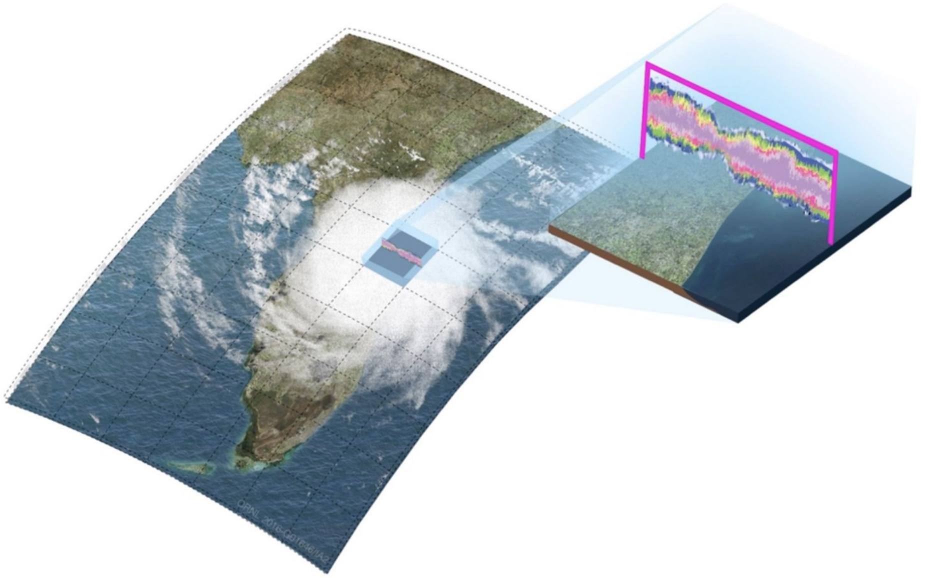
# Precipitation Downscaling with Spatiotemporal Video Diffusion

Prakhar Srivastava[1], Ruihan Yang[1], Gavin Kerrigan[1], Gideon Dresdner[2], Jeremy McGibbon[2], Christopher Bretherton[2] and Stephan Mandt[1]

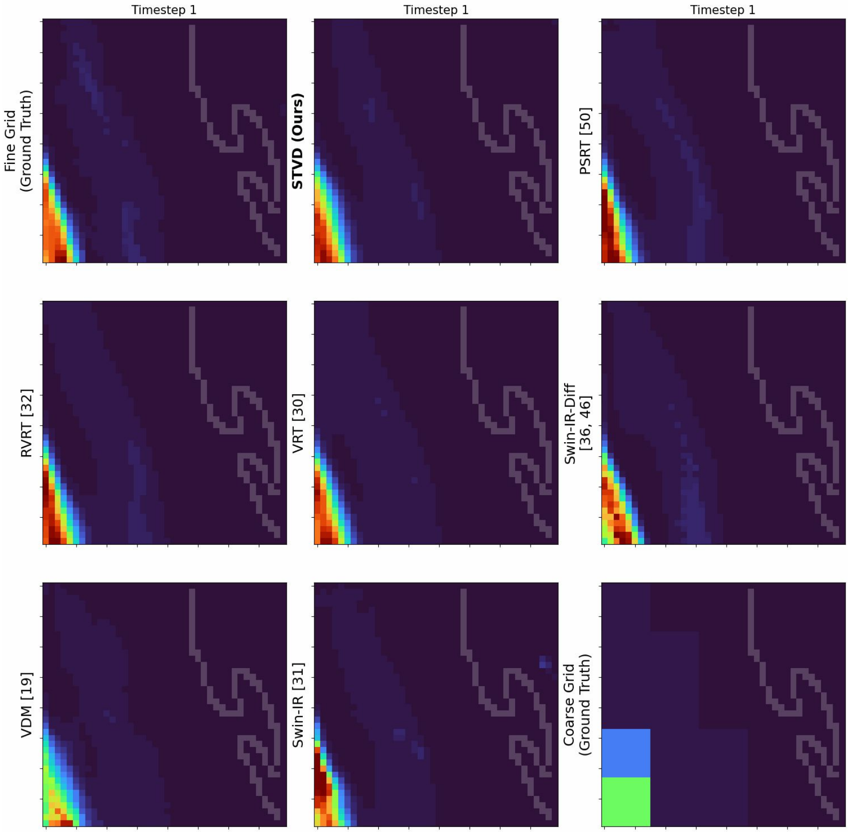
[1]University of California, Irvine [2]Allen Institute for Artificial Intelligence

# Background



- Precipitation patterns are central to human and natural life
- Simulations are challenging due to the multi-scale variability of weather systems and the influence of complex surface features
- Fluid-dynamical models of the global atmosphere are too expensive to run routinely at such fine scales

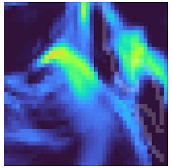
# Goal



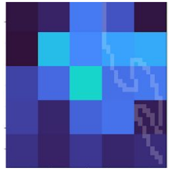

- Traditional downscaling methods are either dynamical or statistical
- Our work builds on vision based super-resolution methods to improve statistical downscaling
- Our objective is to transform a sequence ("video") of low-resolution precipitation frames into a sequence of high-resolution frames

# Deterministic methods and mode averaging

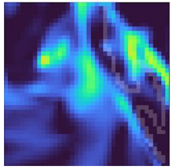
Fine grid Ground Truth



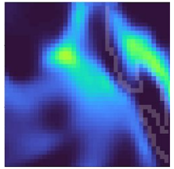
Coarse grid Ground Truth



STVD (Ours)

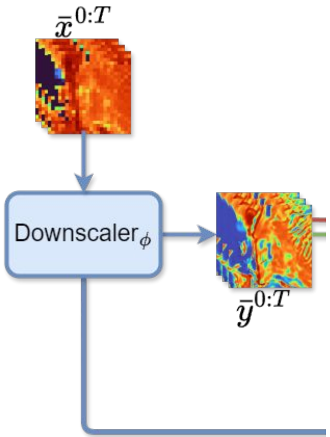


RVRT



- Deterministic methods for these problems often lead to visual artifacts from mode averaging
- A natural alternative to prevent mode averaging is generative modeling
- To that end, we propose **S**patio**T**emporal **V**ideo **D**iffusion (**STVD**)

# Approach

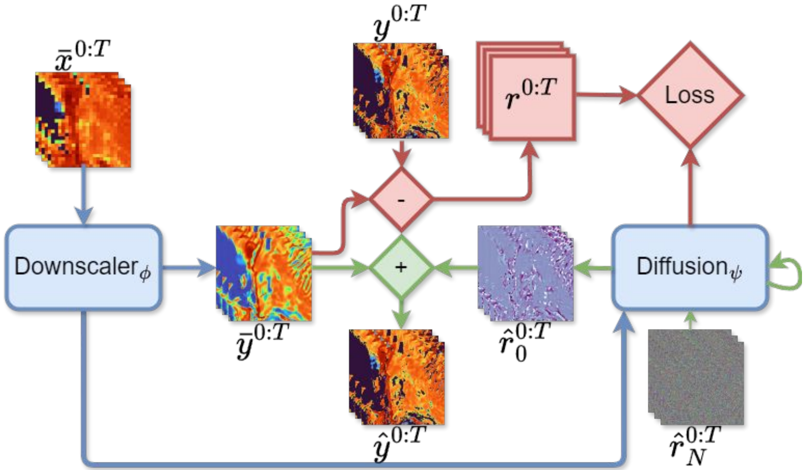



- Our model combines a **mean super-resolved sequence** from deterministic downscaling module with a **stochastic residual sequence** from a conditional diffusion model

- Our model outperforms six strong super-resolution baselines across multiple criteria

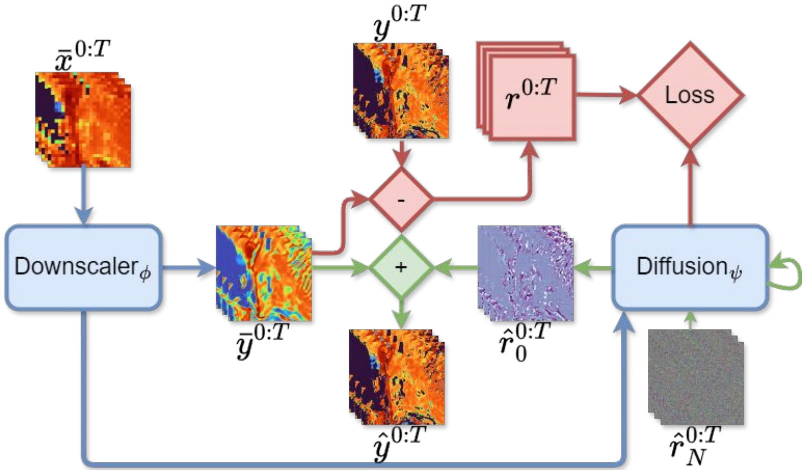
- Our approach captures key characteristics of precipitation and trained end-to-end, inspired from predictive-coding

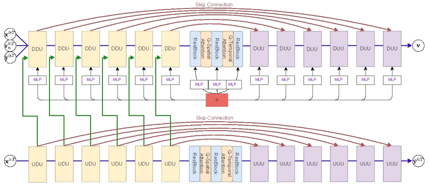
# Approach




- Our model combines a **mean super-resolved sequence** from deterministic downscaling module with a **stochastic residual sequence** from a conditional diffusion model
- Our model outperforms six strong super-resolution baselines across multiple criteria
- Our approach captures key characteristics of precipitation and trained end-to-end, inspired from predictive-coding

# Method

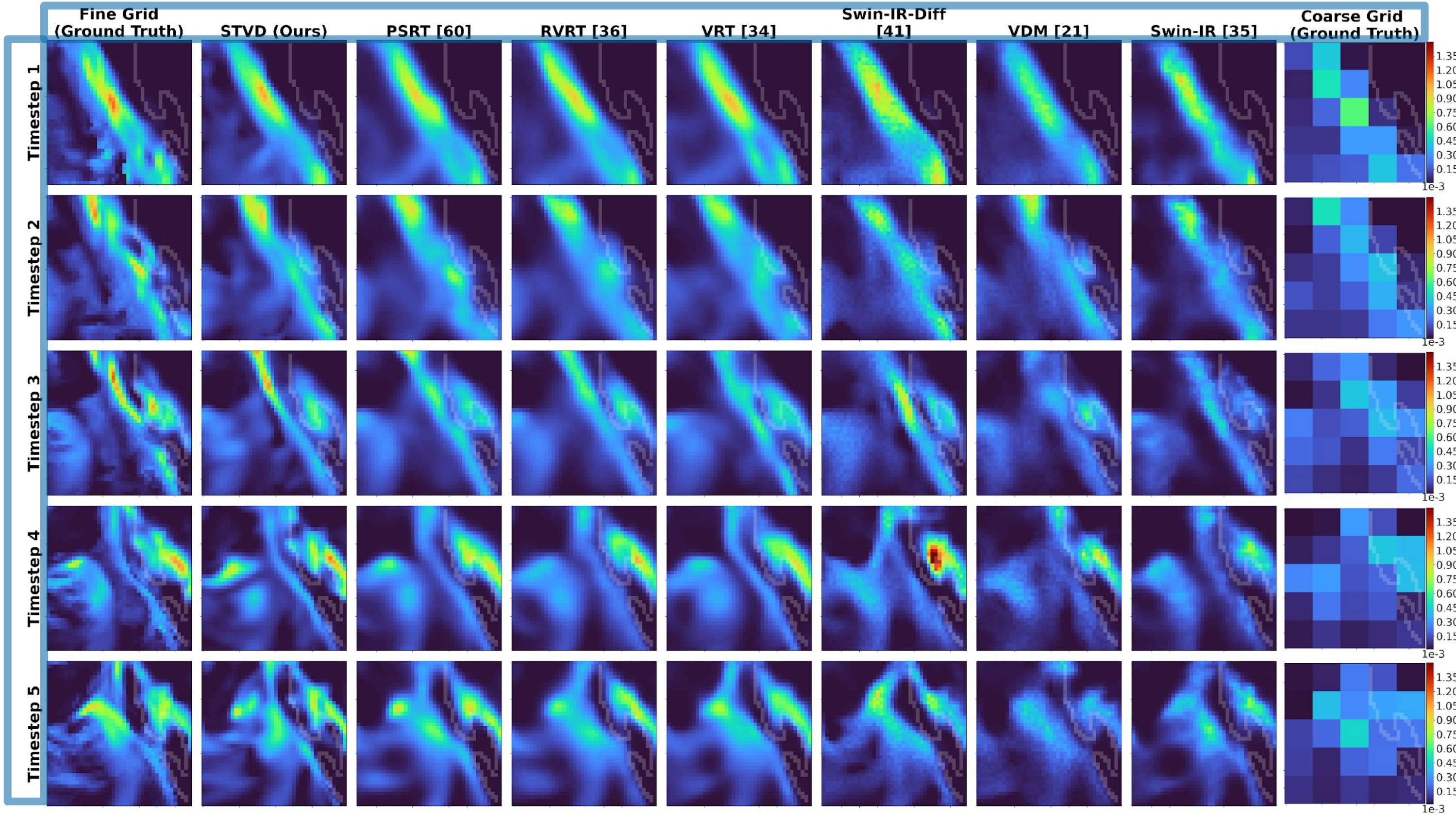


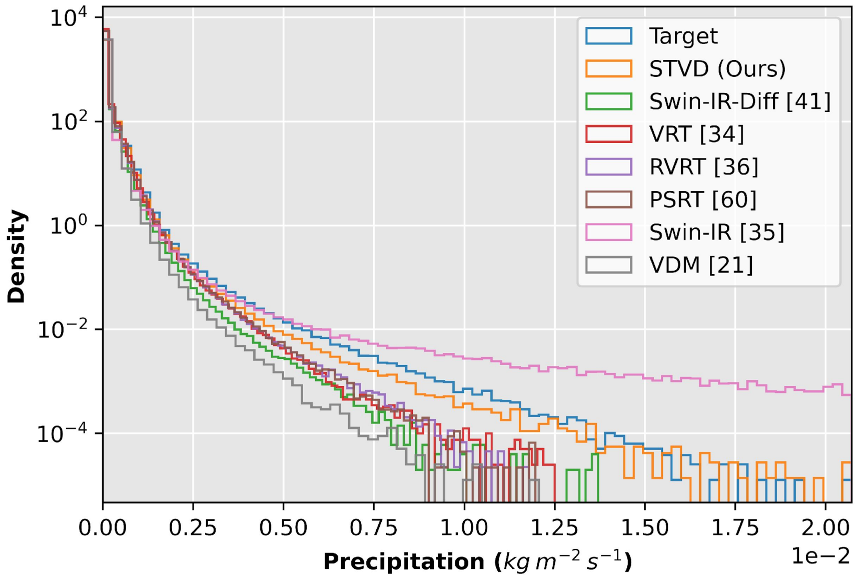

Implemented as UNets with spatio-temporal factorized attention and feature sharing

- $x^{0:T}$ is the coarse-grid sequence and $y^{0:T}$ is the corresponding fine-grid sequence
- Deterministic downscaler yields $\bar{y}^{0:T} = \mu_\phi(x^{0:T})$
- The residuals $r^{0:T} = y^{0:T} - \bar{y}^{0:T}$ modeled through DDPM with the reverse process $p_\psi(r^{0:T}_{n-1}, |r^{0:T}_n, c) = \mathcal{N}\left(r^{0:T}_{n-1} | \mathcal{M}_\psi(\mathrm{r}^{0:\mathrm{T}}_{\mathrm{n}}, n, c), \gamma \mathrm{I}\right)$
- $L(\phi, \psi) = \mathbb{E}_{x^{0:T}, y^{0:T}, n, \epsilon} \left\| v - \mathcal{M}_\psi(r^{0:T}_n, n, c) \right\|^2$

# Experiments

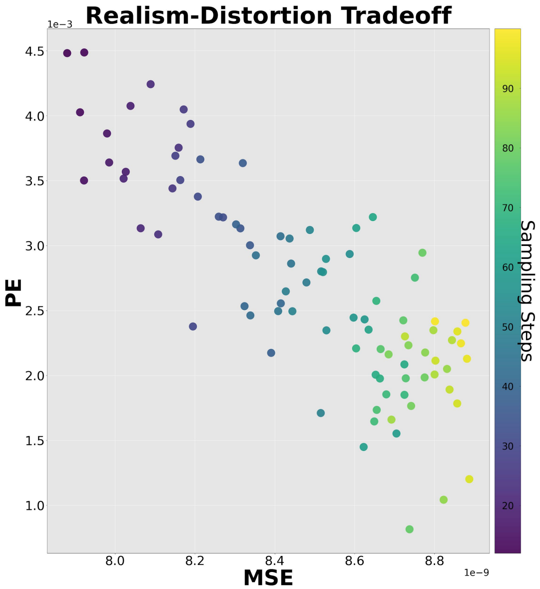


# Results




| | CRPS ($10^{-5}$) | MSE ($10^{-8}$) | EMD ($10^{-6}$) | PE ($10^{-3}$) | SAE ($10^{-6}$) |
|---|---|---|---|---|---|
| **STVD (ours)** | **1.85** | **0.59** | **2.49** | **1.2** | **4.00** |
| PSRT [60] | 2.15 | 0.66 | 4.21 | 3.8 | 6.24 |
| RVRT [36] | 3.55 | 1.73 | 4.33 | 3.6 | 7.39 |
| VRT [34] | 3.58 | 1.74 | 4.61 | 4.0 | 7.39 |
| Swin-IR-Diff [41] | 2.29 | 1.94 | 6.38 | 4.4 | 7.70 |
| VDM [21] | 2.21 | 0.73 | 12.70 | 6.4 | 8.84 |
| Swin-IR [35] | 2.36 | 2.29 | 17.40 | 23.40 | 18.9 |
| STVD-single | 1.81 | 0.62 | 4.64 | 2.3 | 6.09 |
| STVD-3 | 1.96 | 0.68 | 4.94 | 2.6 | 4.99 |
| STVD-1 | 2.05 | 0.72 | 7.19 | 4.1 | 6.87 |

# Realism-Distortion Tradeoff [1]

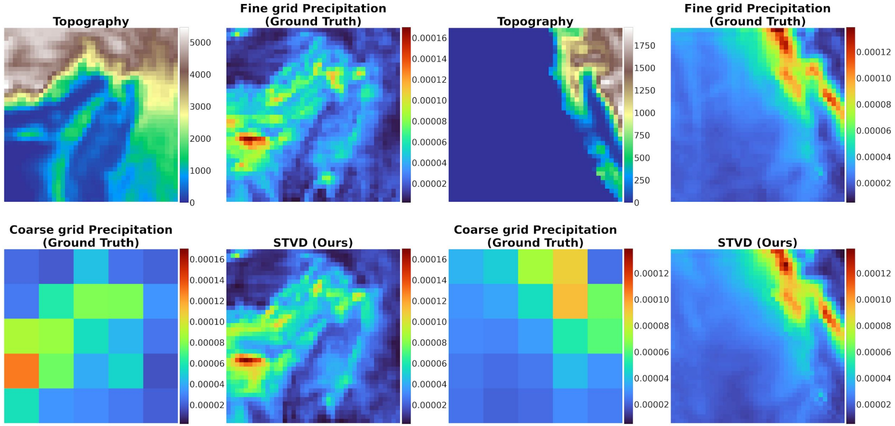



[1] Blau, Yochai, and Tomer Michaeli. "The perception-distortion tradeoff." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Perception Distortion tradeoff is commonly observed in the natural image domain. Perception/Realism measures the model's ability to unconditionally generate realistic samples irrespective of the context. While often measured by FID, we use the model's ability to match the ground truth precipitation histograms as a measure of realism.

- MSE measures accuracy, but conflicts with perceptual quality
- PE proxy for realism, captures extreme events crucial for flood forecasting and disaster mitigation
- More sampling steps increase MSE, but improve realism
- Depending on the application, practitioners can balance the tradeoff effectively

# Annual Trends




- Annual precipitation patterns are crucial for long-term water availability
- Fine-grid topography input enables accurate replication of high-precipitation regions
- Baseline models show faster spectral decay than STVD

# Stochastic Variability

STVD can produce multiple consistent samples that are each compatible with the coarse-grid reference simulation video

To analyze model stochasticity more effectively, we also include a variance map over these sample

In this map, red regions highlight areas of high variance, while blue regions indicate low variance