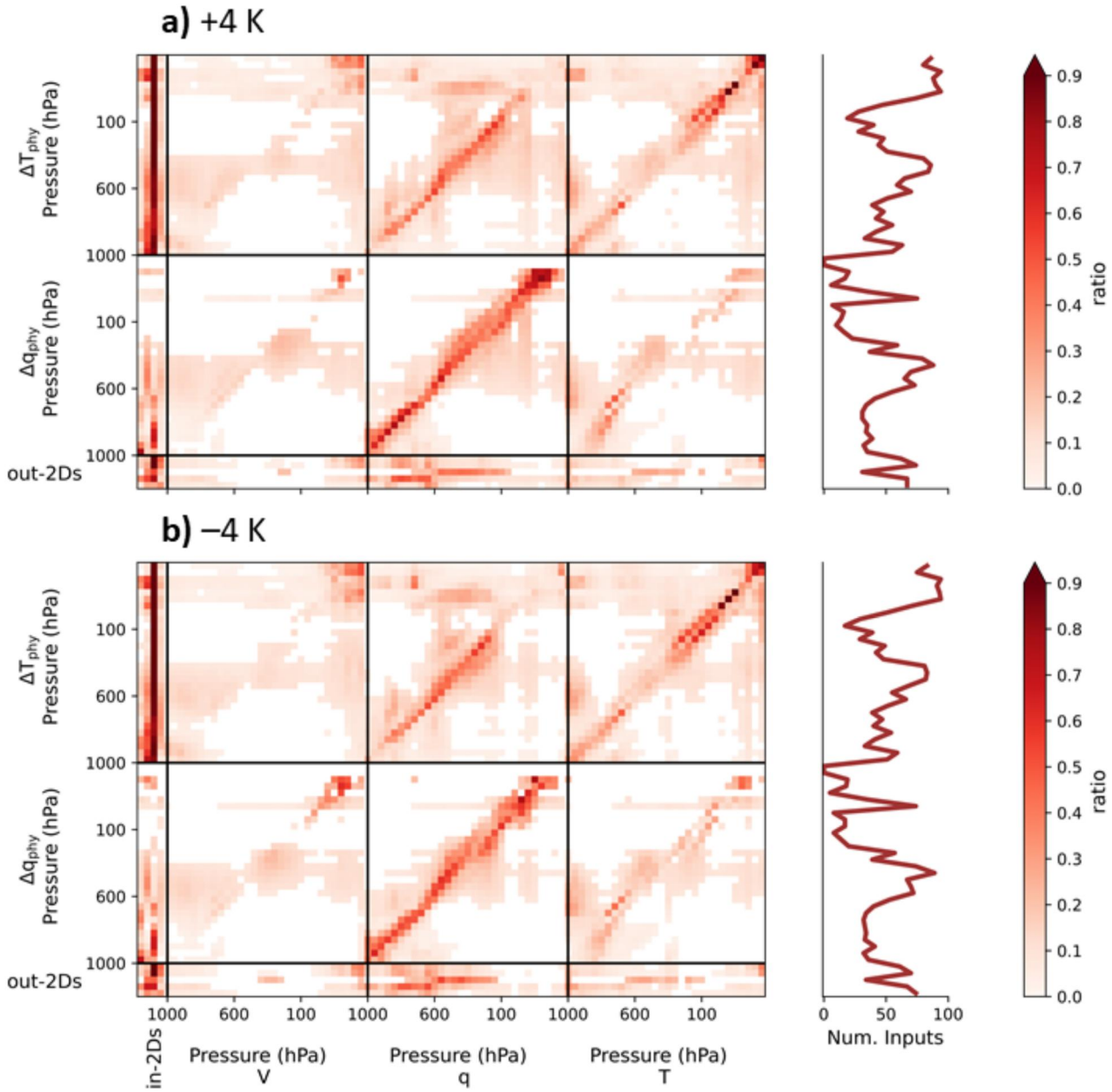

Figure 3. Same as Figure 2, but for (a) warming and (b) cooling of global 4K sea surface temperatures by adding a wavenumber one perturbation to the reference sea surface temperatures in increments of 1 K. Right panels show the number of causal drivers for each output. Both cases have an equivalent mean number of inputs of 48 (51% of the total) compared to the reference climate (+0 K).

(missing variable as a predictor in our NN setup) in determining the climate in the stratosphere (WMO, 2018). These biases were also evident in NNCAM (Rasp et al., 2018).

The better ability of the causally-informed DL-based parameterization, CausalNNCAM, to realistically reproduce the mean ITCZ and characteristic midlatitude storm-track variability of the SPCAM reference simulation is also reflected in surface precipitation and net radiative fluxes (Figure 5), as it accurately resolves the zonal patterns of precipitation and net radiation of SPCAM. In contrast, in Non-causalNNCAM precipitation is substantially underestimated, both mean and variability and a double-ITCZ pathology is evident. CausalNNCAM correctly captures precipitation peaks, though it is somewhat overestimated over the ITCZ and associated with stronger moistening rates (Figure S4 in Supporting Information S1). Similarly, net radiative fluxes at the top of the atmosphere in Non-causalNNCAM are underestimated in the subtropics compared to SPCAM due to the double-

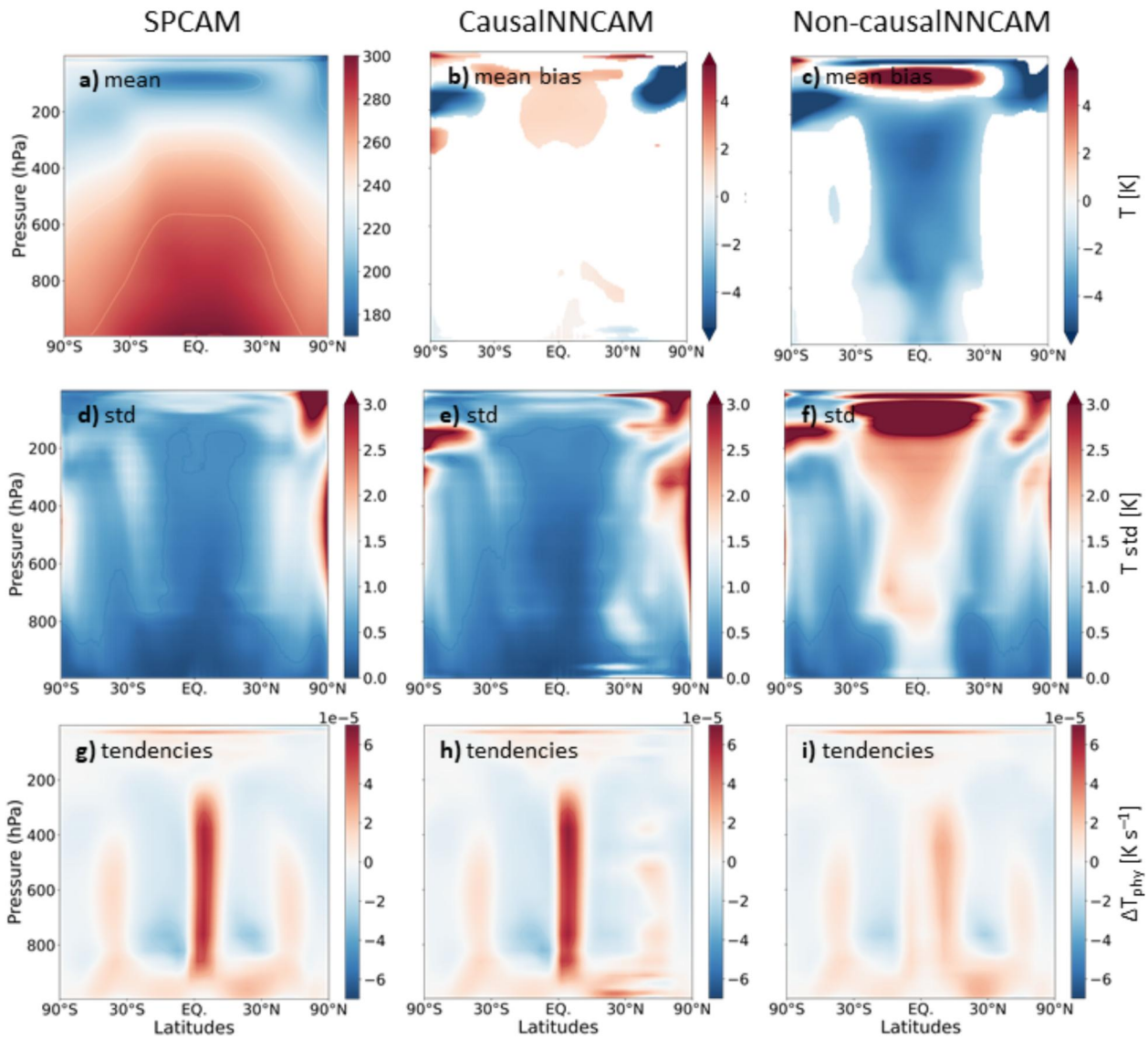

Figure 4. Zonal-mean climatologies of (a)–(c) temperature (T), (d)–(f) T variability (standard deviation; std), and (g)–(i) heating tendencies (ΔT_{phy}) for SPCAM, CausalNNCAM and Non-causalNNCAM. Contour colors for temperature biases (b)–(c) are for statistically significant differences at the 95% confidence interval. Note Non-causalNNCAM biases are not present in NNCAM (Rasp et al., 2018), see Figure S11 in Supporting Information S1.

ITCZ bias, deficiency largely overcome in CausalNNCAM. These results clearly show that CausalNNCAM not only reduces the dimensionality of the DL algorithm, which limits the impact of confounders such as the strong spatial (vertical and zonal) inter-correlations in the atmospheric profiles, but can also match the performance of NNCAM (Rasp et al., 2018). We reiterate that these biases in Non-causalNNCAM were not present in NNCAM (Figure S11 in Supporting Information S1). Nevertheless, both non-causal and causal parameterizations accurately represent the physics of the test set (offline; Figures S7 and S8 in Supporting Information S1), and perform as well as the original NN (Rasp et al., 2018). Details about the disparities between offline and online performances are provided in Section 4.2. In principle, it would be possible to develop a good performing non-causal hybrid model by systematically training a very large number of NNs, as it has been already shown (Rasp et al., 2018).

Causal discovery helps improve DL-based learning of physical processes (i.e., parameterizations) by informing them with causal drivers. Two key open questions are whether such causally-informed neural networks: (a) lead to a reduced complexity of the system (lower dimensionality associated with greater input

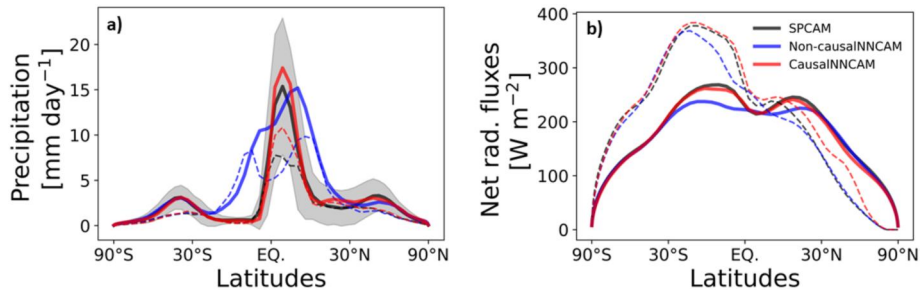

Figure 5. Zonal average climatologies of precipitation (P), and net radiative fluxes at the top of the atmosphere. (a) Mean (thick solid lines) and standard deviation (thin dashed lines) are shown for P. Note the shaded gray area indicates the standard deviation around the mean of SPCAM. (b) Radiative longwave (Q_{lw}^{top}; solid lines) and shortwave (Q_{sw}^{top}; dashed lines) net fluxes are shown. Zonal mean values are area-weighted, that is, cosine (latitudes). Note Non-causalNNCAM biases are not present in NNCAM (Rasp et al., 2018), see Figure S11 in Supporting Information S1.

sparsity); and (b) can make more accurate predictions across climate regimes (improving generalization skills). To address the former question, we explore the performance of equivalent lower dimensional NNs compared to the causally-informed case (i.e., same number of inputs), but applying different methods to select the inputs. Both, randomly- and correlationally-informed NNs show worse performance compared to the causally-informed case for heating and moistening tendencies (Figure S7 in Supporting Information S1). In addition, we use a linear version of the neural network parameterizations by replacing the activation functions with the identity function (i.e., removing non-linearity), to test whether inputs-to-outputs non-linearities are relaxed due to lower dimensionality in the causal case. The linear versions of the parameterizations show, as expected, a substantial drop in their performance compared to the non-linear versions (Figure S7 in Supporting Information S1). Interestingly, however, the performance of the linear parameterizations for both, Non-causalNN and Causally-informedNN, are equivalent. Therefore, lower dimensionality alone explains little of the causally-informed NNs accurate performance, suggesting that it is largely related to the use of causal drivers (i.e., removing spurious links).

Then, we investigate the generalization capabilities of the DL-based parameterizations across ±4 K climates compared to the original climate model (see Supporting Information S1). We find that causally-informed parameterizations retain similar generalization capabilities as the non-causal case, but without any substantial improvement with respect to the latter (Figure S7 in Supporting Information S1). In particular, we find that our DL-based parameterizations, both Non-causalNN and Causally-informedNN, generalize poorly under the +4 K climate, which is in line with previous results (Rasp et al., 2018). DL algorithms usually optimize an objective using a training dataset. NNs make out-of-distribution predictions (extrapolation) such as across different climates, relying on implicit assumptions. There is no inherent guarantee that NNs will accurately generalize far beyond their training data (Beucler et al., 2021), even when using causal drivers. This extrapolation challenge leads to the failure of DL-based parameterizations when confronted with environmental conditions significantly different from their training data range (Rasp et al., 2018). Overall, these results suggest that while neural network parameterizations can be improved in combination with causality, the prediction skills across climates must be enhanced by other approaches (Beucler et al., 2021).

3.3. Neural Nets Explainability

Having demonstrated that causal discovery helps unveil direct physical drivers of subgrid-scale processes and that the causally-informed prognostic simulation accurately represents the climate of the original SPCAM model, we turn here to explaining the predictions of such DL-based parameterizations. Figure 6 shows the feature importance of unresolved processes predictions for both, Causally-informedNN and Non-causalNN parameterizations, under the reference climate (+0 K), using a SHapley Additive exPlanations (SHAP) analysis (Lundberg & Lee, 2017; Shrikumar et al., 2017) (see Supporting Information S1). In both cases, the predominant features are in line with physical understanding. Spurious features, however, are evident in the Non-causalNN case (Figure 6a and Figure S9 in Supporting Information S1). Heating and moistening rates in the lower troposphere are linked to temperature throughout the atmosphere. The absence of a clear pattern—but random links—is suggestive of non-physical spurious correlations. These spurious links are likely the result of the strong inputs-to-outputs inter-

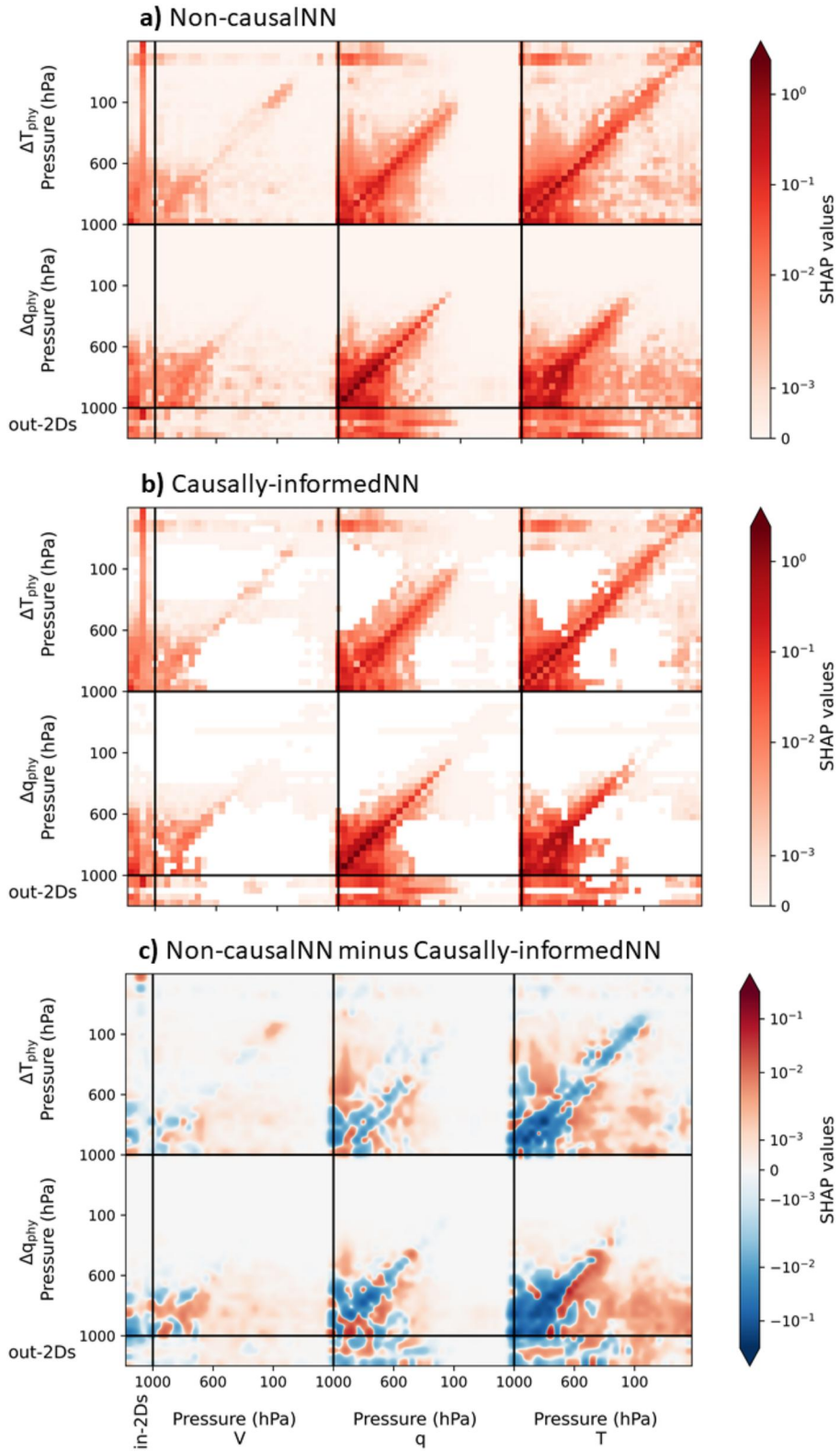
a) Non-causalNN
b) Causally-informedNN
c) Non-causalNN minus Causally-informedNN
ΔT_phy Pressure (hPa)
Δq_phy Pressure (hPa)
out-2Ds
in-2Ds
Pressure (hPa) V
Pressure (hPa) q
Pressure (hPa) T
SHAP values
100
600
1000

Figure 6.

correlation vertically in the atmosphere due to convective processes. By construction, such spurious links are mainly missing in the causally-informed parameterization (Figure 6b), which in turn shows stronger feature importance values for causal drivers compared to the latter (Figure 6c).

4. Discussion

4.1. Causal-Threshold Optimization

Optimizing the causal-threshold to find robust and globally invariant causal drivers poses the unique challenge of lacking a ground truth causal graph, and therefore, relies on both expert knowledge and empirical performance. This work considers two threshold optimization cases: a single optimized quantile-threshold for all outputs (fixed value); and a varying quantile-threshold optimized for each output separately. While the single optimized quantile-threshold may appear more straightforward and "cleaner" (Figure S1 in Supporting Information S1), it may fail to reflect the true complexity of the underlying causal relationships. In contrast, using a varying threshold optimized for each output introduces adaptability during the causal discovery phase and enables it to capture nuances in the data (Figure 2). For example, mild causal relationships between heating tendencies in the upper stratosphere with environmental conditions throughout the atmosphere, present in the varying quantile-threshold, may be associated with structural artifacts of the original SPCAM model (e.g., models with top of the atmosphere below the stratopause present stratosphere-troposphere coupling issues) (Charlton-Perez et al., 2013). The superior empirical performance of the causal parameterization based on the varying threshold approach is associated with the set of causal drivers that better uncover hidden causal dependencies (Figure 2), which may be overlooked by the more rigid single threshold strategy (Figure S1 in Supporting Information S1).

4.2. Hybrid Model Stability and Performance

This work builds on a previous NN (architecture and hyperparameters) based on the same dataset, for which the resulting hybrid model (once the NN is coupled to the coarse climate model) ran stably and accurately represented the climate of the original SPCAM model (Rasp et al., 2018). We find a number of Causal$_q$NNCAM cases with sub-optimal single optimized causal-thresholds ($q \in [0.6, 0.8]$) that were unstable. Moreover, causally-informed parameterizations with the optimal causal-threshold (quantile optimized; Table 1) but with simpler architectures (shallower and less complex) were also unstable once coupled to the host climate model. This result is in agreement with previous work (Rasp et al., 2018). Nevertheless, we find that Non-causalNNCAM, Causal$_{0.59}$NNCAM, and CausalNNCAM run stably without climate drifts (spurious and increasing long-term errors compared to SPCAM; not shown).

The DL-based parameterizations presented here (Non-causalNN, Causally-informed$_{0.59}$NN, and Causally-informedNN) perform as well as the original NN (Rasp et al., 2018) in the test set (offline; Figure S7 in Supporting Information S1). Particularly, we note that Non-causalNN (and Causally-informed$_{0.59}$NN; not shown) accurately captures both the ITCZ and midlatitudes storm-tracks as represented by SPCAM (Figure S8 top row in Supporting Information S1). We find that DL-based parameterizations for surface precipitation and net radiative fluxes following a better architecture found by a systematic hyperparameter tuning using an analogous SPCAM dataset (Hertel et al., 2020), lead to a marginal or negligible performance improvement (Figure S8 bottom row in Supporting Information S1). Notably, there are disparities between offline and online performances (Brenowitz, Henn, et al., 2020). In theory, it would be possible to develop equally good, or even better, performing hybrid models compared to the original NNCAM model (Rasp et al., 2018) if one trains a large number of NNs (Lin et al., 2023; Ott et al., 2020). In spite of that, this work advances DL-based parameterizations for climate models by implementing causal discovery, and demonstrates these newly developed causally-informed NNs better

Figure 6. Feature importance mapping of subgrid-scale processes predictions. Explanations are based on absolute averaged values of the SHapley Additive exPlanations (SHAP; see Supporting Information S1) (Lundberg & Lee, 2017; Shrikumar et al., 2017) for: (a) Non-causalNN; (b) Causally-informedNN and (c) the difference between both parameterizations (smoothed via Gaussian interpolation). Same as in Figure 2, the inputs of the neural networks are given on the x-axes, and the outputs (subgrid-scale processes) are represented on the y-axes. 3–D variables are shown from the surface (10^3 hPa) to the model's top (3 hPa). Absolute averaged SHAP values are shown in symlog scale, and are calculated for over 4,000 random samples of the test set compared to the train set (as background).

respect the underlying physical processes, with improved interpretability and without compromising performance skills (offline and online).

5. Conclusions

Data-driven parameterizations of subgrid-scale processes based on SRMs are able to represent to a good extent the climate of the original simulation once coupled to the coarser climate model (i.e., hybrid model) (Bretherton et al., 2022; Rasp et al., 2018; Watt-Meyer et al., 2021; Yuval & O'Gorman, 2020). Hybrid models can potentially alleviate persistent biases in coarse climate model simulations, and improve future climate projections. However, instabilities in hybrid models have been difficult to overcome and prognostic skills are challenging even in idealized simulations (e.g., aquaplanet settings without topography). It may well be that the sources of such instabilities and prognostic skills are associated with spurious non-physical relationships learned by the ML algorithm due to strong vertical inter-correlations, as well as fitting to noise. Current approaches to achieve stable hybrid models fail to care for the causes (e.g., deepening the deep learning algorithm or ablating the stratosphere) (Brenowitz & Bretherton, 2019; Rasp et al., 2018). An approach that is scalable and can reliably target the causes to overcome these issues would be a key breakthrough for ML-based parameterizations.

Here, we present a novel approach that combines causal discovery and DL to improve climate models and projections. We demonstrate that causal discovery robustly unveils causal (physical) drivers of subgrid-scale processes across different climate regimes, while improving interpretability and trust in the DL algorithm. Our causally-informed data-driven model also runs stably when coupled to the host coarse resolution model and generates a climate (mean and variability) close to the original simulation under the reference climate (within the distribution of the training dataset). We showed that causally-informed NNs prevent obvious spurious links in conventional DL-based parameterizations, leading to greater attention of the algorithm to the physical drivers compared to the latter.

Causal discovery, however, requires expert knowledge. In particular, we provided a solution to optimizing the causal-threshold (i.e., significance of the causal drivers), by running statistics over a number of causal graphs and testing the performance of the related causally-informed NNs. Yet, we avoided a systematic hyperparameter tuning of the original DL algorithm (Rasp et al., 2018) to find a stable and skillful performing hybrid model (Hertel et al., 2020). Moreover, we demonstrate that causal discovery, both PC_1 and linear Lasso regression, can identify key causal drivers of subgrid-scale processes that respect the underlying physical mechanisms (i.e., removing redundant information and non-physical links), for which standard feature selection methods, such as linear correlation, clearly fail. Future work will test this approach in more challenging and realistic setups (e.g., historical simulations with varying forcings and a real topography), as well as extend this method to integrate causality with state-of-the-art advances in deep learning approaches (Camps-Valls et al., 2021).

This work presents a fundamental and novel step in overcoming major challenges of data-driven models of physical processes (e.g., in parameterizations for climate models), paving the way toward improving climate models and projections via causally-based ML techniques. Explicitly using direct drivers in deep learning methods to represent physical processes is a key challenge that our methodology addresses, which in turn helps solve the problem of finding more reliable and reproducible data-driven parameterizations. Furthermore, advances in ML techniques are rapidly offering potential solutions to other limitations, such as generalization capabilities. The combination of causal discovery and deep learning presented here introduces a powerful new approach that opens a new window into process-based representation of complex processes not only for Earth system science but also in other scientific disciplines.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The code used to train the neural networks and to produce all figures of this manuscript is archived on Zenodo: Software - (Solino & Iglesias-Suarez, 2023). An example of SPCAM data is also archived on Zenodo: Data - (Rasp, 2019).

Acknowledgments
Funding for this study was provided by the European Research Council (ERC) Synergy Grant "Understanding and modeling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant Agreement 855187). F.I.-S. is a postdoc of the ELLIS Postdoc Program and acknowledges travel support from the European Union's Horizon 2020 research and innovation programme under ELISE (Grant Agreement 951847). Additionally, T.B. acknowledges funding from the Columbia University sub-award 1 (PG010560-01). P.G. and M.P. acknowledge funding from the National Science Foundation Science and Technology Center, Learning the Earth with Artificial intelligence and Physics, LEAP (Grant 2019625). M.P. acknowledges funding from the US Department of Energy Advanced Scientific Computing Research program (DE-SC0022331). J.R. has received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant 948112). This work used resources of both, the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID 1179 (USMILE), and the supercomputer JUWELS at the Jülich Supercomputing Centre (JSC) under the Earth System Modelling Project (ESM). Open Access funding enabled and organized by Projekt DEAL.

References

Ahmed, F., & Neelin, J. D. (2018). Reverse engineering the tropical precipitation–buoyancy relationship. *Journal of the Atmospheric Sciences*, *75*(5), 1587–1608. https://doi.org/10.1175/JAS-D-17-0333.1

Andersen, J. A., & Kuang, Z. (2012). Moist static energy budget of mjo-like disturbances in the atmosphere of a zonally symmetric aquaplanet. *Journal of Climate*, *25*(8), 2782–2804. https://doi.org/10.1175/JCLI-D-11-00168.1

Arnold, N. P., & Randall, D. A. (2015). Global-scale convective aggregation: Implications for the madden-julian oscillation. *Journal of Advances in Modeling Earth Systems*, *7*(4), 1499–1518. https://doi.org/10.1002/2015MS000498

Benedict, J. J., & Randall, D. A. (2009). Structure of the madden–julian oscillation in the superparameterized cam. *Journal of the Atmospheric Sciences*, *66*(11), 3277–3296. https://doi.org/10.1175/2009JAS3030.1

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., et al. (2021). Climate-invariant machine learning. arXiv preprint arXiv:2112.08440.

Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., et al. (2020). Quantifying progress across different cmip phases with the esmvaltool. *Journal of Geophysical Research: Atmospheres*, *125*(21), e2019JD032321. https://doi.org/10.1029/2019JD032321

Bony, S., Stevens, B., Frierson, D. M. W., Jakob, C., Kageyama, M., Pincus, R., et al. (2015). Clouds, circulation and climate sensitivity. *Nature Geoscience*, *8*(4), 261–268. https://doi.org/10.1038/ngeo2398

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. https://doi.org/10.1175/JAS-D-20-0082.1

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. https://doi.org/10.1029/2019ms001711

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., et al. (2020). Machine learning climate model dynamics: Offline versus online performance. *arXiv*. https://doi.org/10.48550/ARXIV.2011.03081

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., et al. (2022). Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. https://doi.org/10.1029/2021MS002794

Camps-Valls, G., Tuia, D., Zhu, X., & Reichstein, M. E. (2021). *Deep learning for the earth sciences: A comprehensive approach to remote sensing, climate science and geosciences*. Wiley & Sons. Retrieved from https://github.com/DL4ES

Charlton-Perez, A. J., Baldwin, M. P., Birner, T., Black, R. X., Butler, A. H., Calvo, N., et al. (2013). On the lack of stratospheric dynamical variability in low-top versions of the cmip5 models. *Journal of Geophysical Research: Atmospheres*, *118*(6), 2494–2505. https://doi.org/10.1002/jgrd.50125

Collins, W. D., Rasch, P. J., Boville, B. A., Hack, J. J., McCaa, J. R., Williamson, D. L., et al. (2006). The formulation and atmospheric simulation of the community atmosphere model version 3 (cam3). *Journal of Climate*, *19*(11), 2144–2161. https://doi.org/10.1175/JCLI3760.1

D'Andrea, F., Gentine, P., Betts, A. K., & Lintner, B. R. (2014). Triggering deep convection with a probabilistic plume model. *Journal of the Atmospheric Sciences*, *71*(11), 3881–3901. https://doi.org/10.1175/JAS-D-13-0340.1

Del Genio, A. D. (2012). Representing the sensitivity of convective cloud systems to tropospheric humidity in general circulation models. *Surveys in Geophysics*, *33*(3–4), 637–656. https://doi.org/10.1007/s10712-011-9148-9

Del Genio, A. D., & Wu, J. (2010). The role of entrainment in the diurnal cycle of continental convection. *Journal of Climate*, *23*(10), 2722–2738. https://doi.org/10.1175/2009JCLI3340.1

De Rooy, W. C., Bechtold, P., Fröhlich, K., Hohenegger, C., Jonker, H., Mironov, D., et al. (2013). Entrainment and detrainment in cumulus convection: An overview. *Quarterly Journal of the Royal Meteorological Society*, *139*(670), 1–19. https://doi.org/10.1002/qj.1959

Eyring, V., Gillett, N., Rao, K. A., Barimalala, R., Parrillo, M. B., Bellouin, N., et al. (2021). Human influence on the climate system. In V. Masson-Delmotte (Ed.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 423–552). Cambridge University Press. Retrieved from https://www.ipcc.ch/report/ar6/wg1/

Eyring, V., Mishra, V., Griffith, G. P., Chen, L., Keenan, T., Turetsky, M. R., et al. (2021). Reflections and projections on a decade of climate science. *Nature Climate Change*, *11*(4), 279–285. https://doi.org/10.1038/s41558-021-01020-x

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models. In T. F. Stocker (Ed.), *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change* (pp. 741–866). Cambridge University Press. Retrieved from https://www.ipcc.ch/report/ar5/wg1

Frezat, H., Le Sommer, J., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003124. https://doi.org/10.1029/2022MS003124

Galytska, E., Weigel, K., Handorf, D., Jaiser, R., Köhler, R., Runge, J., & Eyring, V. (2023). Evaluating causal arctic-midlatitude teleconnections in cmip6. *Journal of Geophysical Research: Atmospheres*, *128*(17), e2022JD037978. https://doi.org/10.1029/2022JD037978

Gentine, P., Eyring, V., & Beucler, T. (2021). *Deep learning for the parametrization of subgrid processes in climate models*. Wiley. https://doi.org/10.1002/9781119646181.ch21

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. https://doi.org/10.1029/2018GL078202

Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M. A., & Eyring, V. (2022). Deep learning based cloud cover parameterization for icon. *Journal of Advances in Modeling Earth Systems*, *14*(12), e2021MS002959. https://doi.org/10.1029/2021MS002959

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A moist physics parameterization based on deep learning. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002076. https://doi.org/10.1029/2020MS002076

Hertel, L., Collado, J., Sadowski, P., Ott, J., & Baldi, P. (2020). Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, *12*, 100591. https://doi.org/10.1016/j.softx.2020.100591

IPCC. (2021a). *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. In Press) [Book]. https://doi.org/10.1017/9781009157896

IPCC. (2021b). In V. Masson-Delmotte(Ed.), *Annex II: Models. Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change, gutiérrez, j m., a.-m. tréguier* (pp. 423–552). Cambridge University Press. Retrieved from https://www.ipcc.ch/report/ar6/wg1/

Karmouche, S., Galytska, E., Runge, J., Meehl, G. A., Phillips, A. S., Weigel, K., & Eyring, V. (2023). Regime-oriented causal model evaluation of atlantic–pacific teleconnections in cmip6. *Earth System Dynamics*, *14*(2), 309–344. https://doi.org/10.5194/esd-14-309-2023

Khairoutdinov, M., & Randall, D. (2006). High-resolution simulation of shallow-to-deep convection transition over land. *Journal of the Atmospheric Sciences*, *63*(12), 3421–3436. https://doi.org/10.1175/JAS3810.1

Khairoutdinov, M. F., & Randall, D. A. (2001). A cloud resolving model as a cloud parameterization in the ncar community climate system model: Preliminary results. *Geophysical Research Letters*, *28*(18), 3617–3620. https://doi.org/10.1029/2001GL013552

Klocke, D., Brueck, M., Hohenegger, C., & Stevens, B. (2017). Rediscovery of the doldrums in storm-resolving simulations over the tropical Atlantic. *Nature Geoscience*, *10*(12), 891–896. https://doi.org/10.1038/s41561-017-0005-4

Kooperman, G. J., Pritchard, M. S., Burt, M. A., Branson, M. D., & Randall, D. A. (2016a). Impacts of cloud superparameterization on projected daily rainfall intensity climate changes in multiple versions of the community earth system model. *Journal of Advances in Modeling Earth Systems*, *8*(4), 1727–1750. https://doi.org/10.1002/2016MS000715

Kooperman, G. J., Pritchard, M. S., Burt, M. A., Branson, M. D., & Randall, D. A. (2016b). Robust effects of cloud superparameterization on simulated daily rainfall intensity statistics across multiple versions of the community earth system model. *Journal of Advances in Modeling Earth Systems*, *8*(1), 140–165. https://doi.org/10.1002/2015MS000574

Kooperman, G. J., Pritchard, M. S., O'Brien, T. A., & Timmermans, B. W. (2018). Rainfall from resolved rather than parameterized processes better represents the present-day and climate change response of moderate rates in the community atmosphere model. *Journal of Advances in Modeling Earth Systems*, *10*(4), 971–988. https://doi.org/10.1002/2017MS001188

Kretschmer, M., Adams, S. V., Arribas, A., Prudden, R., Robinson, N., Saggioro, E., & Shepherd, T. G. (2021). Quantifying causal pathways of teleconnections. *Bulletin of the American Meteorological Society*, *102*(12), E2247–E2263. https://doi.org/10.1175/BAMS-D-20-0117.1

Kretschmer, M., Cohen, J., Matthias, V., Runge, J., & Coumou, D. (2018). The different stratospheric influence on cold-extremes in Eurasia and North America. *npj Climate and Atmospheric Science*, *1*(1), 44. https://doi.org/10.1038/s41612-018-0054-4

Kretschmer, M., Coumou, D., Donges, J. F., & Runge, J. (2016). Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *Journal of Climate*, *29*(11), 4069–4081. https://doi.org/10.1175/JCLI-D-15-0654.1

Krich, C., Runge, J., Miralles, D. G., Migliavacca, M., Perez-Priego, O., El-Madany, T., et al. (2020). Estimating causal networks in biosphere–atmosphere interaction with the pcmci approach. *Biogeosciences*, *17*(4), 1033–1061. https://doi.org/10.5194/bg-17-1033-2020

Kuang, Z., & Bretherton, C. S. (2006). A mass-flux scheme view of a high-resolution simulation of a transition from shallow to deep cumulus convection. *Journal of the Atmospheric Sciences*, *63*(7), 1895–1909. https://doi.org/10.1175/JAS3723.1

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J., et al. (2021). Future global climate: Scenario-based projections and near term information. In V. Masson-Delmotte (Ed.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 423–552). Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.020

Lin, J., Yu, S., Beucler, T., Gentine, P., Walling, D., & Pritchard, M. (2023). Systematic sampling and validation of machine learning-parameterizations in climate models.

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R. A., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8). https://doi.org/10.1029/2022MS003105

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. In I. Guyon (Ed.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Mapes, B., & Neale, R. (2011). Parameterizing convective organization to escape the entrainment dilemma. *Journal of Advances in Modeling Earth Systems*, *3*(2). https://doi.org/10.1029/2011MS000042

Muller, C., & Bony, S. (2015). What favors convective aggregation and why? *Geophysical Research Letters*, *42*(13), 5626–5634. https://doi.org/10.1002/2015GL064260

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. https://doi.org/10.1029/2018MS001351

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A fortran-keras deep learning bridge for scientific computing. *Scientific Programming*, *2020*, 1–13. https://doi.org/10.1155/2020/8888811

Oueslati, B., & Bellon, G. (2015). The double ITCZ bias in CMIP5 models: Interaction between SST, large-scale circulation and precipitation. *Climate Dynamics*, *44*(3–4), 585–607. https://doi.org/10.1007/s00382-015-2468-6

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The new science of cause and effect*. Basic books.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, *78*(5), 947–1012. https://doi.org/10.1111/rssb.12167

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., et al. (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of Geophysics*, *53*(2), 323–361. https://doi.org/10.1002/2014RG000475

Pritchard, M. S., & Bretherton, C. S. (2014). Causal evidence that rotational moisture advection is critical to the superparameterized madden–julian oscillation. *Journal of the Atmospheric Sciences*, *71*(2), 800–815. https://doi.org/10.1175/JAS-D-13-0119.1

Pritchard, M. S., Bretherton, C. S., & DeMott, C. A. (2014). Restricting 32–128 km horizontal scales hardly affects the mjo in the super-parameterized community atmosphere model v.3.0 but the number of cloud-resolving grid columns constrains vertical mixing. *Journal of Advances in Modeling Earth Systems*, *6*(3), 723–739. https://doi.org/10.1002/2014MS000340

Rasp, S. (2019). Sample spcam dataset. Dataset. *Zenodo*. https://doi.org/10.5281/zenodo.2559313

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Romps, D. M., & Kuang, Z. (2011). A transilient matrix for moist convection. *Journal of the Atmospheric Sciences*, *68*(9), 2009–2025. https://doi.org/10.1175/2011JAS3712.1

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019). Inferring causation from time series in earth system sciences. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-10105-3

Runge, J., Gerhardus, A., Varando, G., Eyring, V., & Camps-Valls, G. (2023). Causal inference for time series. *Nature Reviews Earth & Environment*, *4*(7), 1–19. https://doi.org/10.1038/s43017-023-00431-y

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11), eaau4996. https://doi.org/10.1126/sciadv.aau4996

Runge, J., Petoukhov, V., Donges, J. F., Hlinka, J., Jajcay, N., Vejmelka, M., et al. (2015). Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, *6*(1), 8502. https://doi.org/10.1038/ncomms9502

Runge, J., Petoukhov, V., & Kurths, J. (2014). Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of Climate*, *27*(2), 720–739. https://doi.org/10.1175/JCLI-D-13-00159.1

Scher, S., & Messori, G. (2019). How global warming changes the difficulty of synoptic weather forecasting. *Geophysical Research Letters*, *46*(5), 2931–2939. https://doi.org/10.1029/2018GL081856

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., & Siebesma, A. P. (2017). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*(1), 3–5. https://doi.org/10.1038/nclimate3190

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634. https://doi.org/10.1109/jproc.2021.3058954

Sherwood, S. C., Bony, S., & Dufresne, J.-L. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, *505*(7481), 37–42. https://doi.org/10.1038/nature12829

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *arXiv*. https://doi.org/10.48550/arXiv.1704.02685

Siew, P. Y. F., Li, C., Sobolowski, S. P., & King, M. P. (2020). Intermittency of arctic–mid-latitude teleconnections: Stratospheric pathway between autumn sea ice and the winter north atlantic oscillation. *Weather and Climate Dynamics*, *1*(1), 261–275. https://doi.org/10.5194/wcd-1-261-2020

Solino, B., & Iglesias-Suarez, F. (2023). EyringMLClimateGroup/iglesias-suarez23jgr_CausalNNCAM: Causally-informed deep learning to improve climate models and projections (v1.1). Software. *Zenodo*. https://doi.org/10.5281/zenodo.8239403

Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, *9*(1), 62–72. https://doi.org/10.1177/089443939100900106

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., et al. (2019). Dyamond: The DYnamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*(1), 61. https://doi.org/10.1186/s40645-019-0304-z

Stommel, H. (1951). Entrainment of air into a cumulus cloud. *Journal of the Atmospheric Sciences*, *8*(2), 127–129. https://doi.org/10.1175/1520-0469(1951)008⟨0127:EOAIAC⟩2.0.CO;2

Stull, R. B. (1993). Review of non-local mixing in turbulent atmospheres: Transilient turbulence theory. In H. Kaplan, N. Dinar, A. Lacser, & Y. Alexander (Eds.), *Transport and diffusion in turbulent fields: Modeling and measurement techniques* (pp. 21–96). Springer Netherlands. https://doi.org/10.1007/978-94-011-2749-3_2

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., et al. (2021). Climate model projections from the scenario model intercomparison project (scenariomip) of cmip6. *Earth System Dynamics*, *12*(1), 253–293. https://doi.org/10.5194/esd-12-253-2021

Tibshirani, R. (1996). Regression Shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tompkins, A. M. (2001). On the relationship between tropical convection and sea surface temperature. *Journal of Climate*, *14*(5), 633–637. https://doi.org/10.1175/1520-0442(2001)014⟨0633:OTRBTC⟩2.0.CO;2

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022). Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development*, *15*(9), 3923–3940. https://doi.org/10.5194/gmd-15-3923-2022

Warner, J. (1970). On steady-state one-dimensional models of cumulus convection. *Journal of the Atmospheric Sciences*, *27*(7), 1035–1040. https://doi.org/10.1175/1520-0469(1970)027⟨1035:OSSODM⟩2.0.CO;2

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., et al. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15). https://doi.org/10.1029/2021gl092555

WMO. (2018). *Scientific assessment of ozone Depletion: 2018*. Tech. Rep.).

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. https://doi.org/10.1038/s41467-020-17142-3

References From the Supporting Information

Jones, T. R., Randall, D. A., & Branson, M. D. (2019). Multiple-instance superparameterization: 2. The effects of stochastic convection on the simulated climate. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3521–3544. https://doi.org/10.1029/2019MS001611

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*. https://doi.org/10.48550/ARXIV.1412.6980