AI Emulation of Stochastic Sudden Stratospheric Warming with Interpretable Latent Structure

C. Daniel Boscu[1*], Daniel Hernandez[1*], Fabio Alvarez Ventura[1*], Justin Finkel[1,4], Ashesh Chattopadhyay[2], Pedram Hassanzadeh[1,3], and Dorian S. Abbot[1]

*These authors contributed equally to this work.

[1]Department of Geophysical Sciences, University of Chicago, Chicago, 60637, IL
[2]Department of Applied Mathematics, University of California, Santa Cruz, Santa Cruz, 95064, CA
[3]Committee on Computational and Applied Mathematics, University of Chicago, Chicago, 60637, IL
[4]Data Science Institute, University of Chicago, Chicago, 60637, IL

Key Points:

- A ResNet-inspired Conditional VAE faithfully emulates the stochastic Holton–Mass model, accurately capturing short-term forecasts, steady-state statistics, and rare SSW transitions.
- The emulator reproduces key SSW statistics, including return periods, committor probabilities, and lead times, with close agreement to the physical model.
- PCA of the learned latent space reveals four well-separated, physically interpretable dynamical regimes, demonstrating that the model internalizes the metastable structure and transition pathways of the system.

Corresponding author: Dorian S. Abbot, abbot@uchicago.edu

Abstract

Extreme weather events like sudden stratospheric warmings (SSWs) are rare yet impactful, and pose significant modeling challenges due to their infrequent occurrence in historical data. AI-based emulators offer a fast and data-efficient alternative to traditional numerical models, so long as they can reliably represent internal variability, in particular regime transitions. In this study, we develop a deep learning architecture tailored to emulate the stochastic Holton-Mass model of stratospheric variability, and investigate its latent space. The model has two stable regimes: a strong and a weak polar vortex state arising from nonlinear wave-mean flow interactions. Weak stochastic forcing excites rare transitions between the two states. The architecture we use is a ResNet-inspired Conditional Variational Autoencoder (CVAE) with 6-layer encoder/decoder stacks and state conditioning. To handle stochasticity, we use a KL-annealed training procedure and a specially weighted loss function that balances reconstruction and latent regularization. The emulator, trained on 300,000 days of simulation data, faithfully represents model dynamics, including steady-state distributions as well as regime transition rates and precursors. Principal Component Analysis (PCA) of the latent space reveals a striking separation into clusters corresponding to four physical regimes: strong vs. weak polar vortex, and stable vs. transition-prone. This degree of unsupervised regime separation in latent space is rare for deep generative models, particularly in high-dimensional, stochastic systems. This work contributes a scalable, interpretable emulator architecture for stochastic climate dynamics and introduces a latent space probing framework for diagnosing what AI models internalize about rare events. Our findings suggest that effectively designed deep emulators can not only accelerate simulation but may also uncover physically meaningful manifolds of variability through latent space interrogation. Future directions include incorporating rare-event sampling and developing disentangled latent models to further enhance interpretability and control.

1 Plain Language Summary

Sudden Stratospheric Warmings (SSWs) are disruptions of the stratospheric polar vortex, which are intermittent, difficult to predict, and impactful for extreme winter weather. Seeking to advance computationally efficient and interpretable data-driven forecasting for this important system, we trained a generative Artificial Intelligence emulator on an idealized stratospheric model. Visualization of the latent space revealed well-separated clusters corresponding to the system's current and near-future states (strong vs. weak vortex). Although generative mdoels are generally seen as “black boxes”, our study introduces an architecture and analysis method that may generalize to other applications.

2 Introduction

The stratospheric polar vortex, although neglected by most early-generation physical models, is an important component of polar and midlatitude winter climate, especially over subseasonal-to-seasonal timescales. This is most dramatically seen in sudden stratospheric warming (SSW) events, in which the breakdown of the polar vortex alters the jet stream configuration (Baldwin & Dunkerton, 2001) and can trigger extreme cold outbreaks. Accurate representation of these rare regime transitions requires incorporating stratospheric dynamics (Charlton & Polvani, 2007).

Vortex disruptions are linked to outbreaks of extreme weather, including severe cold snaps (Kautz* et al., 2020), making its early and accurate prediction a key goal for forecasting. Typically, the polar vortex is simulated using discretized physical equations, an example being the early Holton-Mass (HM) model (Holton & Mass, 1976), which captures aspects of stratospheric variability through a wave-mean flow interaction. The HM model provides a robust physical understanding, but can be computationally expensive for answering detailed statistical questions such as rare event probabilities. More realistic models suffer this problem even more, limiting their utility for rapid ensemble forecasting or long-term climate projections. This computational bottleneck highlights the need for efficient alternatives.

In recent years, Machine Learning (ML) has emerged as a framework to accelerate complex simulations. This paper explores the possibility of applying ML to emulate the behavior of the polar vortex, starting from the HM model but with potential for scaling to more complex models. We aim to determine whether an ML approach can improve forecasting speed while maintaining the accuracy required for practical use.

We use the same model version as Finkel et al. (2021). The state of the system is expressed as a high dimensional vector $\mathbf{X}(t)$, which encodes a complex-valued perturbation streamfunction Ψ and the zonal wind U, discretized into 25 vertical levels with equal widths in the log-pressure coordinate z:

$$\begin{aligned}\mathbf{X}(t) = [&\mathrm{Re}\{\Psi(\Delta z,t)\},\ldots,\mathrm{Re}\{\Psi(z_{\mathrm{top}}\}-\Delta z,t),\\ &\mathrm{Im}\{\Psi(\Delta z,t)\},\ldots,\mathrm{Im}\{\Psi(z_{\mathrm{top}}-\Delta z,t)\},\\ &U(\Delta z,t),\ldots,U(z_{\mathrm{top}}-\Delta z,t)] \in \mathbb{R}^d = \mathbb{R}^{75}\end{aligned} \tag{1}$$

Physically, each entry of $\mathbf{X}$ is a Fourier coefficient for a horizontally varying field with single wavenumbers in the zonal and meridional directions, with detailed description and analysis available in Holton and Mass (1976); Yoden (1987); Finkel et al. (2021), and Finkel et al. (2022).

The system has two stable equilibria: a strong and a weak vortex state, arising from nonlinear wave–mean flow interactions and a height-dependent radiative cooling (Holton & Mass, 1976).

The addition of stochastic forcing to represent fast, unresolved processes like gravity waves enough to drive occasional transitions between the two states, with the strong-to-weak zonal wind transition qualitatively representing an SSW event(Birner & Williams, 2008; Finkel et al., 2021).

When the stochastic forcing is weak, $U(z)$ has a strongly bimodal probability distribution functions (PDFs) for $z \gtrsim 10$ km. We define the strong and weak states A and B based on $U(30$ km): We label these regimes A and B and define them by zonal-wind thresholds:

$$A = \{\mathbf{X} : U(\mathbf{X})(30\,\mathrm{km}) \geq u_A := 53.8\,\mathrm{m\,s^{-1}}\} \tag{2}$$

$$B = \{\mathbf{X} : U(\mathbf{X})(30\,\mathrm{km}) \leq u_B := 21.4\,\mathrm{m\,s^{-1}}\}. \tag{3}$$

The thresholds u_A and u_B are chosen based on the stable equilibria of the unforced system, which we sometimes refer to as "(equilibrium) point (**a** or **b**)", but the results do not depend sensitively on the thresholds.

To capture the effects of stochasticity, we need a stochastic machine learning model. A deterministic one would underestimate the variability and just collapse into one of these states depending on the initial condition. We use a Conditional Variational Autoencoder (CVAE) to model the conditional probability distribution $P(x_{t+1} \mid x_t)$, producing a distribution of plausible future states (Chattopadhyay et al., 2023). The latent space provides a probabilistic representation of unresolved variability, while conditioning on the input state x_t preserves physical context. We apply KL annealing and posterior-collapse mitigation (Wang et al., 2023) to balance reconstruction accuracy with stochastic variability, reproducing the rare regime transitions characteristic of SSWs.

Beyond accurate emulation, we seek to interpret the model's representation by examining the latent space. Through principal component analysis of the latent space, we find that different dynamical regimes naturally separate into clusters corresponding to strong-vortex, weak-vortex, and pre-transition states. The model distinguishes physically distinct behaviors without explicit supervision. This structure allows us to probe how the emulator captures regime transitions and provides insight into what the model has learned about the system's stochastic dynamics.

3 Methods

The following section presents our machine learning pipeline, including the training dataset (generated from the Holton-Mass model); the emulator's architecture (a ResNet-Inspired Convolutional Variational AutoEncoder(CVAE)); and some tailored training methods to jointly optimize the emulator's short-term forecast skill and long-term statistics.

3.1 Data

The training data is a 3×10^5-day simulation of the Holton-Mass model (Finkel et al., 2021), using the Euler-Maruyama stochastic integrator with a numerical time step of 0.005 days and an output frequency of once per day.

We classify each day in the $3{\times}10^5$-day time series of both the Holton-Mass Model and the Emulator into one of four dynamical categories: A, B, C, $A \to B$, and $B \to A$. The latter two categories, "transition points", are defined as follows. A day belongs to the $A \to B$ (AB) set if

1. $U_t \geq u_A$,
2. $U_{t+1} \leq u_A$
3. The system next visits B before returning to A. More formally, if $\tau_A = \min\{s \geq t : U_s \geq u_A\}$ and $\tau_B = \min\{s \geq t : U_s \leq U_B\}$ denote the next-return time to A and B respectively after t, then $\tau_B < \tau_A$.

Similarly, $B \to A$ (BA) points must exit B on the next timestep and then complete the transition into A. These requirements ensure that AB and BA samples represent genuine regime transition events rather than brief threshold excursions. The initial two categories, "A" and "B," represent the equilibrium points we defined earlier, and will never meet the third requirement but might occasionally meet the second if close to the corresponding state boundary. If $u_A > U_t > u_B$, then the day is contained in the set C, which corresponds to the region C that was defined in section 2, the introduction.

3.2 Model Architecture

3.2.0.1 Motivation. The Holton–Mass (HM) system is stochastic and metastable, so a useful emulator must output a *distribution* over next states rather than a single point estimate. We therefore model the conditional law $p_\theta(x_{t+1} \,|\, x_t)$ with a Conditional Variational Autoencoder (CVAE). The CVAE provides (i) a principled probabilistic formulation (ELBO training) and (ii) a minimal, easy-to-train architecture that nonetheless reproduces rare transitions and long-term statistics. Among probabilistic sequence models, this is the simplest configuration we found that achieves strong short-term forecast skill and accurate steady-state densities without heavy architectural complexity.

3.2.0.2 Notation. Let $x_t = [\Psi_t, U_t] = [\mathrm{Re}\{\Psi_t\}, \mathrm{Im}\{\Psi_t\}, U_t] \in \mathbb{R}^{75}$ denote the HM state at day t (Section 1). The emulator learns an encoder $q_\phi(z \mid x_t)$ with latent $z \in \mathbb{R}^{32}$ and weights ϕ, and a decoder $p_\theta(x_{t+1} \,|\, \Psi_t, z)$ with weights θ.

3.2.0.3 Conditioning mechanism. We implement conditioning by *concatenation*: the sampled latent z is concatenated with Ψ_t before entering the decoder. This injects physical context directly into the generative path with minimal overhead and proved more

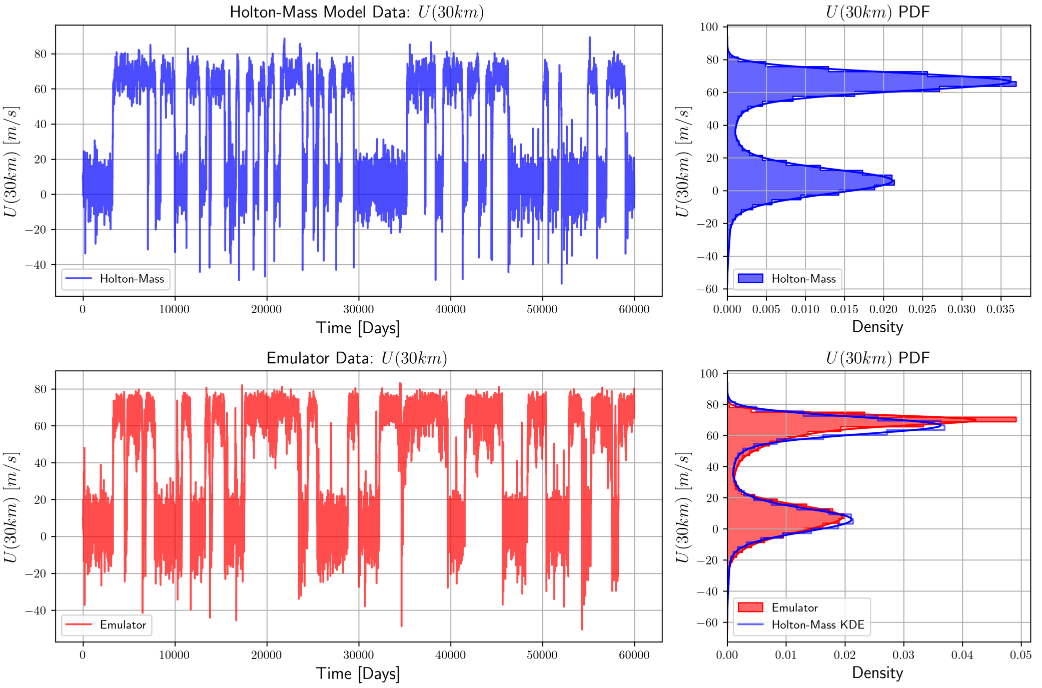


Figure 1. Time series forecast and Probability Density Function(PDF) of $U(30\text{ km})$, over 60,000 days and 10^6 days, respectively, for the Holton-Mass Model (blue) and the emulator (red). The PDF of the emulator is accompanied by the kernel density estimate (KDE) of the Holton-Mass model's PDF over the same 10^6 days.

stable than conditional priors or feature-wise modulation in our setting. We deliberately exclude U from the conditioning input for improved performance, as found in different experimental setups.

3.2.0.4 Residual network structure. Both the encoder and decoder adopt a residual multilayer perceptron (ResNet-MLP) design. Each layer applies a linear transformation followed by a ReLU activation and an identity skip connection, producing a deep but numerically stable mapping that preserves gradient flow through six layers. Specifically, the encoder transforms the input state $x_t \in \mathbb{R}^{75}$ through six hidden layers of width 1024, each followed by a residual addition:

$$x_{l+1} = \text{ReLU}(W_l x_l + b_l) + x_l, \tag{4}$$

ensuring that information from earlier layers is directly available to later ones. This residual coupling prevents vanishing gradients and allows the encoder to learn smooth transformations that capture nonlinear dependencies among vertical levels of the Holton–Mass state.

The final two heads output the mean $\mu(x_t) \in \mathbb{R}^{32}$ and log-variance $\log \sigma^2(x_t) \in \mathbb{R}^{32}$ defining the approximate posterior $q_\phi(z \mid x_t)$. An analogous residual structure is used in the decoder: after concatenating z with the conditioning vector Ψ_t, the combined input is passed through six residual fully connected layers of width 1024, each computing

$$h_{l+1} = \text{ReLU}(W_l h_l + b_l) + h_l, \tag{5}$$

culminating in a linear output layer that predicts the next state $\hat{x}_{t+1} \in \mathbb{R}^{75}$.

This residual MLP configuration offers three key benefits: (i) improved gradient propagation and training stability compared with plain feed-forward stacks, (ii) the ability to represent both near-identity and highly nonlinear mappings without tuning layer depth, and (iii) empirical robustness, with training converging reliably without normalization layers or dropout. Together, these properties make the ResNet-style encoder–decoder an efficient and stable backbone for the conditional VAE.

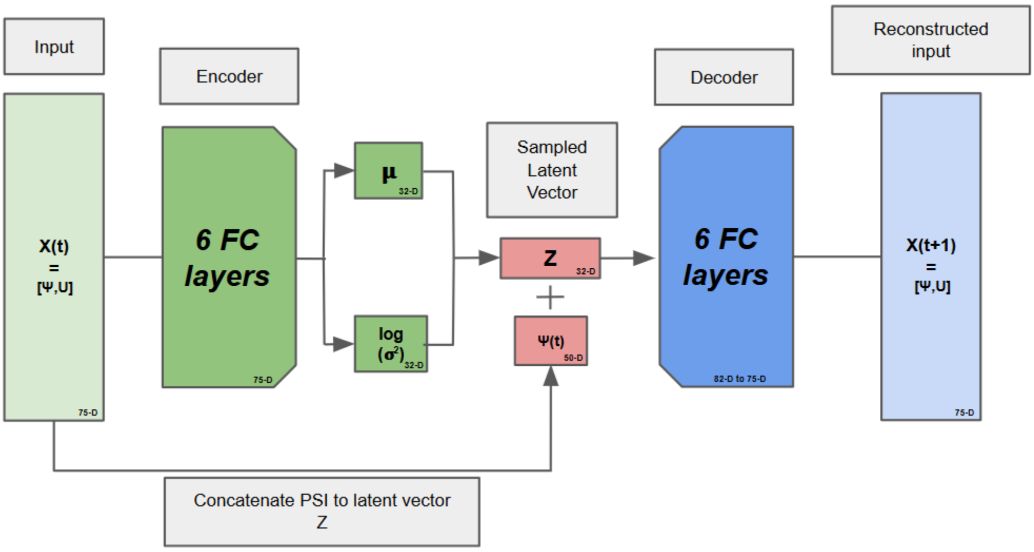


Figure 2. The encoder (6 fully connected layers) maps x_t to 32-dimensional latent mean and log-variance vectors. A latent variable z is sampled and concatenated with Ψ_t to form a 82-D decoder input. The decoder (6 fully connected layers) outputs the predicted next state x_{t+1}. This simple probabilistic design implements and captures stochastic regime transitions in the Holton–Mass model.

Both encoder and decoder are 6-layer fully connected (FC) stacks:

- **Encoder:** $\mathbb{R}^{75} \to \mathbb{R}^{32}$ mean vector $\mu(x_t)$ and $\mathbb{R}^{32}$ log-variance vector $\log \sigma^2(x_t)$.
- **Latent Sampling:** $z = \mu + \sigma \odot \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I_{32})$.
- **Decoder input:** $[z; \Psi_t] \in \mathbb{R}^{32+50} = \mathbb{R}^{82}$.
- **Decoder:** $\mathbb{R}^{82} \to \mathbb{R}^{75}$ predictive mean for x_{t+1}.

Hidden-layer widths and activations follow standard MLP practice (ReLU activations) and are kept constant across encoder/decoder for simplicity. The likelihood is taken as a factorized Gaussian with fixed variance, and we use a robust reconstruction loss that is equivalent to a Huber penalty on the decoder mean (see 3.3.0.1 Loss Function).

3.2.0.5 Objective (ELBO with robust reconstruction). We optimize the evidence lower bound (ELBO) with a KL-annealed weighting:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(z \mid x_t)}[\log p_\theta(x_{t+1} \mid \Psi_t, z)] - \beta \, D_{\mathrm{KL}}(q_\phi(z \mid x_t) \parallel p(z)) \,, \tag{6}$$

with $p(z) = \mathcal{N}(0, I_{32})$ and a cycling annealing schedule $\beta = [0.01, 0.3]$ to avoid posterior collapse. For $\log p_\theta(x_{t+1}|\Psi_t, z)$ we use a robust Huber reconstruction,

$$\mathcal{L}_{\mathrm{rec}}(x_{t+1}, \hat{x}_{t+1}) = \sum_{j=1}^{75} \mathrm{Huber}_\delta\big((x_{t+1})_j - (\hat{x}_{t+1})_j\big), \tag{7}$$

where $\hat{x}_{t+1}$ is the decoder output and δ is the Huber threshold. The KL term uses the closed form for Gaussians:

$$D_{\mathrm{KL}} = \tfrac{1}{2}\sum_{k=1}^{32}\left(\mu_k^2 + \sigma_k^2 - \log\sigma_k^2 - 1\right). \tag{8}$$

3.3 Training

3.3.0.1 Loss Function The loss function of a CVAE is made up of two components: the reconstruction loss and the KL divergence loss. The reconstruction loss is the Smooth L1 Loss function, which replaces the MSE for large prediction errors,

$$L(x) = \begin{cases} \frac{1}{2}(x-y)^2, \mid x-y \mid < 1 \\ \mid x-y \mid -\frac{1}{2}, \mid x-y \mid \geq 1 \end{cases}$$

Recall that the key objective of this model is to emulate the Polar Vortex, including its extreme events such as SSWs. Compared to the Smooth L1 Loss, MSE also squares large residuals, heavily penalizing high errors. If trained too much, the model may overfit to the outliers, in this case the transitions, degrading performance. In turn, the Smooth L1 Loss also improves stability when transitions are present. The KL divergence loss transforms the true posterior distribution, $p_\theta(z|x)$ with parameters θ, to an approximate standard normal distribution form, $q_\phi(z|x)$ with learned parameters ϕ. This approximation allows the latent space to sample noise directly and easily from the same standard normal distribution, $\mathcal{N}(0, I_{32})$, allowing for decoder-only inference that halves the inference time with a negligible difference in performance. To mitigate potential posterior collapse during training, we implement a cycling linear annealing schedule (Fu et al., 2019) with coefficient β ranging from 0.01 to 0.3 and cycling every 100 epochs.

3.3.0.2 Training parameters The model was trained using all 250,000 days from our dataset and in batches of size 1024 with a validation training set of 50,000 days. This is based on the physical intuition that the average return period of the vortex breakdown is slightly lower, at around 700-800 days as per the HM physical model. Through experimentation, the best learning rate was found to be around 10^{-4}, where if we increased the learning rate, the model would become increasingly unstable, and if we reduced the learning rate, the model would take too long to train. The training process is done over 1500 epochs, of which we choose the model with the best long-term statistics through multiple sessions.

3.3.0.3 Choosing the Best Model Minimizing a loss function on short-term statistics, like the Smooth L1 loss, does not guarantee fidelity of long-term statistics. To obtain one with improved long-term statistics, we iterated through multiple training sessions with the same parameters, and chose the best model by the euclidean distance of three long-term metrics of an inference at each saved epoch: exponential fit error of the return periods, the range error of the return periods, and the KL divergence error between the probability distributions(i.e. how much information was lost as the emulator tried to approximate the physical probability distribution). Note that each training session was different due to inherent stochasticity.

3.3.0.4 Finetuning During training, we found that the emulator was slightly overweighting state A, in terms of the marginal PDF of $U(30\text{ km})$. Initially, our proposed solution went as follows: run an emulator inference of $3{\cdot}10^5$ days with the best model, find the number of days the system was in state A and call it d, and then sample $\frac{d}{2}$ state A days to remove from the dataset. Subsequently, we finetuned the best model using the recently modified dataset with a learning rate of 10^{-8}. Visually, the difference between the two PDFs remained the same, but the method was numerically verified to improve the model by the KL divergence error. The proposed solution lowered the error from 0.119 to 0.097, which was averaged over 1000 inferences. The difference implies that the emulator was able to capture $\sim 18.5\%$ more information in its approximation of the physical probability distribution than previously done.

4 Results

4.1 Evaluating Emulator Skill Overview

To comprehensively evaluate the AI emulator, we assess its predictive accuracy and ability to reproduce the dynamical behavior of the Holton–Mass model. The evaluation is structured as:

- **Short-term accuracy:** Emulator's single time step predictive ability and short forecast skill using root mean square error (RMSE) of the zonal wind.
- **Long-term dynamics:** Climatological comparison of statistics of the emulator against the physical model, including:
  - Steady-state probability density functions with respect to zonal wind and IHF,
  - Return period distributions quantifying regime persistence times,
  - Transition duration distributions characterizing timescales.
- **SSW risk quantifiers:** The emulator's ability to capture key predictability metrics, is evaluated using the committor function's q^+ (probability of transitioning to state B before A) and the conditional mean first passage time η^+ (lead time to SSW onset).
- **Latent space structure:** PCA is used to examine the internal representation learned by the CVAE.

All long-term statistics are computed using up to 10^6 days of data, as the emulator became unstable beyond 1.4 million days.

4.1.0.1 Short-term accuracy Fig. 3 shows the results of testing the model's ability to predict the next time step from a given initial state. Fig. 4 shows test predictions from state B, while Fig. 3 shows test predictions from state B. Both figures show that the emulator captures the correct trend of the Holton Mass model throughout vertical levels, though the emulator tends to overestimate the change at higher states, owing to the greater wind speeds at higher levels. Conversely, the emulator's predictions are most accurate at the lowest levels, implying that higher levels or higher value wind speeds are more difficult to emulate. The rightmost panel shows the emulator's mean deviation predictions after 10,000 inferences. This panel shows that the mean of the predictions trace the Holton mass faithfully, and that variability once again increases as vertical height increases.

The forecast error is computed as:

4.1.0.2 RMSE Calculation

$$\mathrm{RMSE}(t) = \sqrt{\frac{1}{N_s}\sum_{s=1}^{N_s}\left(\bar{y}_{t,s} - y_{t,s}^{\mathrm{true}}\right)^2} \tag{9}$$

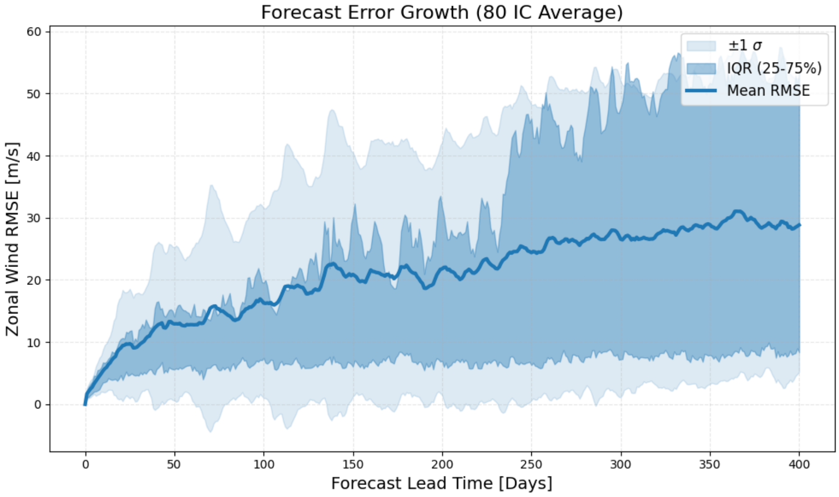
$$\bar{y}_{t,s} = \frac{1}{N_e}\sum_{e=1}^{N_e} y_{t,s,e} \tag{10}$$

$$\overline{\mathrm{RMSE}}(t) = \frac{1}{N_{\mathrm{IC}}}\sum_{i=1}^{N_{\mathrm{IC}}} \mathrm{RMSE}_i(t) \tag{11}$$

where:

- t: forecast lead time [days], $t \in \{0, 1, \ldots, 400\}$
- $N_s = 25$: number of spatial grid points (zonal wind levels)
- $N_e = 50$: number of ensemble members
- $N_{\text{IC}} = 80$: number of initial conditions
- $y_{t,s,e}$: prediction at time t, point s, ensemble member e $(401 \times 25 \times 50)$
- $\bar{y}_{t,s}$: ensemble mean prediction at time t, point s (401×25)
- $y_{t,s}^{\text{true}}$: ground truth at time t, point s (401×25)
- $\text{RMSE}_i(t)$: RMSE for initial condition i at time t $(401,)$
- $\overline{\text{RMSE}}(t)$: mean RMSE across all initial conditions $(401,)$

The aggregated error statistics are visualized over time t. The solid curve represents the mean RMSE across all initial conditions . To quantify the spread of the error, the uncertainty bounds in the plot represent the interquartile range (IQR) and one standard deviation ($\pm 1\sigma$) of the RMSE values across the $N_{IC} = 80$ initial conditions.



1

Figure 3. Forecast error growth (RMSE) averaged over 80 initial conditions from random initial conditions. The dark blue line shows the mean RMSE, with the medium blue region indicating the interquartile range (IQR, 25-75%) and the light blue region showing $\pm 1\sigma$ uncertainty bounds.

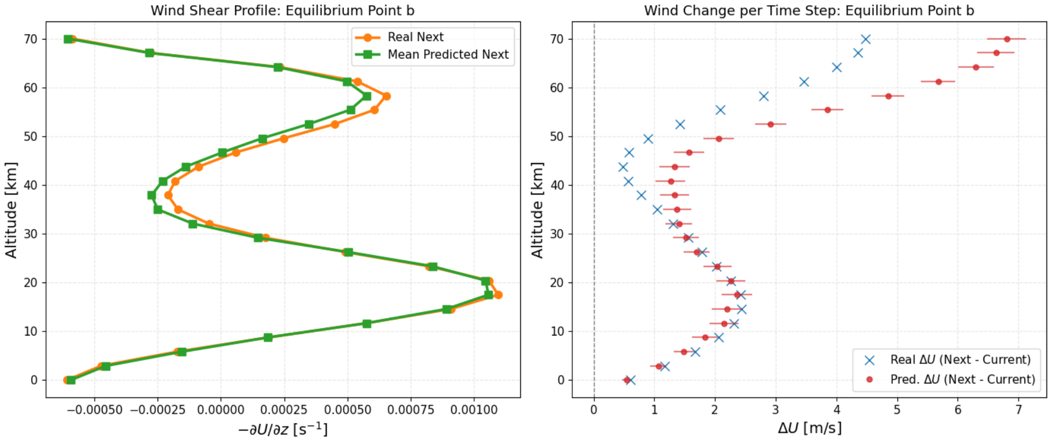


Figure 4. (a) Vertical derivative of zonal wind (shear). (b) Ensemble-mean time derivative of a 1000-member ensemble prediction, according to the emulator (red) and the Holton-Mass model (blue). Error bars show 2σ intervals.

4.1.0.3 Long-term Dynamics The long-term dynamics, or "climatology", of the system is also of primary interest. Fig. 5 shows excellent agreement between the emulator's steady-state density and that of the HM model, in particular the bimodal structure with respect to the zonal wind(U) and Integrated Heat Flux(IHF), which is defined as

$$IHF(z\ km) = \int_{0\ km}^{z\ km} e^{-z/H}\overline{v'T'}dz \propto \int_{0\ km}^{z\ km} |\Psi|^2 \frac{\partial \varphi}{\partial z}dz.$$

The emulator also captures to a great extent the elliptical level sets of the Holton-Mass data, as seen in Fig. 5, regardless of the corresponding state.

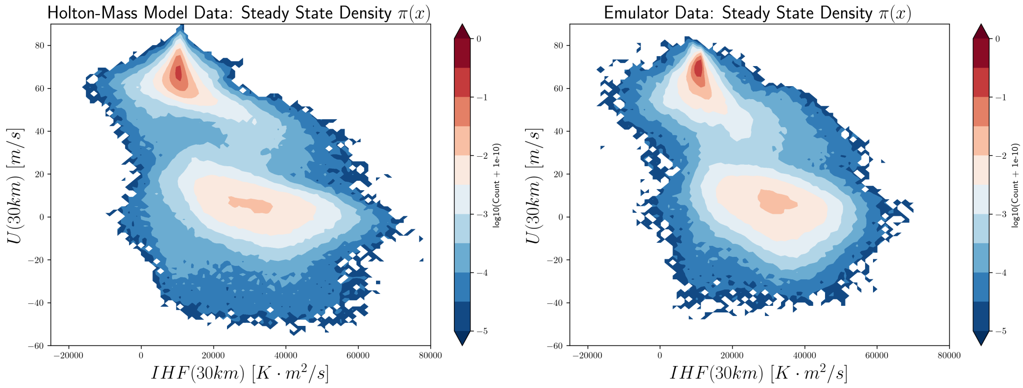


Figure 5. Two-dimensional projections of the PDF of the Holton-mass (left) and the AI emulator (right) with respect to U(30 km) and IHF(30 km).

Return periods are essential for comparing the dynamical behavior of climate models. If two models produce similar distributions of regime persistence times, they are in strong agreement regarding the likelihood and temporal characteristics of extreme transitions. To quantify this, we estimate return periods by scanning $U(30\,\text{km})$ for crossings between the B-state threshold u_B and the A-state threshold u_A. Each time the trajectory exits one state and subsequently entered the opposite one, we recorded the elapsed number of days τ.

To analyze the distribution of these persistence times, we compute the empirical complementary cumulative distribution function (CCDF) of the τ values using fixed 500-day bins. This uniform binning avoids overweighting the large number of short-duration events and provides a balanced view of the full range of return periods. For each model, we apply 1000 bootstrap resamplings of the τ values and recompute the CCDF to obtain a median estimate and a 95% confidence interval at each bin.

The resulting persistence distributions from the emulator and the Holton–Mass model are in close agreement across most durations. Their median CCDFs track each other closely over several orders of magnitude. Differences appear primarily in the tail, where the emulator exhibits slightly higher probabilities of longer persistence times, consistent with its tendency to generate marginally longer-lived states. Overall, the emulator captures the essential structure of the return-period distribution while reproducing the variability and uncertainty of the underlying stochastic dynamics.

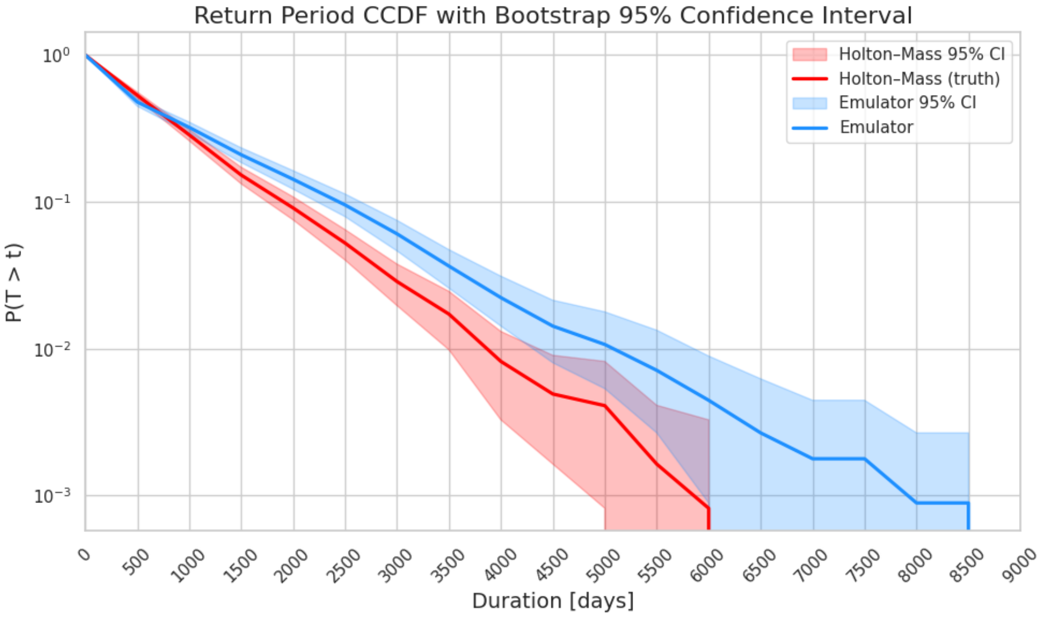


Figure 6. Complementary cumulative distribution functions (CCDFs) of regime persistence (durations between transitions) for the Holton–Mass model (red) and the emulator (blue). Solid lines show the median empirical CCDFs, and the shaded regions indicate 95% bootstrap confidence intervals based on 1000 resamplings. The figure uses fixed 500-day bins and logarithmic axes to highlight differences in the tail behavior of the return-period distribution.

Transition durations present a core skill that the model should have, as they represent “suddenness” of SSW events and hence the typical time available to prepare for a forecasted event. Overall, the two models have virtually the same structure in the distribution, with the emulator being slightly skewed to the left and thus slightly faster transitions.

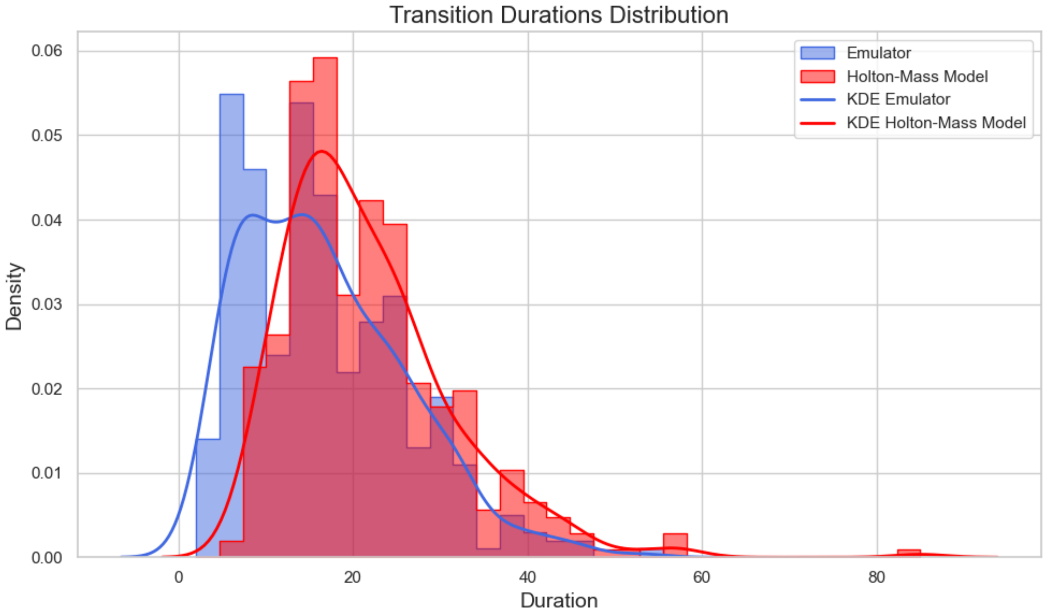


Figure 7. Distribution of transition durations (days) between vortex regimes. Overlaid histograms and KDE curves compare the emulator (blue) and the physical HM model (red).

4.2 Committor and lead time, the SSW risk quantifiers

In the following section, we plot two risk quantifiers for both the Holton-Mass model and the emulator. The committor, q^+ is the probability that given an initial condition, the system first reaches state B before state A. Suppose that it does. The expected time it takes to get there is called the conditional mean first passage time, or lead time: η^+.

4.2.0.1 Committor function Fig. 8 shows the SSW committor of the HM model projected onto U and IHF over 10^6 days. The data is sliced into 100 bins with respect U and IHF, and the committor function for each bin is computed as follows,

$$q^+_{ij} = \frac{\sum_x R_{ij}(x)}{\sum_x R_{ij}(x) + \sum_x U_{ij}(x)},$$

where $R(x)$ is the number of realized transitions that stem from an AB point x in the bin, $U(x)$ is the number of brief threshold excursions that stem from a non-AB point x in the bin, i is the ith bin of U(30 km), j is jth bin of IHF(30 km), and the sums are taken over all x in the corresponding ij bin. In the low IHF region, the model does not have as smooth of a probability outline, as shown by the staggered walk with respect to the zonal wind and the few artifacts with high probability. This is, however, overshadowed by dominant pattern that is the negative correlation between U and q^+ and the positive correlation between IHF and q^+. The striking similarity between committors is emphasized by the 50% probability level set that mimics the sharpness and movement pattern of its physical counterpart.

4.2.0.2 Lead time metrics of SSW Fig. 9 displays the lead time of some SSW. The data, as previously done in 4.2.0.1, is sliced into 100 bins with respect U and IHF, and the lead time function only takes points x from the C region that are already part of some AB transition. Then, the lead time function for each bin is computed as follows,

$$\eta^+_{ij} = \frac{\sum_x N_{ij}(x)}{M},$$

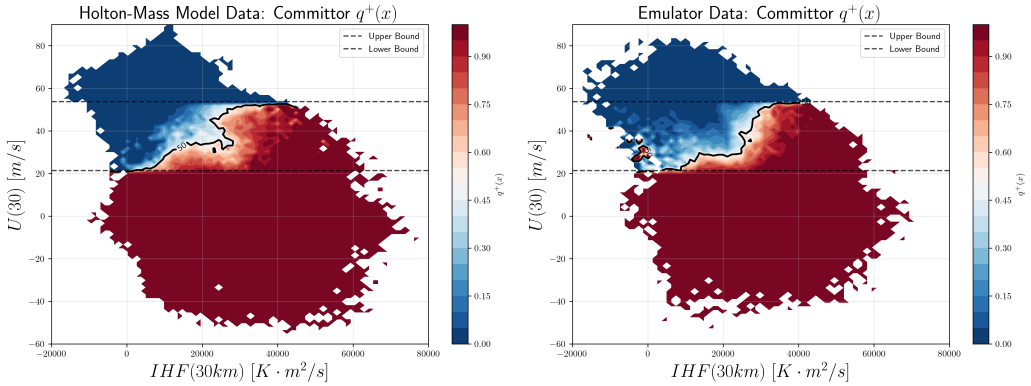


Figure 8. Committor functions according to the HM model (a) and the emulator (b), with respect to $U(30\text{ km})$ and IHF(30 km). The two dashed lines are the bounds of state A and state B, respectively, and the black line is the level set $q^+ = 0.5$

where $N(x)$ is the number of days until the system reaches state B from x, i is the ith bin of $U(30\text{ km})$, j is jth bin of IHF(30 km), and M is the number of transitions in the ij bin. Our model seems to struggle with low IHF values, as seen by the lower presence of high lead times when compared to the Holton–Mass model. Nevertheless, the model captures the level set structure reasonably well as $U(30\text{ km})$ decreases and IHF(30 km) increases.

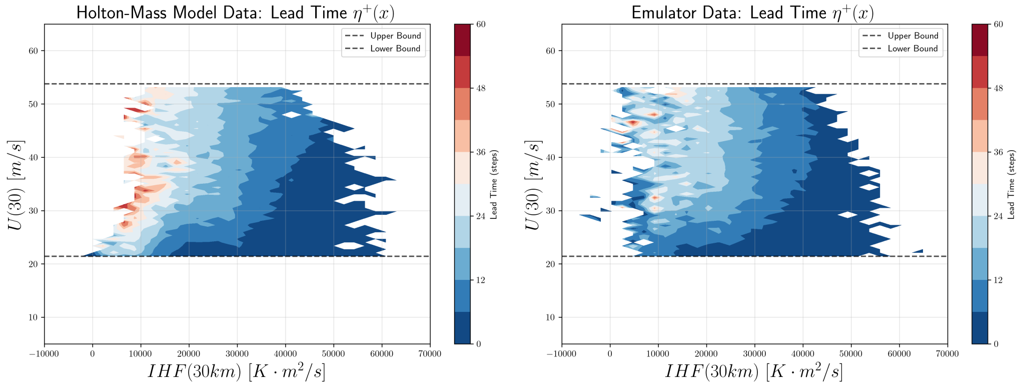


Figure 9. Lead time with respect to $U(30\text{ km})$ and IHF(30 km) of (a) the Holton-Mass Model, and (b) the emulator. The two dashed lines represent the bounds of state A and state B, respectively.

4.3 Visualizing and Interpreting the Latent Space

We examine the latent representation to make sense of what the model has internalized. Because the CVAE encodes each input state x_t into a 32-dimensional latent vector z, we apply Principal Component Analysis (PCA) to the latent mean vectors to identify dominant directions of variability. PCA provides a linear, interpretable reduction that can reveal geometric organization within the latent manifold.

4.3.0.1 Motivation. If the CVAE successfully captures the metastable regime structure of the HM dynamics, its latent space could?Should? (which should I use?) reflect the same physical organization: two stable basins (A and B) and two transition pathways ($A \to B$ and $B \to A$). Demonstrating such separation provides evidence that the model's internal representation is dynamically meaningful and not merely a statistical compression.

4.3.0.2 Principal Component Analysis. Since the four classes of days occur with unequal frequency, we balance them by sampling an equal number of points from each equal to the smallest class size. For the Holton-Mass model this minimum is 756 points ($A \rightarrow B$), and for the Emulator it is 905 points ($B \rightarrow A$). We therefore draw 756 (HM) or 905 (Emulator) points from each of the A, B, AB, and BA sets. This guarantees balance between the four regimes and prevents PCA from being biased toward more common states.

For each sampled day, we compute the latent mean vector $\mu(x_t)$ from the encoder and perform PCA on the resulting 32-dimensional dataset. The first three principal components explain approximately 99.87% of the total variance in the Physical Model (PC1 $\approx$ 99.61%, PC2 $\approx$ 0.17%, PC3 $\approx$ 0.09%) and 99.88% in the Emulator (PC1 $\approx$ 99.67%, PC2 $\approx$ 0.13%, PC3 $\approx$ 0.08%). Two rotated 3D views of this projection are shown in Fig. 10. The latent geometry is striking: the four dynamical regimes form distinct clusters in the (PC1, PC2, PC3) subspace.

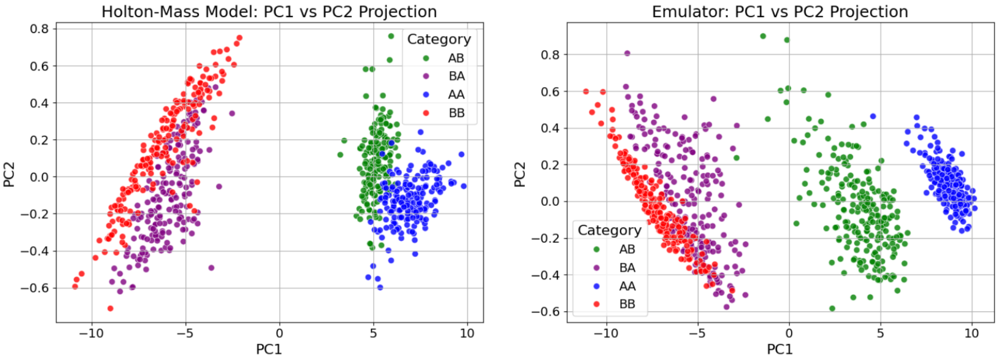


Figure 10. Projection of the latent mean vectors onto the first three principal components for the Physical Model (left) and Emulator (right). Points are colored by their physical regime label.

4.3.0.3 Four distinct regimes. The clusters correspond to:

1. ***A* (Strong vortex, stable)** — concentrated at large positive PC1
2. ***B* (Weak vortex, stable)** — concentrated at large negative PC1
3. ***AB* (Pre-SSW transition)** — located between A and B but displaced in PC2, reflecting a weakening vortex
4. ***BA* (Recovery transition)** — the reverse of AB, reflecting a strengthening vortex

Together these four groups form an elongated, quasi-two-dimensional manifold along PC1 with secondary curvature in PC2–PC3.

4.3.0.4 Physical interpretation. This emergent organization indicates that the ResNet-CVAE captures not only the two stable states of the HM model but also the dynamical pathways connecting them. The fact that such structure arises without providing regime labels during training suggests that by learning the conditional distribution $p(x_{t+1} \,|\, x_t)$, the network implicitly discovers physically meaningful manifolds of variability. This supports the broader motivation of this work: well-designed generative emulators can replicate stochastic dynamics while also revealing internal representations that mirror the underlying physical regime structure.

5 Conclusions

The emulator shows a strong capacity to capture the system's metastable states. However, an analysis of the plots demonstrates some of the limitations in error distribution and vertical modeling.

The vertical structure of the zonal wind, in Fig. 4, focuses only on equilibrium point **b**. Though the behavior is similar for other states (not shown), the assessment is still qualitative. This suggests that the structural behavior at equilibrium point **b** is not just an anomaly but representative of the entire phase space. The vertical change of the zonal wind difference, as seen in Fig. 4, shows that the model makes predictions in the middle atmosphere the best. This phenomenon suggests that the mid level accuracy could mask larger deviations near the top and bottom of the model, a factor that must be considered when interpreting aggregate skill scores. In regions with low IHF values, the emulator faces slight difficulty in fully capturing the dynamics, as seen in the steady state density, committor, and lead time. The existence of these limitations does not take away from the emulator's high degree of fidelity. Rather, it merely displays points of future work, especially since the paper put forward only one of the many viable hyperparameter regimes.

Furthermore, the PCA analysis of the latent space shows that our CVAE emulator does more than match short-term forecasts and long-term statistics: it learns a low-dimensional, dynamically meaningful representation of the Holton–Mass system. By projecting the 32-dimensional latent means onto the leading principal components, we find four well-separated clusters associated with strong and weak vortex states and their respective pre-transition pathways, effectively revealing an emergent reaction coordinate for SSW-like regime changes. Our results demonstrate that VAEs can encode slow dynamical variables and transitions pathways in their latent space in a stochastic, metastable climate toy model where rare transitions are central to the problem. For researchers interested in rare-event dynamics, this provides a concrete, data-driven handle on the geometry of regime transitions, and for those working on interpretable ML, it offers an example where a relatively simple probabilistic emulator yields a latent manifold that maps cleanly onto physically defined regimes rather than remaining a black box.

In a broader context, this work fits alongside emerging efforts to use AI both to accelerate climate simulations and to better understand extremes. Recent reviews emphasize that AI-based methods for climate extremes will only be trusted if their speed is combined with physical fidelity and interpretability (Materia et al., 2024). Large-scale climate emulators such as ACE2, which reproduce subseasonal-to-decadal variability and phenomena like SSWs at global scale (Watt-Meyer et al., 2025), show that autoregressive ML models can stably emulate complex atmospheric dynamics, though their internal representations are often opaque. At the same time, generative AI approaches such as FM-Cast for SSW ensemble prediction demonstrate how probabilistic deep learning can rival or exceed operational systems for real-world events while remaining computationally efficient (Tao et al., 2025). Our emulator contributes at the "idealized building-block" level that these communities need: it demonstrates, in a controlled setting provided by the Holton–Mass model, that a carefully designed stochastic deep model can faithfully capture transition statistics and SSW risk quantifiers while exposing a latent space in which regimes and precursors are clearly organized. Researchers working on operational SSW prediction and AI weather models may be most interested in how this latent structure could inform feature design, regime-aware training, or rare-event sampling strategies in more realistic systems, while ML theorists and climate dynamicists may see it as a testbed for future work on disentangled or physics-informed latent variables that enable targeted manipulation of SSW likelihood or lead time.

In summary, this study demonstrates that ResNet-Inspired Conditional VAEs are able to closely emulate the stochastic Holton-Mass model. The emulator reproduces with close agreement both the steady-state density and risk quantifiers such as return periods, committor, and lead time. At the same time, the latent space PCA contributes by displaying four well-separated clusters with clear physical meanings, discovering a model that is not only a black box. Nevertheless, the analysis raises opportunities for further research: improving the vertical error structure and dealing with the difficulties of low-IHF regions. These findings indicate that ResNet-Inspired Conditional VAEs can accelerate climate simulations and internalize physically meaningful structure, which allows for promising future research in modeling stratospheric extremes.

Appendix A Alternative Strategies Explored

We conducted several training experiments using different setups, including a number of approaches that failed to deliver the desired results: forecasting day-to-day differences, predicting with inputs spanning multiple days, reducing the latent dimensionality, and augmenting the latent space with both U and Ψ.

We intuited that the model may learn the daily difference in U and Ψ. We have tried including this objective in two setups. One setup had the label changed to the day-to-day differences, and the other setup had the reconstruction loss changed such that it compares the Holton-Mass and emulator differences. Neither proved to be true. To improve the machine learning model's capacity to emulate, we removed the assumption of a Markovian process and conducted an experiment by using multiple days,

$$x(t), x(t-1), \dots$$

to predict $x(t{+}1)$. This technique unexpectedly resulted in an emulator that constantly jumped between states, so we stuck with $x(t)$ only.

Reducing the latent dimensionality, even as low as one, increased instability. Nevertheless, some epochs had better inferences that even our best model up to 100k days, after which they would go unstable. Since we prioritized stability, we decided to stay at a latent dimension of 32. As mentioned in section 3.3.0.2, conditioning on both Ψ and U instead of only Ψ crippled the emulator, scoring the highest possible error in all long-term metrics. We intuit that the problem may rise from the PDF bimodality as we go up into the stratosphere. Early on into the stratosphere, the zonal wind has a unimodal distribution, but as we increase the altitude, the stochastic forcing weakens, quickly transitioning into a bimodal distribution within a few kilometers.

Acknowledgments

This work was funded in part by the United States National Science Foundation through award NSF RISE–2425898. We acknowledge the support of undergraduate summer Polsky Research Fellowships from the Institute for Climate and Sustainable Growth at the University of Chicago. We acknowledge support from a University of Chicago Quad Undergraduate Research Scholarship. This research was funded in part by the Institute for Climate and Sustainable Growth at the University of Chicago.

References

Baldwin, M. P., & Dunkerton, T. J. (2001). Stratospheric harbingers of anomalous weather regimes. *Science*, *294*, 581–584. doi: 10.1126/science.1063315

Birner, T., & Williams, P. D. (2008). Sudden stratospheric warmings as noise-induced transitions. *Journal of the Atmospheric Sciences*, *65*(10), 3337 - 3343. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/65/10/2008jas2770.1.xml` doi: 10.1175/2008JAS2770.1

Charlton, A. J., & Polvani, L. M. (2007). A new look at stratospheric sudden warmings. part i: Climatology and modeling benchmarks. *J. Climate*, *20*, 449–469. doi: 10.1175/JCLI3996.1

Chattopadhyay, A., Pathak, J., Nabizadeh, E., Bhimji, W., & Hassanzadeh, P. (2023). Long-term stability and generalization of observationally-constrained stochastic data-driven models for geophysical turbulence. *Environmental Data Science*, *2*. doi: 10.1017/eds.2022.30

Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., & Weare, J. (2021, Nov). Learning forecasts of rare stratospheric transitions from short simulations. *Monthly Weather Review*, *149*(11), 3647–3669. doi: 10.1175/mwr-d-21-0024.1

Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., & Weare, J. (2022). Data-driven transition path analysis yields a statistical understanding of sudden stratospheric warming events in an idealized model. *Journal of the Atmospheric Sciences*. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/aop/JAS-D-21-0213.1/JAS-D-21-0213.1.xml` doi: 10.1175/JAS-D-21-0213.1

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., & Carin, L. (2019). *Cyclical annealing schedule: A simple approach to mitigating kl vanishing.* Retrieved from `https://arxiv.org/abs/1903.10145`

Holton, J. R., & Mass, C. (1976). Stratospheric vacillation cycles. *Journal of Atmospheric Sciences*, *33*(11), 2218 - 2225. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/33/11/1520-0469_1976_033_2218_svc_2_0_co_2.xml` doi: 10.1175/1520-0469(1976)033⟨2218:SVC⟩2.0.CO;2

Kautz*, L.-A., Polichtchouk*, I., Birner, T., Garny, H., & Pinto, J. G. (2020). Enhanced extended-range predictability of the 2018 late-winter eurasian cold spell due to the stratosphere. *Quarterly Journal of the Royal Meteorological Society*, *146*(727), 1040-1055. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3724` doi: https://doi.org/10.1002/qj.3724

Materia, S., Weerts, A., Lang, D., & Balsamo, G. (2024). Artificial intelligence for climate extremes: A review. *Nature Reviews Earth & Environment*. doi: 10.1038/s43017-024-00565-y

Tao, J., O'Neill, P., Fletcher, C., & et al. (2025). Fm-cast: A generative ai framework for stratospheric sudden warming ensemble forecasting. *Geophysical Research Letters*. (In press)

Wang, Y., Blei, D. M., & Cunningham, J. P. (2023). *Posterior collapse and latent variable non-identifiability.* Retrieved from `https://arxiv.org/abs/2301.00537`

Watt-Meyer, O., Duncan, D., Brenowitz, N., & et al. (2025). Ace2: A stable deep learning atmospheric emulator for subseasonal-to-decadal climate variability. *Science Advances*. (In press)

Yoden, S. (1987). Dynamical aspects of stratospheric vacillations in a highly truncated model. *Journal of Atmospheric Sciences*, *44*(24), 3683 - 3695. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/44/24/1520-0469_1987_044_3683_daosvi_2_0_co_2.xml` doi: 10.1175/1520-0469(1987)044⟨3683:DAOSVI⟩2.0.CO;2