# Supporting Information for "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, 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, USA

[4]University of Lausanne, Institute of Earth Surface Dynamics, Lausanne, Switzerland

[5]University of California, Department of Earth System Science, Irvine, USA

[6]NVIDIA Corporation, Santa Clara, 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

**Contents of this file**

1. Text S1 to S4
2. Table S1
3. Figures S1 to S11

**Introduction** This text complements the description of the concept of this study, the SPCAM model setup, neural networks specifications, causal discovery optimization, and neural nets explainability provided on the main manuscript.

**Text S1. Superparameterized Community Atmosphere Model v3.0 (SPCAM)** The superparameterization component (SP) is spun-up at the beginning of the simulation and subcycles every 20 seconds given the large-scale tendencies (Collins et al., 2006). At the end of every time step (30 minutes), the horizontal mean of state variables for temperature, moisture and condensate from the SP component update the resolved fields in its host (Benedict & Randall, 2009). Note that unlike traditional parameterizations, the SP component runs continuously throughout the simulation after the initial spin-up, which adds a "memory" effect to the subgrid-scale processes affecting the large-scale tendencies. The memory of subgrid-scale processes is not explicitly treated during the training of the neural networks (NNs); rather we assume its importance is secondary (Jones et al., 2019). After the SP component, the radiation scheme is called using the explicitly resolved vertical distribution of clouds among the large-scale resolved fields. Finally, surface fluxes are computed following a simple bulk scheme using the host model's coarse state fields, and then the dynamical core.

Solar insolation follows a diurnal cycle in perpetual Southern Hemisphere solstice. Sea surface temperatures (SSTs) are imposed following a zonally symmetric setup but with a shift in maximum temperatures five degrees North of the equator (Andersen & Kuang, 2012):

$$SST(\phi) = 2 + \frac{27}{2}(2 - \zeta - \zeta^2), \tag{1}$$

with $SSTs$ in Celcius, latitudes ($\phi$) in degrees, and

$$\zeta = \begin{cases} sin^2(\pi \frac{\phi - 5}{110}) & 5 < \phi \leq 60 \\ sin^2(\pi \frac{\phi - 5}{130}) & -60 \leq \phi < 5 \\ 1 & \text{if } |\phi| < 60 \end{cases}. \tag{2}$$


0148-0227/23/$9.00

Sensitivity experiments comprise global changes in SSTs ($\pm 4$ K) by adding a wavenumber one perturbation to the reference ($+0$ K) SSTs in increments of 1 K:

$$SST'(\lambda, \phi) = 3\, cos\Big(\frac{\lambda \pi}{180}\Big)\, cos\Bigg(0.5\pi \frac{\frac{\phi\pi}{180} - 5}{30}\Bigg)^2 \quad \text{if} \quad -25 \leq \phi \leq 35, \tag{3}$$

with longitudes ($\lambda$) in degrees. The SPCAM model source code used here, including the neural network implementation via the Fortran-Keras Bridge (Ott et al., 2020), is available at `https://gitlab.com/mspritch/spcam3.0-neural-net` (causalcoupler branch; commit hash: 5ebff0a6). Further details of the SPCAM model are provided elsewhere (Khairoutdinov & Randall, 2001; Collins et al., 2006; Pritchard et al., 2014; Pritchard & Bretherton, 2014).

**Text S2. Neural network setup** Building on previous hybrid modeling work (Rasp et al., 2018), we develop and train all NNs used here with Keras (`https://keras.io/`) built on top of Tensorflow 2 (`https://www.tensorflow.org/`). All NNs were trained for 18 epochs using: a batch size of $1,024$; the LeakyReLU activation function set to $max(0.3x, x)$; the Adam optimizer (Kingma & Ba, 2014); a starting learning rate of $1 \times 10^3$ subsequently divided by 5 every 3 epochs; and a mean squared error loss function.

Each input field was normalized by subtracting its mean across samples and then dividing it by its maximum range, whereas the outputs were normalized to bring them to the same order of magnitude (Table S1).

The computational costs of CausalNNCAM and Non-causalNNCAM are virtually the same. However, these hybrid models, consisting of 65 single-output NNs coupled to the host model via the Fortran-Keras Bridge, result in similar computational costs compared to SPCAM. Therefore, CausalNNCAM and Non-causalNNCAM are approximately 20 times slower than a former hybrid model consisting in one multi-output NN (Rasp et al., 2018). Nevertheless, we emphasize that our study is a proof of concept focused on the added value of combining causality and ML, and envisage an overall speed up of the hybrid models could be attained by developing a method that enables multi-output causally-informed NNs, as well as advanced software engineering solutions and coding (e.g., efficiently coupling machine learning algorithms with Fortran code).

**Text S3. Causal drivers optimization** We optimize the causal threshold during the causal discovery phase (i.e., finding the most important causal drivers of the subgrid-scale processes in SPCAM). The causal drivers determine the input layer of the causally-informed NNs, while the rest of the

NNs setup remains unchanged (see above). Two optimization approaches are considered: 1) a single optimized causal threshold for all outputs; and 2) a varying optimized causal threshold for each output.

The single optimized causal threshold approach focuses on $\Delta T_{phy}$ and $\Delta q_{phy}$ at the level closest to the surface (992 hPa), where the causally-informed NNs show worst offline performance compared to the non-causal NNs (reference case). The single optimized threshold is chosen following two conditions:

1. $R^2_{causally-informedNN(thr)} \geq R^2_{non-causalNN} \leftrightarrow R^2_{non-causalNN} > 0$
2. $\max(thr)$.

The NNs were trained using the reference simulation (+0 K), and their performance was computed applying the coefficient of determination ($R^2$) to the test sets of all simulations considered (−4 K, +0 K and +4 K). While the first condition selects those causally-informed NNs that perform at least as good as its non-causal NN counterpart (as long as it shows some skill), the second condition selects the causally-informed NN with the most stringent threshold (minimum number of causal drivers).

There are two causal threshold definitions considered (Fig. S1):

$$causal'(Y_t^j) = X_{t-\tau}^i : P(\, X_{t-\tau}^i \rightarrow Y_t^j \in causal_g(Y_t^j)\,) > quantile \tag{4}$$

and,

$$causal'(Y_t^j) = X_{t-\tau}^i : \Big(\frac{\#(\, X_{t-\tau}^i \in causal_g(Y_t^j)\,)}{N_g}\Big) > ratio. \tag{5}$$

The quantile-based definition considers causal drivers of each output ($causal'(Y_t^j)$), the inputs ($X_{t-\tau}^i$) for which their probability of being causally-linked to the given output is greater than a certain quantile (threshold). The ratio-based definition considers causal drivers of each output, the inputs for which their ratio of being causally-linked to the given output across the model's grid ($N_g$ = 8,192) is greater than a certain value (threshold). Figure S10 shows the $R^2$ value for all causally-informed NNs explored compared to the non-causal NNs for both, $\Delta T_{phy}$ and $\Delta q_{phy}$. While using the definition of the ratio-based threshold no causally-informed NN explored met the above conditions, we find the optimal 0.59 value for the quantile-based approach (See Table 1; Causal-threshold: quantile 0.59).

Using the quantile-based threshold definition, a varying optimized causal threshold for each output is achieved using the Grid Search algorithm of the SHERPA package (Hertel et al., 2020). We explore a threshold range of 0.-0.95 at intervals of 0.05. For each output, 20 NNs were trained –one per threshold step– using the reference simulation (+0 K), and the optimal threshold is based on the minimum mean squared error of the validation set (6 floating points).

Figure S7 shows the offline performance ($R^2$) of the Causally-informed$_{0.59}$NN (single optimized causal-threshold) and Causally-informedNN (varying optimized causal-threshold) cases for $\Delta T_{phy}$ and $\Delta q_{phy}$, using the test set of the reference simulation (+0 K). Although Causally-informed$_{0.59}$NN shows similar offline performance compared to Causally-informedNN, its hybrid model (Causal$_{0.59}$NNCAM) shows a double Intertropical Convergence Zone bias (see discussion in main text and Fig. S5 and S6).

**Text S4. Neural nets explainability** We use the SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) game theoretic approach to explain the predictions of subgrid-scale processes by the NNs. Although the computation of the exact Shapley value is challenging, there are different SHAP built-on methods to approximate it. The algorithm used here is DeepExplainer, specifically tailored for deep learning models, which is based on the Deep Learning Important FeaTures (DeepLIFT) (Shrikumar et al., 2017). DeepExplainer decomposes the prediction (output) of a neural network on the different inputs. This is achieved by approximating the difference of the output from a distribution of background outputs with regard to the difference of the input from a distribution of background inputs. The complexity of the method scales linearly with the number of background samples ($n$), and the variance of the expectation estimates scales by approximately $1/\sqrt{n}$. Although 1000 samples would give already a very good approximation to the exact Shapley value, we use 4096 random samples across 1440 time-steps (∼1 month) and the full horizontal model grid (8192 points; latitude by longitude).

## References


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. Retrieved from `https://journals.ametsoc.org/view/journals/clim/25/8/jcli-d-11-00168.1.xml` doi: 10.1175/JCLI-D-11-00168.1

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. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/66/11/2009jas3030.1.xml` doi: 10.1175/2009JAS3030.1

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. Retrieved from `https://journals.ametsoc.org/view/journals/clim/19/11/jcli3760.1.xml` doi: 10.1175/JCLI3760.1

Hertel, L., Collado, J., Sadowski, P., Ott, J., & Baldi, P. (2020). Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, *12*, 100591. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2352711020303046` doi: https://doi.org/10.1016/j.softx.2020.100591

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. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001611` doi: https://doi.org/10.1029/2019MS001611

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. Retrieved from

https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001GL013552 doi: https://doi.org/10.1029/2001GL013552

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

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *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

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. doi: https://doi.org/10.1155/2020/8888811

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. Retrieved from https://journals.ametsoc.org/view/journals/atsc/71/2/jas-d-13-0119.1.xml doi: 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 superparameterized 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. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014MS000340 doi: https://doi.org/10.1002/2014MS000340

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.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *CoRR*, *abs/1704.02685*. Retrieved from http://arxiv.org/abs/1704.02685

**Table S1.** Summary of neural networks inputs and output fields.

| Inputs | Units | Outputs | Units | Normalization |
|---|---|---|---|---|
| Temperature, $T(p)$ | K | Temperature tendencies, $\Delta T_{phy}(p)$ | $Ks^{-1}$ | $C_p$ |
| Specific humidity, $q(p)$ | $kgkg^{-1}$ | Moistening tendencies $\Delta q_{phy}(p)$ | $kgkg^{-1}s^{-1}$ | $L_v$ |
| Meridional wind, $V(p)$ | $ms^{-1}$ | Net shortwave radiative heat flux at TOA, $Q_{sw}^{top}$ | $Wm^{-2}$ | $10^{-3}$ |
| Surface pressure, $P_{srf}$ | Pa | Net longwave radiative heat flux at TOA, $Q_{lw}^{top}$ | $Wm^{-2}$ | $10^{-3}$ |
| Incoming solar radiation, $Q_{sol}$ | $Wm^{-2}$ | Net shortwave radiative heat flux at the surface, $Q_{sw}^{srf}$ | $Wm^{-2}$ | $10^{-3}$ |
| Sensible heat flux, $Q_{sen}$ | $Wm^{-2}$ | Net longwave radiative heat flux at the surface, $Q_{lw}^{srf}$ | $Wm^{-2}$ | $10^{-3}$ |
| Latent heat flux, $Q_{lat}$ | $Wm^{-2}$ | Precipitation, P | $kgm^{-2}d^{-1}$ | $1.728\times10^{6}$ |

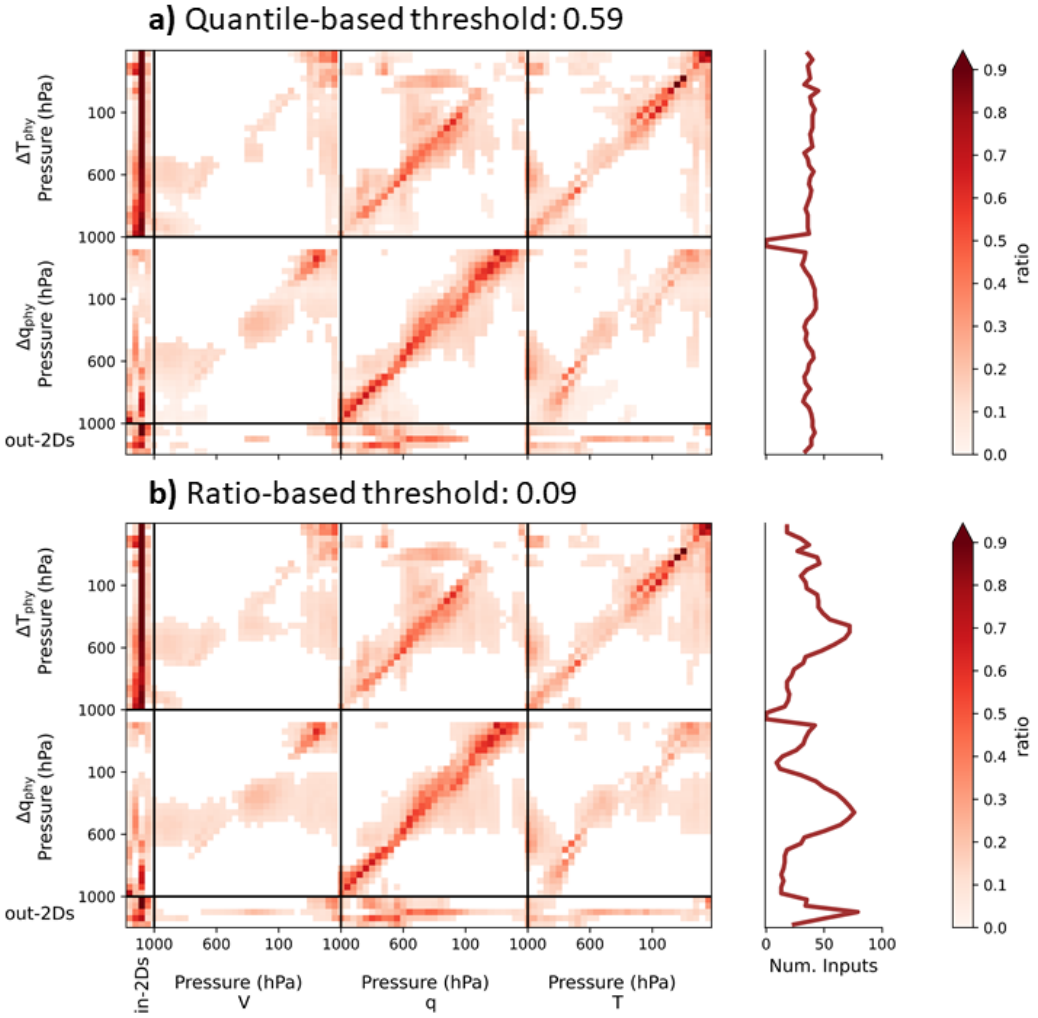


**Figure S1.** Same as Fig. 2, but for single optimized **a**) quantile-based threshold (0.59), and **b**) ratio-based threshold (0.09). Right panels show the number of causal drivers for each output, with a mean number of inputs of 36 (39 % of the total) and 35 (38 % of the total) for the quantile-based-threshold of 0.59 and the ratio-based threshold of 0.09, respectively.

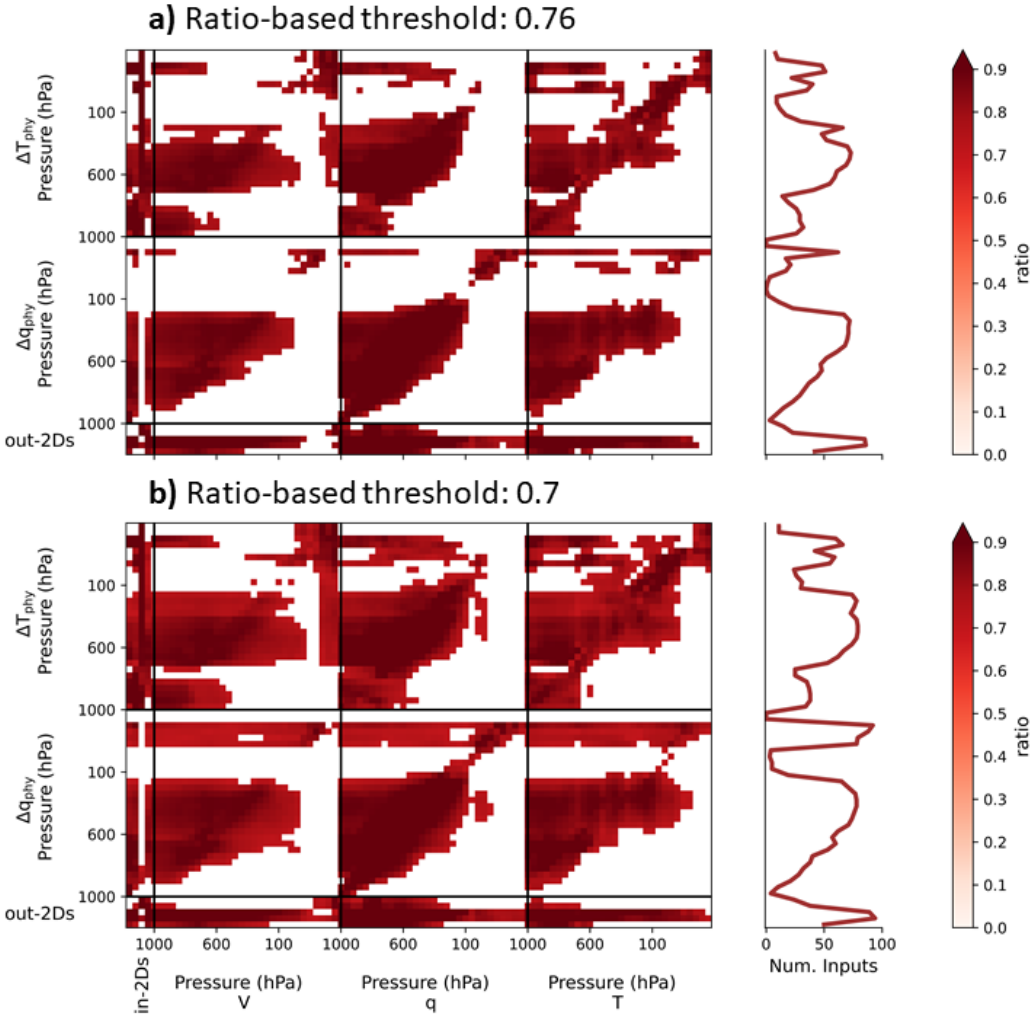


**Figure S2.** Same as Fig. S1, but for Pearson correlation ratio-based thresholds of **a**) 0.76, and **b**) 0.7. Right panels show the number of causal drivers for each output, with a mean number of inputs of 36 (39 % of the total) and 48 (51 % of the total) for the ratio-based thresholds of 0.76 and 0.7, respectively.

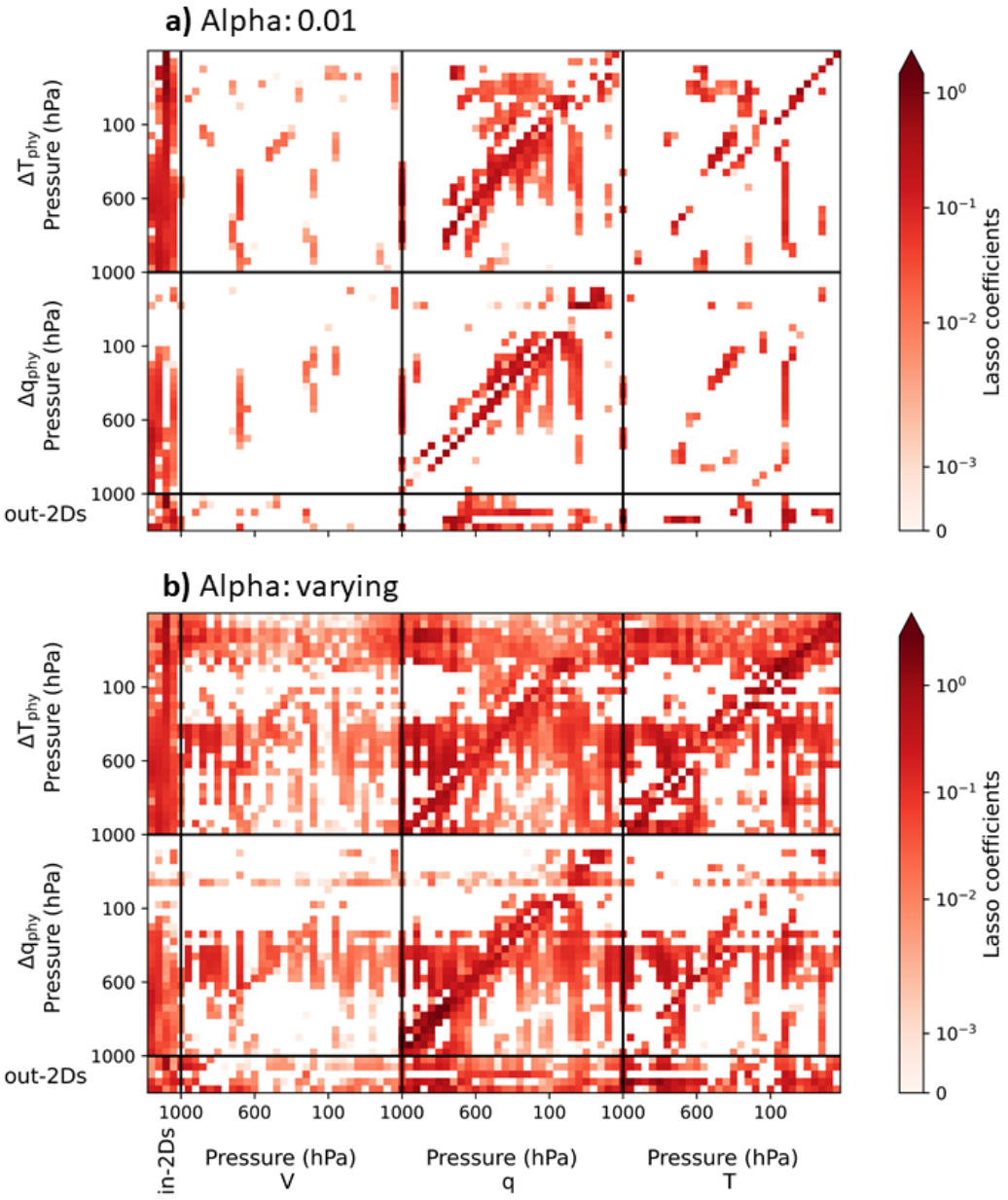


**Figure S3.** Same as Fig. S1, but for Lasso regression with **a**) 0.01 alpha and **b**) varying alpha. Note varying alpha in **b**) is chosen to obtain a similar number of inputs as in the quantile-optimized causal-threshold case (Fig. 2).

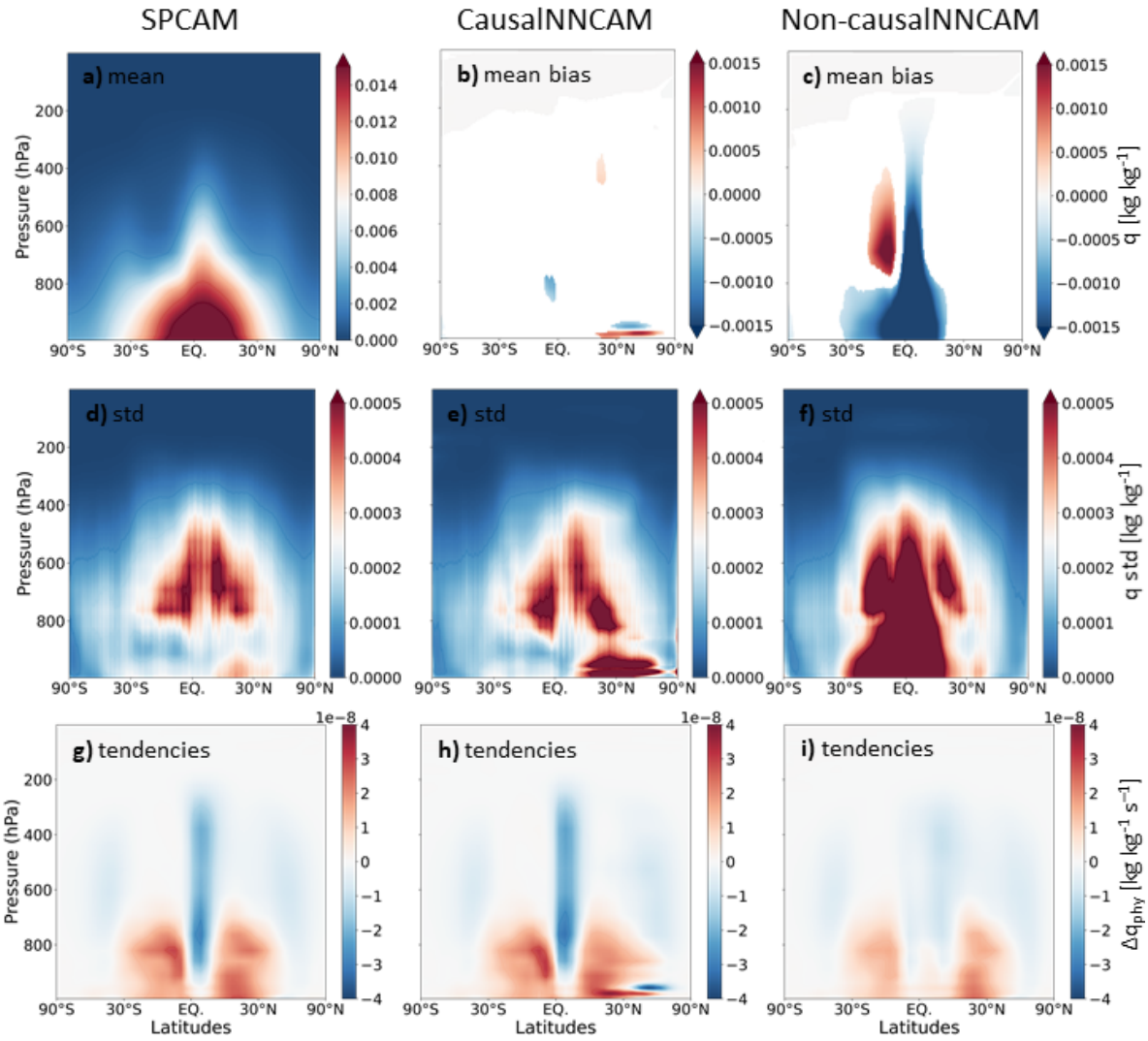


**Figure S4.** Same as Fig. 4, but for specific humidity ($q$) and moistening rates ($\Delta q_{phy}$).

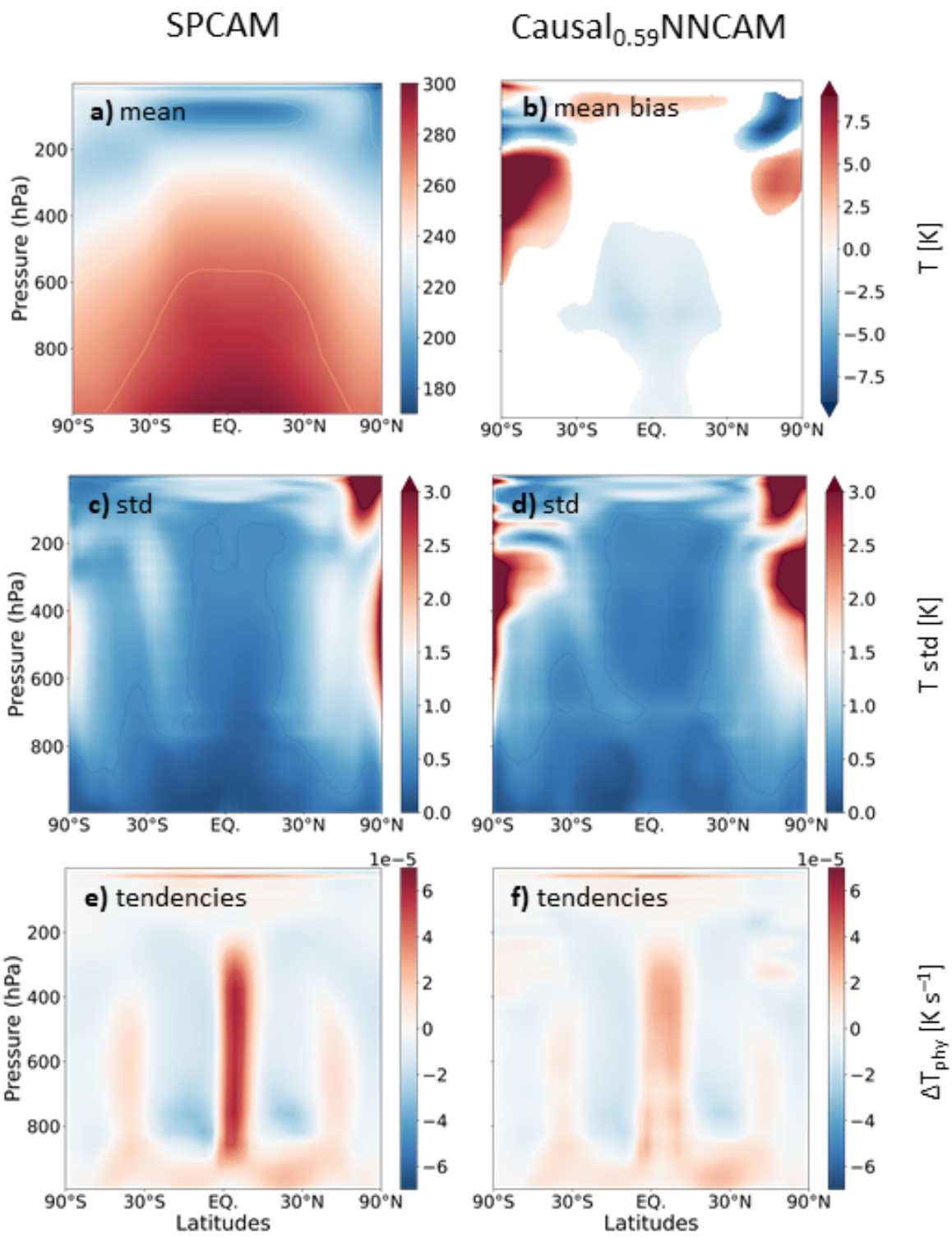


**Figure S5.** Same as Fig. 4, but for Causal$_{0.59}$NNCAM.

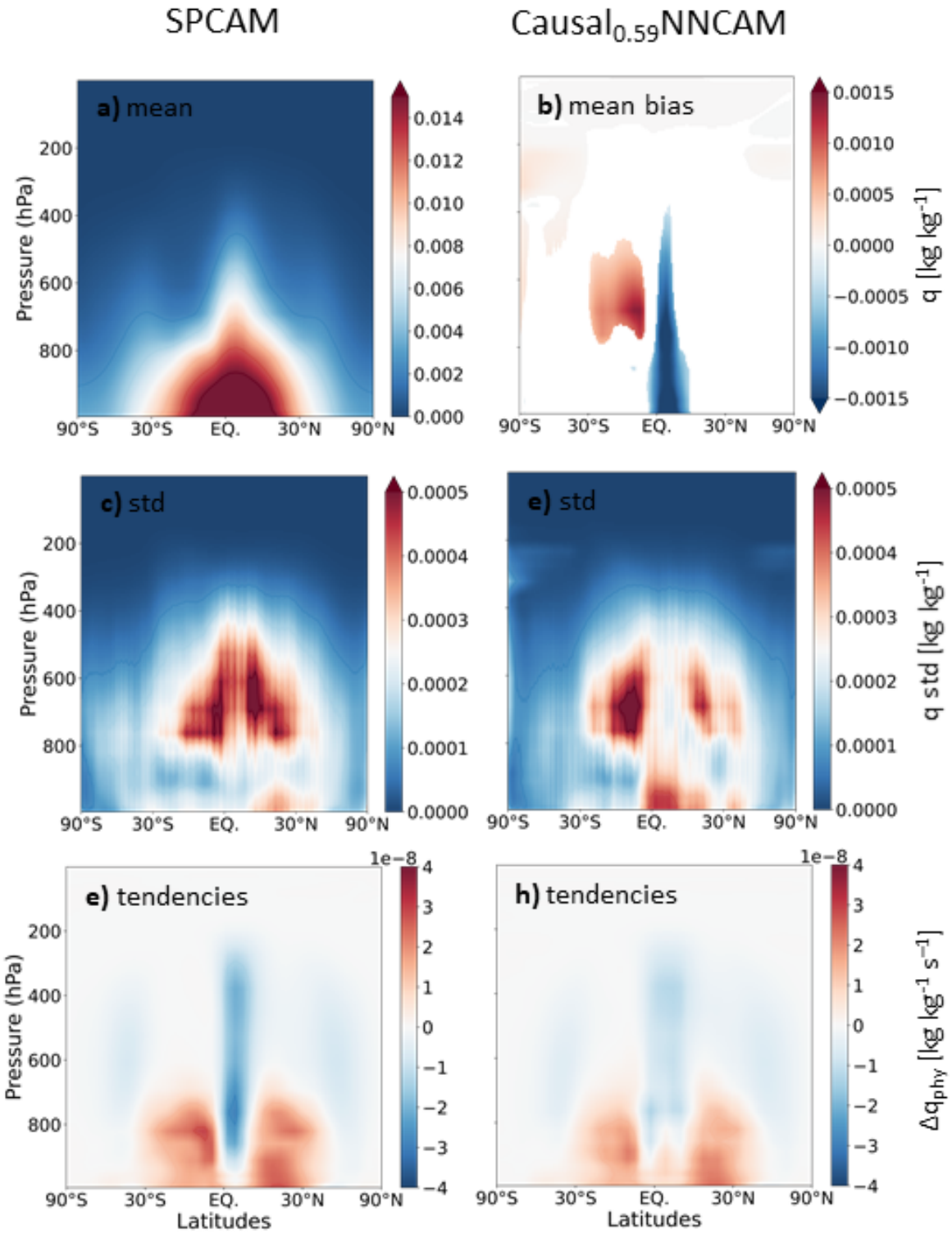


**Figure S6.** Same as Fig. S4, but for Causal$_{0.59}$NNCAM.

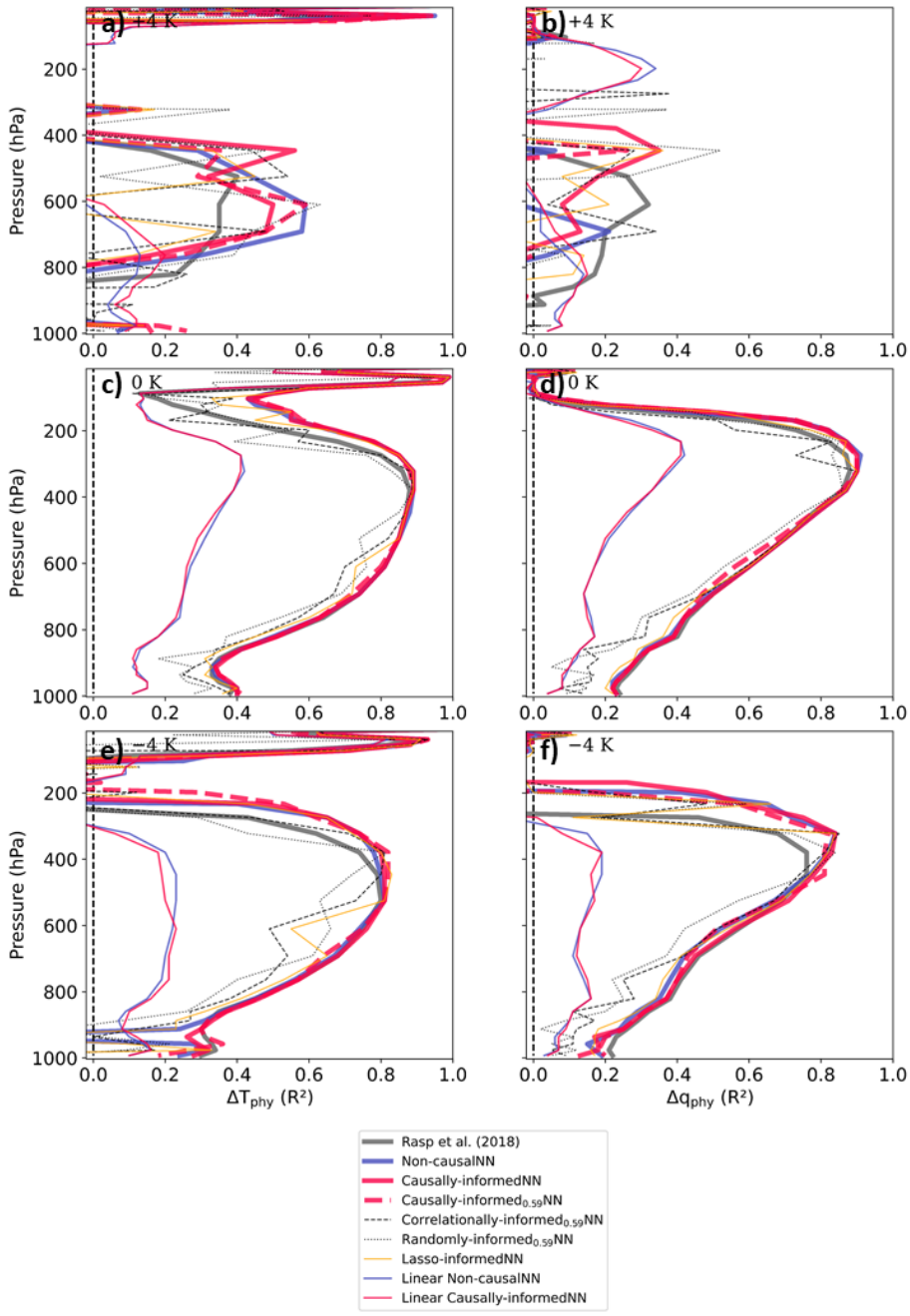


**Figure S7.** Vertically resolved coefficient of determination ($R^2$), averaged horizontally and in time, for heating rates ($\Delta T_{phy}$) and moistening rates ($\Delta q_{phy}$) of neural network (NN) parameterizations trained on the reference climate (+0 K). $R^2$ is calculated using the test sets of each SPCAM simulation case (−4 K, +0 K and +4 K). Linear version of the neural network parameterizations (thin solid lines) are for the activation identity functions. Correlationally-informed (dashed black lines) and randomly-informed (dotted black lines) NNs use the same number of inputs as in the single optimized causal-threshold ($q$=0.59) case. Lasso-informed (solid yellow line) NN use similar number of inputs as in the quantile-optimized causal-threshold case (Fig. S3b).

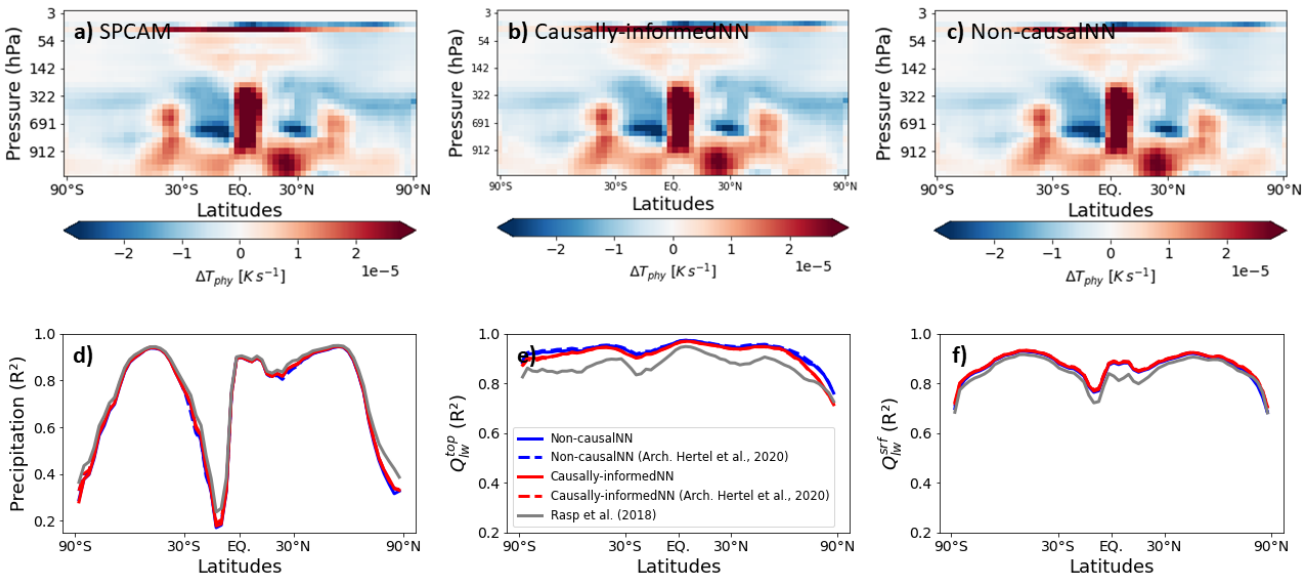


**Figure S8.** (**top row**) Zonal-mean climatologies of heating tendencies ($\Delta T_{phy}$), and (**bottom row**) latitudinally resolved coefficient of determination ($R^2$) of surface precipitation ($P$) and net longwave radiative heat fluxes ($Q_{lw}^{top}$, $Q_{lw}^{srf}$).

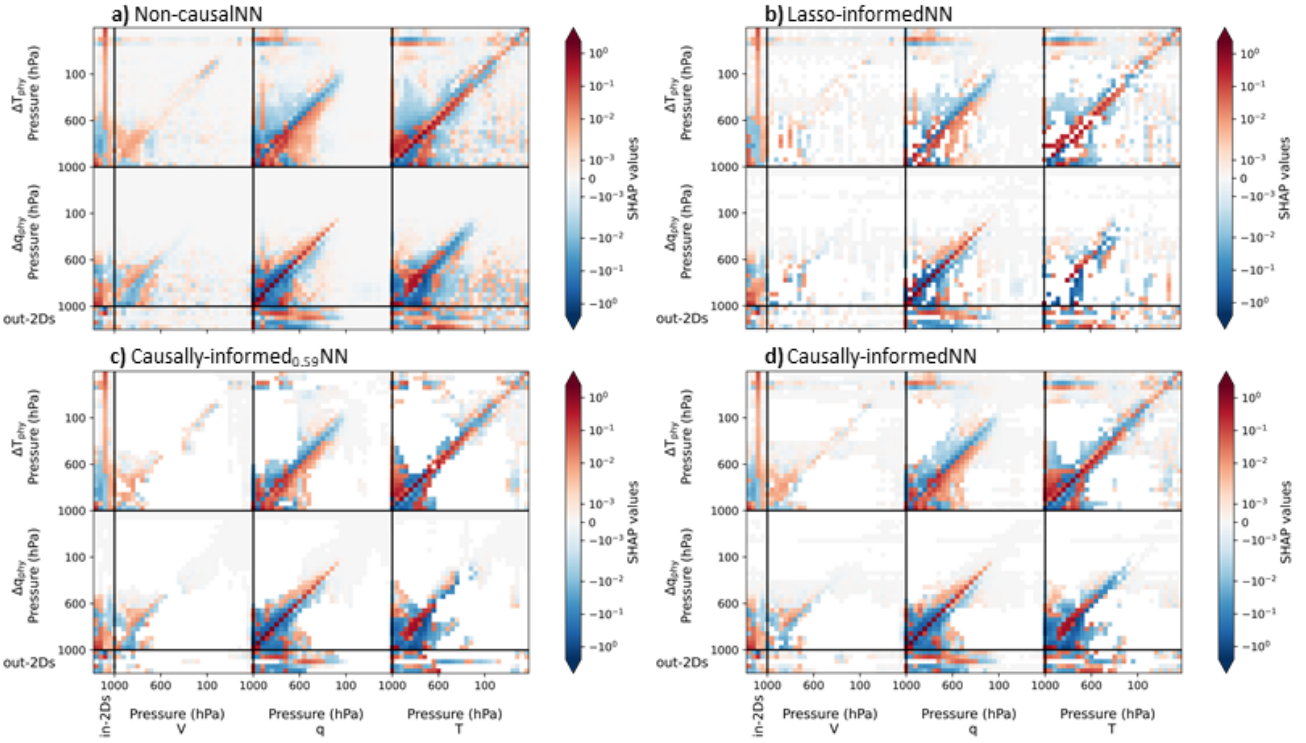


**Figure S9.** Same as Fig. 6, but including mean SHAP value sign for: **a**) Non-causalNN; **b**) Lasso-informedNN (Fig. S3b); **c**) Causally-informed$_{0.59}$NN; and **d**) Causally-informedNN.

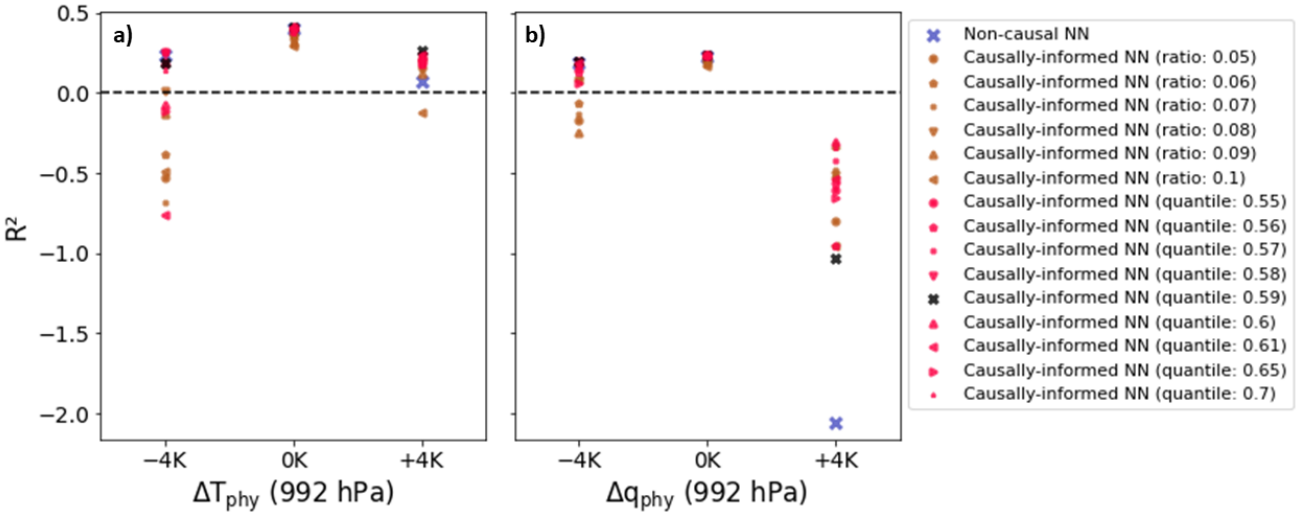


**Figure S10.** Coefficient of determination ($R^2$) of **a**) $\Delta T_{phy}$ and **b**) $\Delta q_{phy}$ at the surface (992 hPa) for the Non-causal neural network (NN) and the causally-informed NN using a number of thresholds for both approaches, ratio- (brown) and quantile-based (red). The reference SPCAM simulation (+0 K) was used for training. $R^2$ is computed using the test sets of each simulation case (−4 K, +0 K and +4 K). The optimal single threshold is marked with a black cross.

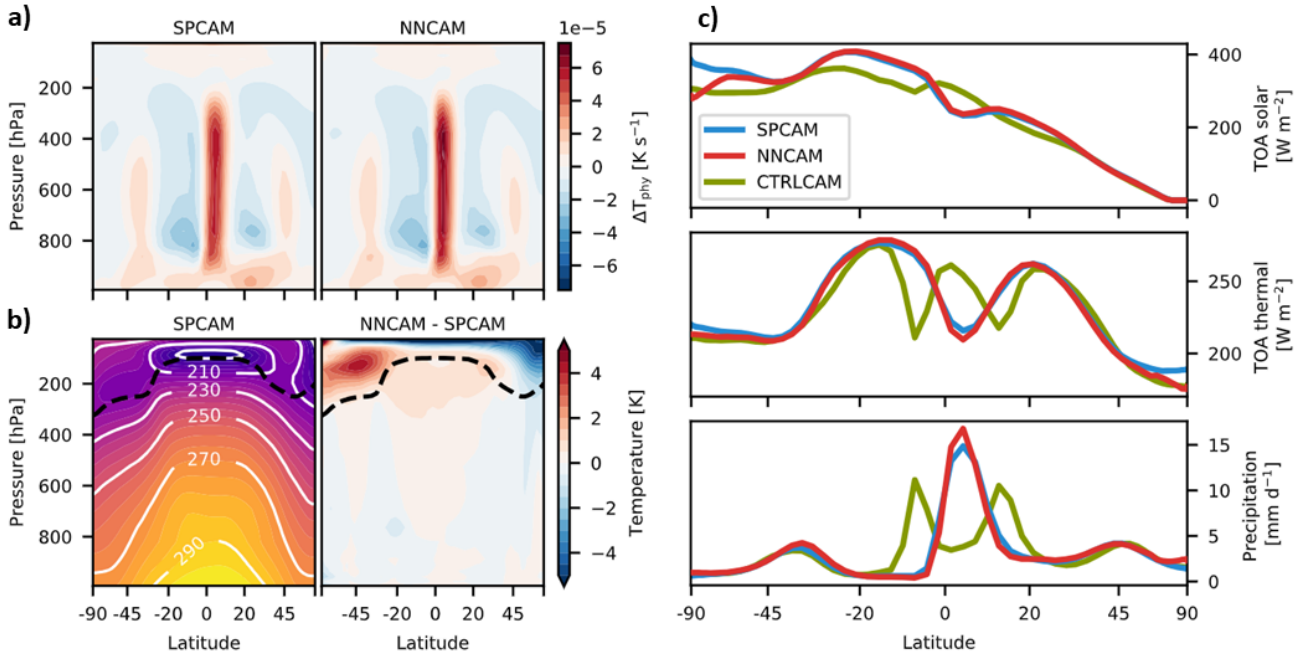


**Figure S11.** Adapted from Fig. 1 in (Rasp et al., 2018). a) Zonal mean convective and radiative subgrid heating rates $\Delta T_{phy}$. b) Zonal mean temperature T of SPCAM and NNCAM biases. Black dashed line shows the mean tropopause. c) Latitudinally resolved mean shortwave and longwave net fluxes at the top of the atmosphere and precipitation. Zonal mean values are area-weighted. 