Supporting Information for “Non-Linear Dimensionality Reduction with a Variational Encoder Decoder to Understand Convective Processes in Climate Models”
Gunnar Behrens 1,2, Tom Beucler 3, Pierre Gentine 2,4, Fernando Iglesias-Suarez 1, Michael Pritchard 5, Veronika Eyring 1,6

1Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
2Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA
3Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, VD 1015, Switzerland
4Earth Institute and Data Science Institute, Columbia University, New York, NY 10027, USA
5Department of Earth System Science, University of California Irvine, Irvine, CA, USA
6University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

Contents of this file
1. Tables S1 to S11
2. Figures S1 to S19

Introduction
The supporting information are structured as follows and each section can be read individually:

In section S.1 we show the hyperparameters of VED and explain how we conducted the search for a suitable set of hyperparameters of VED. Furthermore we discuss the
used VED output normalization dictionary. In section S.2 we show additional figures for the general evaluation of VED and other reference networks. This section further describes differences in reproduction skill if either the VED or the output normalization of Rasp, Pritchard, and Gentine (2018) is used. Furthermore we describe differences in the interpretability between the VED’s latent space and a principal component analysis on the large-scale variables in this section. Also we show that the latent space exploration with conditional averages can be conducted on the five original latent dimension. Section S.3 shows one alternative VED and a conditional VAE structure and discusses their strengths and limitations. We describe in subsection A) the VED\(_{X \rightarrow Y}\) and in subsection B) a conditional VAE (cVAE). Section S.4 comprises the tables of all generated 2D SP or CAM variables with our generative modeling approach. Additionally the squared Pearson correlation coefficients R\(^2\) between the latent nodes and vertical heating, moistening, specific humidity and temperature profiles in space-time and time are shown in this section respectively.
S.1 VED Hyperparameters based on a Hyperparameter Search and Normalisation

In earlier experiments we found that the output normalisation used in Rasp et al. (2018) was not well-suited for the optimization of a VED during training. With their output normalisation dictionary, the VED focused solely on the reproduction of radiative fluxes in $Y$ and lacked skill with respect to heating and moistening profiles. Therefore we had to re-scale the output normalisation dictionary for $Y$ and implement a suitable scaling for the extended variable list $O$. The vertical profiles of temperature, specific humidity and specific humidity tendency are normalised by long-term (3 month) standard deviations of the near surface model level. In the case of temperature tendencies, the standard deviation on the 845 hPa level is used due to the dominant variability of convection related temporal temperature changes on this level near the upper limit of the planetary boundary layer in SP data. The remaining 2D variables of radiative properties, precipitation rates and surface pressure are standardised.

We initially performed a hyperparameter search (random search) with 120 trials for a VAE$_{X \rightarrow X}$, which was trained on large-scale climate variables $X$ to reproduce $X$. Table S2 shows the hyperparameter range for a hyperparameter search over a sequence of 1 month of SP data.

The best-performing encoder and decoder hyperparameter settings; 6 hidden layers, 463 nodes in the first and last hidden layer and a latent space width of 5 nodes; were fixed for the development of the VED presented in the paper. To account for shifts in suitable learning rates and batch size due to the additional sub-grid-scale output variables $Y$, we
conducted a second hyperparameter search (random search) for our main VED specifically over a sequence of 1 month of SP data, see Table S3.

After that, we fixed the initial learning rate and batch size and conducted further sensitivity tests with respect to the latent space width (which are documented in section 3) of VED. The choice of activation functions in the hidden layers is based on small initial experiments with VED, which showed enhanced emulation skill if the last hidden layer was elu-activated (exponential linear unit).

S.2 Evaluation of VED and the Reference Networks

If we use the output normalization of reference ANN to investigate the sensitivity of the VED performance as a function of latent space width, then we observe similar asymptotic behaviour as in Figure 2, see Figure S1. The VED shows an improved emulation skill compared to the reference linear model with fixed layer widths of 256 nodes. The difference in performance between the VED and reference ANN increases if the output normalization of Rasp et al. (2018) is used, which points to the fact that the VED output scaling weights SP variables \( Y \) differently. The VED has a decreased performance compared to reference ANN, but is converging to a similar level of emulation with increasing latent space width.

The latent space of VED can be explored with the computation of conditional averages in a 2D PCA compressed submanifold as it is shown in Figure 5. However this analysis can be complemented with an inspection of the 5 latent dimensions itself. To visualize the five dimensional latent space, we projected two latent variables onto each other. This results in 20 spanned submanifolds of different latent variables and five projections of one latent variable onto itself oriented along the main diagonal in Figure S11 to S14. These 2D submanifolds of two different latent variables are often characterised by two or three
centers of action with a strong concentration of samples (Figure S11). In most cases there is a weak linear connection between the latent variables, except for latent variable 2 (Large-scale variations along mid latitude storm tracks) and latent variable 5 (Deep Convection), as can be seen in Figure S11. The projection of these two latent variables is also characterised by a pronounced separation of samples with negligible convective processes (no precipitation, Figure S12) and deep convective samples. This shows that the convective strength of the samples can be gauged with these two latent variables of VED and is not relying on a PCA as postprocessing step. Moreover the latent variables itself can be utilised to investigate large-scale geographic variability. One particular example for that, is the projection of latent variable 1 (Global temperature variations) and 2 (Large-scale variability along mid latitude storm tracks). In this submanifold we see two separated maxima of solar insolation (Figure S13) and two areas with no solar insolation (night-time conditions). If we compare this distribution to the conditional averages of the surface air temperature (Figure S14), we observe that one solar insolation maximum is associated with a minimum in surface air temperatures below 275K, which can be only observed in polar latitudes. The combination of solar insolation with anomalous cold temperatures is a clear evidence that the respective samples are originating from austral polar or subpolar latitudes due to the austral summer solar forcing of the SPCAM simulations. In contrast the other minimum in surface air temperatures in this projection of latent variable 1 and 2 is associated with no solar insolation. This suggests that the corresponding samples are coming from the boreal high latitudes (due to constant polar night conditions). These two examples illustrate that the interpretation of convective processes and large-scale drivers of convective predictability is possible on the latent variables of VED itself and not relying
on the PCA postprocessing step. Furthermore the latent space of VED can be used to investigate longstanding hypotheses of atmospheric science. As an example we can focus on the projection of latent variable 5 (Deep Convection) onto latent variable 1 (Global temperature variations). The strongest precipitating samples in Figure S12 are situated in the middle of the conditional distribution of latent variable 1 (Figure S11 and not in the right tail of the marginal distribution, which suggests that strong precipitation is not occurring in the regions with the highest surface air temperatures. This hypothesis can be evaluated with Figure S14, where the region with the strongest precipitation is associated with conditional averages of surface air temperatures of around 295K in this projection. These temperatures are around 5K colder than the maximum of the conditional averages seen for this particular projection, which is in agreement with the original hypothesis. Overall these results indicate the power of the VED with respect to the interpretability and meaningfulness of the latent space and stored physical concepts in the lower-order manifold.

S.3 Alternative VED and cVAE Structure

A) VED\(_{X \rightarrow Y}\)

VED\(_{X \rightarrow Y}\) closely mirrors the original SP with similar output variables to those of the reference ANN. It uses a set of convection related CAM climate variables \(X\) as input to the network, except for meridional wind profiles which were additionally used in Rasp et al. (2018). For this variational network, we couple the encoder to a regular feed forward neural net with 3 hidden layers. The resulting variational network VED\(_{X \rightarrow Y}\) (Figure S15) reproduces the convection-related SP output variables \(Y\) used in Rasp et al. (2018). The concatenated output vector \(Y\) has a length of 65 (65 output nodes). It contains the
vertical profiles of temperature $dT(p)/dt$ and specific humidity tendencies $dq(p)/dt$, the shortwave / longwave fluxes at the model top / surface $Q_{sw/lw \ top/surf}$ and the precipitation rate $\text{precip}$. The coupled decoding feed forward neural net has three hidden layers with 353 nodes in each layer. The associated loss function is given in Equation 1.

\[
\text{VED loss}_X \rightarrow Y = \text{reconstruction loss}_X \rightarrow Y + \lambda \ \text{KL loss} \quad (1)
\]

The reconstruction loss (Equation 2) of VED$_X \rightarrow Y$ is defined as the MSE between the emulated $Y_{\text{emul}}$ and $Y$.

\[
\text{reconstruction loss}_X \rightarrow Y = \frac{1}{M} \times \frac{1}{N} \sum_{i=1}^{M=65} \sum_{j=1}^{N=\text{batch size}} (Y_{ij} - Y_{ij}^{\text{emul}})^2 \quad (2)
\]

\[
\text{KL loss} = \frac{1}{2} \times \frac{1}{N} \sum_{j=1}^{N=\text{batch size}} \sum_{k=1}^{K=\text{latent space width}} \left[ -1 - \ln \sigma_{jk}^2 + \mu_{jk}^2 + \sigma_{jk}^2 \right] \quad (3)
\]

\[
\lambda \in \mathbb{R}_+
\quad (4)
\]

The hyperparameters used for VED$_X \rightarrow Y$ are displayed in Table S5, and the model architecture is illustrated in Figure S15.

VED$_X \rightarrow Y$ (test MSE = 0.157) reproduces the mean statistics with increased skill compared to VED (test MSE = 0.165) using the VED output normalization. The emulation skill of the spatio-temporal tropical variability is of the order of that of VED and slightly reduced with respect to reference ANN. However we see a decreased interpretability of the latent space of VED$_X \rightarrow Y$ in comparison to VED, which is a major disadvantage of
the VED\textsubscript{X→Y} network architecture. The 2D PCA compressed latent space of VED\textsubscript{X→Y} generally shows a weak minimum to maximum distribution mostly focusing on the magnitude of convective processes (see Figure S16) and faintly on geographic variability with respect to multiple sub-grid-scale and large-scale climate variables (see Figure S17, as an example for surface air temperatures). Samples from the two poles with anomalously cold surface air temperatures are not well separated in the 2D PCA compressed latent space of VED\textsubscript{X→Y}, in contrast to that seen for VED (see Figure S17). We observe one surface air temperature minimum in the 2D PCA compressed latent space of VED\textsubscript{X→Y}. The minimum comprises samples from the austral high latitudes to the right and from boreal latitudes to the left. These low surface air temperatures are compressed within a very small fraction of the 2D PCA compressed latent space of VED\textsubscript{X→Y} surrounded by mid-latitude temperatures in close distance. For VED we see a clearly improved adaption to these large-scale meridional temperature variations with well separated zones of austral and boreal polar samples. Likewise we see for VED that samples with increased precipitation are concentrated into two centers of action and the 2D PCA compressed latent space illustrates strong gradients with respect to conditional averages of precipitation, which is not the case for VED\textsubscript{X→Y} (Figure S16). This lack of interpretability of the latent space is a general limitation of VED\textsubscript{X→Y} compared to VED or even ED for the identification of driving large-scale climate conditions and related convective processes globally.

B) cVAE

In general, a conditional VAE (cVAE) predicts the distribution of a set of output variables conditioned on the input variables. The general model configuration of cVAE’s enables the propagation of information about the state of output variables and also input
variables through the latent space to the conditional decoder (Sohn et al., 2015). For the task to realistically reproduce \( Y \) and gain insights on the interpretability of the latent space, we construct one possible cVAE. The sub-grid-scale variable vector \( Y \) is fed into the encoder together with large-scale CAM variables \( X \), as can be seen in Figure S15. \( X \) is an additional input to the decoding part of the network. As a result of that the latent space should illustrate a pronounced dependence on the sub-grid-scale input features \( Y \) rather than on large-scale CAM variables \( X \). The cVAE’s loss function is defined as:

\[
cVAE \text{ loss} = \text{reconstruction loss}_{cVAE} + \lambda \text{ KL loss}
\]

(5)

The associated reconstruction loss is defined as the MSE between \( Y_{emul} \) and \( Y \), as can be seen in Equation 6.

\[
\text{reconstruction loss}_{cVAE} = \frac{1}{M} \times \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} (Y_{ij} - Y_{emul}^{ij})^2
\]

(6)

\[
\text{KL loss} = \frac{1}{2} \times \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{K} [-1 - \ln \sigma_{jk}^2 + \mu_{jk}^2 + \sigma_{jk}^2]
\]

(7)

\[
\lambda \in \mathbb{R}_+
\]

(8)

The used hyperparameters are displayed in Table S6 and the model architecture can be seen in Figure S15.

Due to its deviating model architecture in comparison to the constructed VEDs (VED and VED\(_{X \rightarrow Y}\)), which are not trained with SP sub-grid-scale variables \( Y \) as input data, the
cVAE has an advantage against all evaluated models in training mode (during the model optimization). This advantage in training mode reflects in a strongly improved emulation skill of this network compared to the reference ANN. The MSE of cVAE in training mode with respect to SP training, validation or test data (0.049 / 0.050 / 0.050) is more than half as small as the one of reference ANN (0.133 / 0.135 / 0.135) using the VED output normalization. We observe similar emulation capabilities for the related coefficients of determination $R^2$ of the lower tropospheric specific humidity and temperature tendencies. More than 96% of the horizontal grid points have a $R^2$ value larger than 0.7 for 700 hPa temperature tendencies in the case of the cVAE in training mode. For cVAE, only 38% of the grid points exceed a coefficient of determination of 0.7. Nevertheless the emulation capabilities in test mode, where only the CAM climate variables $X$ are fed into cVAE, are remarkably weaker than for all other evaluated networks. This is one clear disadvantage of the “brute-force training strategy” of the cVAE with our architecture, where we train the encoder and decoder together. The strong decrease in emulation skill between training and test mode suggests that the largest portion of optimization goes into the emulation of the sub-grid-scale variables $Y$. Another discouraging point is the overall poorly developed interpretability of the latent space of cVAE with respect to essential sub-grid-scale and climate variables like outgoing longwave radiation, solar insolation or surface air temperature. cVAE is not capable to distinguish between day and nighttime conditions in its latent space. This is a crucial benchmark of all other evaluated models. Overall, cVAE focuses in its latent space exclusively on variations in convective moistening and heating tendencies or the related formation of precipitation. This clearly limits the interpretability of drivers of convective predictability in the latent space of
cVAE. Furthermore it suggests that key information about the background climate state of convective processes are dominantly propagated through the additional link of $X$ to the decoder of cVAE (see Figure S15). This leads to the fact that the encoding of large-scale information in the latent space of cVAE in training mode is clearly outperformed by a traditional PCA on the climate variables $X$. Despite these discouraging results, we think cVAE could be upgraded towards a generative and stochastic parameterization of SP. Pan et al. (2020) described that their initial cVAE structure exhibited large differences in performance between the training and test mode too. Therefore they developed a step-wise concept, where first the decoder is trained on $X$, then the encoder on $Y$ and later the entire network on the complete variable list $O$. With this concept of step-wise training they were able to drastically improve the emulation abilities of their cVAE (Pan et al., 2020). In our case this upgraded training strategy might result in an enhanced interpretability of the latent space of cVAE with respect to large-scale drivers of convective predictability similar to results shown in this study for VED.

S.4 Generated SP/CAM Variables with $z_{\text{translation}}/z_{\text{median}}$ and Squared Pearson Correlation $R^2$ Plots between Latent Nodes and Vertical Profiles

This section comprises the Tables S7-S11 of generated 2D variables in $X$ and $Y$ for each latent node with our generative modeling approach. Additionally the squared Pearson correlation $R^2$ between the Nodes 1 to 5 and vertical profiles of $dq/dt$, $dT/dt$, $q$ and $T$ are displayed for space-time series (Figure S18) or time series (Figure S19) respectively. For Figure S18 the Pearson correlation is computed based on the concatenated space-time series (with the shape $[\text{horizontal grid-cells } H \times \text{ time steps } P, \text{ latent space width } K \text{ or output variable size } M]$) of the latent nodes and profiles in $O$, which means that these
arrays include information about the large-scale geographic variability, e.g., the large meridional temperature and specific humidity contrasts between the tropics and poles. For Figure S19 the Pearson correlation is calculated in each horizontal grid-cell between the time series (with the shape \([P, K \text{ or } M]\) of the latent nodes and output profiles in \(O\). As a second step the median of the Pearson correlation coefficients is calculated across all horizontal grid-cells \(H\).

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. doi: 10.1175/JCLI-D-11-00168.1

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., & Lee, J. (2020). Improving Seasonal Forecast using Probabilistic Deep Learning.

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

Sohn, K., Yan, X., & Lee, H. (2015). Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems, 2015-Janua*, 3483–3491
| Hyperparameter of VED | Values |
|-----------------------|--------|
| Learning Rate         | 0.00074594 |
| Training / learning rate decrease | 40 epochs, learning decrease every 7<sup>th</sup> epoch by factor 5 |
| Batch size            | 714 |
| Latent Space Width    | 5 nodes |
| Node Size of Encoder  | [64,463,463,232,116,58,29,5] |
| Node Size of Decoder  | [5,29,58,116,232,463,463,129] |
| Activation Encoder    | [Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, Lambda] |
| Activation Decoder    | [Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, ELU] |
| KL Annealing          | Linear annealing from 2<sup>nd</sup> to 7<sup>th</sup> epoch |

Table S1. Hyperparameters and architecture of the final VED which uses large-scale CAM variables \( \mathbf{X} \) to investigate simulated convective processes of SP \( \mathbf{Y} \) together with driving climate conditions.

July 7, 2022, 6:23pm
### Table S2.

Hyperparameter range of search for initial VAE, which reproduces large-scale climate variables $X$ with $X$ as input data set. The hyperparameter search was conducted over 120 trials and 30 epochs with a learning rate decrease after every $5^{th}$ epoch by a factor 5.

| Hyperparameter of VAE$_{X\rightarrow X}$ | Values          |
|----------------------------------------|-----------------|
| Initial learning rate                  | $10^{-5}$ to $5 \times 10^{-4}$ |
| Batch size                             | 200 to 8192     |
| Latent Space Width                     | 2 to 5 nodes    |
| Node Size of first or last hidden layer of Encoder or Decoder | 300 to 500       |
| Depth of Encoder or Decoder in hidden layers | 5 to 7 hidden layers |

### Table S3.

Hyperparameter range of search for VED, the main model in this study. The hyperparameter search was conducted over 80 trials and 20 epochs with one learning rate decrease after the $10^{th}$ epoch.

| Hyperparameter of VED | Values          |
|-----------------------|-----------------|
| Initial learning rate | $5 \times 10^{-5}$ to $5 \times 10^{-3}$ |
| Batch size            | 200 to 8000     |
| Network | Training MSE | Validation MSE | Test MSE |
|---------|--------------|----------------|----------|
| VED     | 0.162        | 0.165          | 0.165    |
| EDD     | 0.162        | 0.165          | 0.165    |
| LR      | 0.242        | 0.244          | 0.243    |
| ANN     | 0.133        | 0.135          | 0.135    |

**Table S4.** Mean squared errors (MSE) of predicted sub-grid-scale SP variables $Y$ of the VED, ED, LR, reference ANN on the training, validation and test data sets (3 month of SP data) using the VED output normalization.

**Figure S1.** Similar to Figure 2, mean squared error as a function of Latent Space Width of the VED for the test (solid cyan), validation (dashed-dotted cyan) and training data set (dashed cyan curve) using the output normalization of the reference ANN (Rasp et al., 2018) as y-axis. The horizontal solid blue / black line represents the MSE scores of the reference ANN (Rasp et al., 2018) / a linear version of this network (Reference Linear Model) on test data with fixed layer width of 256 nodes in the 9 hidden layers.
Figure S2. Wheeler Kiladis diagram based on tropical outgoing longwave radiation $[15^\circ N - 15^\circ S]$ of SP (a), of ED predictions (b) and the absolute difference of spatio-temporal wave spectra ED - SP (c) for 1 year of SP data.

Figure S3. Wheeler Kiladis diagram based on tropical outgoing longwave radiation $[15^\circ N - 15^\circ S]$ of SP (a), of reference ANN predictions (b) and the absolute difference of spatio-temporal wave spectra reference ANN - SP (c) for 1 year of SP data.
Figure S4. Fixed Sea Surface Temperature (SST) forcing of the SPCAM simulation following Andersen and Kuang (2012). The blue / red zonal lines indicate the region of Northern / Southern mid latitudes between 60° N/S and 35° N/S. The green lines indicate the deep tropics with the ITCZ between 10° S and 10° N.

Figure S5. Scatter plot with isolines and histograms of the Joint and Conditional distributions of the PCA compressed latent space of VED (left) and ED (right). The plot is based on 100000 randomly picked samples from CAM test data. The 1st / 2nd PC of the resulting compressed latent space is the x-axis / y-axis in the respective subplot.
Figure S6. Latent Space clustering of VED for precipitation (left), outgoing longwave radiation ($Q_{lw\ top}$) (left middle), shortwave heat flux at the model top ($Q_{sw\ top}$) (right middle) and Surface Air Temperature ($T_{surf}$) (right column). The first row illustrates the clustering in the PCA compressed latent space with respect to the SP / CAM variables on global scales (as seen in Figure 4). The lower rows depict the Latent Space clustering in the evaluated regions Northern Mid Latitudes (2nd row), Tropics (3rd row) and Southern Mid Latitudes (4th row). The x-axis / y-axis represents the 1st / 2nd leading PC of the global / regional latent space.
Figure S7. Latent Space clustering of ED for precipitation (left), outgoing longwave radiation ($Q_{lw, top}$) (left middle), shortwave heat flux at the model top ($Q_{sw, top}$) (right middle) and Surface Air Temperature ($T_{surf}$) (right column). The first row illustrates the clustering in the PCA compressed latent space with respect to the SP / CAM variables on global scales (as seen in Figure 4). The lower rows depict the Latent Space clustering in the evaluated regions Northern Mid Latitudes (2nd row), Tropics (3rd row) and Southern Mid Latitudes (4th row). The x-axis / y-axis represents the 1st / 2nd leading PC of the global / regional latent space.

July 7, 2022, 6:23pm
Figure S8. Conditional averages of solar insolation (upper left), precipitation (upper right), outgoing longwave radiation (lower left) and surface air temperature (lower right panel) in the submanifold spanned by the first two leading PC’s of the large-scale variables $X$. Similar to Figure 5 the conditional averages are computed based on 1000000 randomly selected samples from the test data set.
Figure S9. Latitude-Longitude plot of the first (upper left, PC1) and second (upper right, PC2) leading principal component of a PCA on the large-scale variables $X$ and respective large-scale and sub-grid-scale variables of the test data set for a particular time step.
Figure S10. Latitude-Longitude plot of the latent variables of the VED (Latent Node 1 to 5) and respective large-scale and sub-grid-scale variables of the test data set for the same time step as in Figure S9.
Figure S11. 2D Density (blue contours) and scatter (light green) plots of the projection of two latent variables of the five dimensional latent space of VED in combination with the marginal distributions of the two latent variables. The first row represents the projection of Node 1 (Global temperature variations) onto all other four latent variables and itself. The second row shows the plots for Node 2 (Large-Scale variations along mid latitude storm tracks). The third row represents Node 3 (Shallow Convection). The forth and fifth row shows the plots for Node 4 (Mid latitude storm track) and Node 5 (Deep Convection). All plots are based on 50000 randomly selected samples from the test data.
Figure S12. Similar to Figure S11 a projection of one latent variable on all other latent variables and itself. The color coding reveals the conditional average of precipitation based on 1000000 randomly selected samples from the test data set. The first row shows the 2D projection of the first latent variable (Node 1, Global Temperature Variations) and all other latent variables. The second / third / fourth and fifth row depicts the projections of Node 2 (Large-Scale Variations along the mid latitude storm tracks) / Node 3 (Shallow Convection) / Node 4 (Mid latitude storm tracks) and Node 5 (Deep Convection).
Figure S13. Similar to Figure S12, but for the conditional averages of solar insolation in the projections.
Figure S14. Similar to Figure S12, but for the conditional averages of the surface air temperature in the projections.
| Hyperparameter | Values |
|----------------|--------|
| Learning Rate  | 0.00018238 |
| Training / learning rate decrease | 40 epochs, learning decrease every 7<sup>th</sup> epoch by factor 5 |
| Batch size     | 714    |
| Latent Space Width | 5 nodes |
| Node Size of Encoder | [64,463,463,232,116,58,29,5] |
| Node Size of Decoder | [5,353,353,353,65] |
| Activation Encoder | [Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, Lambda] |
| Activation Decoder | [Input, ReLU, ReLU, ReLU, ELU] |
| KL Annealing   | Linear annealing from 2<sup>nd</sup> to 7<sup>th</sup> epoch |

**Table S5.** Hyperparameters and architecture of the constructed \( \text{VED}_{X \rightarrow Y} \) which uses large scale CAM variables \( X \) to simulate SP variables \( Y \).
Figure S15. Combined schematic of the architecture of VED (green), VED\(_{X \rightarrow Y}\) (light blue) and cVAE (purple arrows and network parts). The network structures in light blue are used for all variational networks with varying hyperparameters.

Figure S16. The 2D PCA compressed latent space of the VED (left) and VED\(_{X \rightarrow Y}\) (right panel) and associated conditional average of precipitation of projected SP test data (similar to Figure 5). The x-axis / y-axis in all subplots indicates the 1\(^{st}\) / 2\(^{nd}\) leading PC of the 5D latent space in the respective panels.
Figure S17. The 2D PCA compressed latent space of the VED (left) and $\text{VED}_{X \rightarrow Y}$ (right panel) and associated conditional average of surface air temperature of projected SP test data (similar to Figure 5). The x-axis / y-axis in all subplots indicates the $1^{st}$/ $2^{nd}$ leading PC of the 5D latent space in the respective panels.
| Hyperparameter          | Values                                                                 |
|------------------------|------------------------------------------------------------------------|
| Learning Rate          | 0.00096133                                                             |
| Training / learning    | 40 epochs, learning decrease every 7th epoch by factor 5               |
| Batch size             | 666                                                                    |
| Latent Space Width     | 5 nodes                                                                |
| Node Size of Encoder   | [[65,64],457,457,228,114,57,29,5]                                      |
| Node Size of Decoder   | [5,29,57,114,228,457,457,65]                                           |
| Activation Encoder     | [Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, Lambda]                    |
| Activation Decoder     | [Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, ELU]                       |
| KL Annealing           | Linear annealing from 2nd to 7th epoch                                 |

**Table S6.** Hyperparameters and architecture of the constructed cVAE which uses sub-grid-scale SP variables $Y$ and large-scale CAM variables $X$ to simulate sub-grid-scale SP variables $Y$. 

July 7, 2022, 6:23pm
### Latent Node 1

| Global Temperature variations | 10\(^{th}\) perc | 25\(^{th}\) perc | 50\(^{th}\) perc | 75\(^{th}\) perc | 90\(^{th}\) perc |
|------------------------------|------------------|------------------|------------------|------------------|------------------|
| \(Q_{sw\ top}\left[\frac{W}{m^2}\right]\) | 4                | 115              | 451              | 36               | 6                |
| \(Q_{sw\ surf}\left[\frac{W}{m^2}\right]\) | -1               | 49               | 284              | 24               | 1                |
| \(Q_{lw\ top}\left[\frac{W}{m^2}\right]\) | 181              | 221              | 241              | 260              | 275              |
| \(Q_{lw\ surf}\left[\frac{W}{m^2}\right]\) | 55               | 12               | 28               | 60               | 44               |
| \(\text{precip}\left[\frac{mm}{h}\right]\) | 0.11             | 0.01             | 0.03             | 0.12             | 0.07             |
| \(P_{surf}\left[hPa\right]\) | 933              | 982              | 995              | 995              | 989              |
| \(Q_{sol}\left[\frac{W}{m^2}\right]\) | 15               | 263              | 748              | 45               | 8                |
| \(Q_{sens}\left[\frac{W}{m^2}\right]\) | 25               | 9                | 3                | 9                | 12               |
| \(Q_{lat}\left[\frac{W}{m^2}\right]\) | 52               | 19               | 39               | 85               | 163              |

**Table S7.** Generated shortwave and longwave heat flux at the model top / surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of \(z_{\text{median}}\) (4\(^{th}\) column, 50\(^{th}\) perc) and \(z_{\text{translation}}\) of the 10\(^{th}\), 25\(^{th}\), 75\(^{th}\) and 90\(^{th}\) percentile of latent node 1 (Global Temperature variations).
Latent Node 2

| Large-scale variability along the mid latitude storm tracks | 10\(^{th}\) perc | 25\(^{th}\) perc | 50\(^{th}\) perc | 75\(^{th}\) perc | 90\(^{th}\) perc |
|-----------------------------------------------------------|------------------|-----------------|-----------------|-----------------|-----------------|
| \(Q_{\text{sw top}} [\text{W m}^{-2}]\)                  | 987              | 1092            | 451             | 57              | 158             |
| \(Q_{\text{sw surf}} [\text{W m}^{-2}]\)                 | 773              | 845             | 284             | 30              | 49              |
| \(Q_{\text{lw top}} [\text{W m}^{-2}]\)                  | 252              | 249             | 241             | 205             | 173             |
| \(Q_{\text{lw surf}} [\text{W m}^{-2}]\)                 | 85               | 73              | 28              | 44              | 13              |
| \(\text{precip} [\text{mm h}^{-1}]\)                     | -0.01            | 0.00            | 0.03            | 0.12            | 0.15            |
| \(P_{\text{surf}} [\text{hPa}]\)                         | 983              | 989             | 995             | 993             | 992             |
| \(Q_{\text{sol}} [\text{W m}^{-2}]\)                     | 1214             | 1347            | 748             | 125             | 443             |
| \(Q_{\text{sens}} [\text{W m}^{-2}]\)                    | 19               | 9               | 3               | 6               | 23              |
| \(Q_{\text{lat}} [\text{W m}^{-2}]\)                     | 83               | 55              | 39              | 86              | 101             |

**Table S8.** Generated shortwave and longwave heat flux at the model top / surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of \(z_{\text{median}}\) (4\(^{th}\) column, 50\(^{th}\) perc) and \(z_{\text{translation}}\) of the 10\(^{th}\), 25\(^{th}\), 75\(^{th}\) and 90\(^{th}\) percentile of latent node 2 (Large-scale variability along mid latitude storm tracks).
| Latent Node 3 |  |
|----------------|---|
| Shallow Node 3 | Convection |
| $Q_{sw\ top} \ [\frac{W}{m^2}]$ | 5  
| $Q_{sw\ surf} \ [\frac{W}{m^2}]$ | 2  
| $Q_{lw\ top} \ [\frac{W}{m^2}]$ | 178  
| $Q_{lw\ surf} \ [\frac{W}{m^2}]$ | 3  
| precip $\ \ [\frac{mm}{h}]$ | 0.08  
| $P_{surf} \ [hPa]$ | 985  
| $Q_{sol} \ [\frac{W}{m^2}]$ | 15  
| $Q_{sens} \ [\frac{W}{m^2}]$ | -15  
| $Q_{lat} \ [\frac{W}{m^2}]$ | -33  
| $Q_{sw\ top}$ | 134  
| $Q_{sw\ surf}$ | 49  
| $Q_{lw\ top}$ | 220  
| $Q_{lw\ surf}$ | 8  
| precip | 0.05  
| $P_{surf}$ | 999  
| $Q_{sol}$ | 329  
| $Q_{sens}$ | -11  
| $Q_{lat}$ | -13  
| $Q_{sw\ top}$ | 451  
| $Q_{sw\ surf}$ | 284  
| $Q_{lw\ top}$ | 241  
| $Q_{lw\ surf}$ | 28  
| precip | 0.03  
| $P_{surf}$ | 995  
| $Q_{sol}$ | 748  
| $Q_{sens}$ | 3  
| $Q_{lat}$ | 39  
| $Q_{sw\ top}$ | 1200  
| $Q_{sw\ surf}$ | 925  
| $Q_{lw\ top}$ | 251  
| $Q_{lw\ surf}$ | 73  
| precip | 0.04  
| $P_{surf}$ | 991  
| $Q_{sol}$ | 1488  
| $Q_{sens}$ | 30  
| $Q_{lat}$ | 207  
| 90th perc | 1112  
| 50th perc | 838  
| 75th perc | 255  
| 90th perc | 72  

Table S9. Generated shortwave and longwave heat flux at the model top and surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of $z_{median}$ (4th column, 50th perc) and $z_{translation}$ of the 10th, 25th, 75th and 90th percentile of latent node 3 (Shallow Convection).
Table S10. Generated shortwave and longwave heat flux at the model top / surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of \( z_{\text{median}} \) (4\textsuperscript{th} column, 50\textsuperscript{th} perc) and \( z_{\text{translation}} \) of the 10\textsuperscript{th}, 25\textsuperscript{th}, 75\textsuperscript{th} and 90\textsuperscript{th} percentile of latent node 4 (Mid latitude frontal systems).
### Latent Node 5

|                | 10$^{th}$ perc | 25$^{th}$ perc | 50$^{th}$ perc | 75$^{th}$ perc | 90$^{th}$ perc |
|----------------|----------------|----------------|----------------|----------------|----------------|
| $Q_{sw \, \text{top}}$ [W m$^{-2}$] | 206            | 577            | 451            | 172            | 7             |
| $Q_{sw \, \text{surf}}$ [W m$^{-2}$] | 80             | 327            | 284            | 109            | 0             |
| $Q_{lw \, \text{top}}$ [W m$^{-2}$] | 188            | 208            | 241            | 254            | 266           |
| $Q_{lw \, \text{surf}}$ [W m$^{-2}$] | 24             | 26             | 28             | 93             | 113           |
| precip [mm h$^{-1}$] | 0.60           | 0.24           | 0.03           | 0.01           | -0.01         |
| $P_{surf}$ [hPa]       | 989            | 989            | 995            | 999            | 998           |
| $Q_{sol}$ [W m$^{-2}$] | 489            | 1036           | 748            | 264            | 4             |
| $Q_{sens}$ [W m$^{-2}$] | 4              | 2              | 3              | 3              | 6             |
| $Q_{lat}$ [W m$^{-2}$] | 57             | 51             | 39             | 64             | 80            |

**Table S11.** Generated shortwave and longwave heat flux at the model top / surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of $z_{\text{median}}$ ($4^{th}$ column, 50$^{th}$ perc) and $z_{\text{translation}}$ of the 10$^{th}$, 25$^{th}$, 75$^{th}$ and 90$^{th}$ percentile of latent node 5 (Deep Convection).
Figure S18. Squared Pearson correlation coefficient (linear explained variance) $R^2$ between the latent nodes of VED and predicted vertical profiles of specific humidity tendency ($dq/dt$), temperature tendency ($dT/dt$), specific humidity ($q$) and temperature ($T$) in space-time (which features the large meridional gradients of $q$ and $T$). The light blue line resembles the $R^2$ value for latent node 1 / Global Temperature variations. The dark blue / black / dark cyan / bronze curve denotes the explained variance of latent node 2 (Large-scale variability along storm tracks) / 3 (Shallow Convection) / 4 (Mid latitude frontal system) / 5 (Deep Convection).
Figure S19. Median Squared Pearson correlation coefficient (linear explained variance) $R^2$ between the latent nodes of VED and predicted vertical profiles of specific humidity tendency ($dq/dt$), temperature tendency ($dT/dt$), specific humidity ($q$) and temperature ($T$) in time (without large meridional gradients of $q$ and $T$). The light blue line resembles the median $R^2$ value for latent node 1 / Global Temperature variations. The dark blue / black / dark cyan / bronze curve denotes the median explained variance of latent node 2 (Large-scale variability along storm tracks) / 3 (Shallow Convection) / 4 (Mid latitude frontal systems) / 5 (Deep Convection).