CloudFlow: A Flow Matching Model to Generate High-Resolution Cloud Structures

Tim Reichelt [1] **Philip Stier** [1]

Abstract

Our limited understanding of clouds is the dominant source of uncertainty in future climate predictions. Understanding how changing atmospheric environmental conditions constrain cloud organisational patterns and their radiative effects is key to understanding their impact on future climate change. We present CloudFlow, a flow matching model that is able to generate high-resolution cloud structures conditioned on coarse-scale atmospheric conditions. Our model generates realistic cloud structures that match the spectra and distributions of the original high-resolution scenes. CloudFlow introduces a new modeling regime to study how atmospheric environmental conditions impact cloud morphologies which will contribute to an improved understanding of cloud feedbacks, the cloud response to a changing climate and its effect on climate itself.

1. Introduction

A fundamental problem of climate science is predicting the expected amount of warming with increased CO_2 concentrations in the atmosphere. A key quantity of interest in the climate community is the Equilibrium Climate Sensitivity (ECS), the expected amount of equilibrium warming for a doubling of CO_2 concentrations in the atmosphere. Our best estimates of ECS have uncertainties ranging from around 2 Kelvin to 5 Kelvin (Arias et al., 2021). This wide uncertainty range has significant impacts on policy planning: if the goal is to constrain warming to two degrees an ECS on the low end would allow us to emit around twice as much carbon as an ECS on the higher end (Bony et al., 2015).

The dominant source of uncertainty in ECS estimates is due to cloud feedbacks, the response of cloud patterns and their radiative properties due to increased surface temperatures (Arias et al., 2021). Key to constraining cloud feedbacks is our ability to understand the processes that govern cloud formation and organisation at the mesoscale, which are the scales ranging from approximately 5 kilometres to several hundred kilometres (Bony et al., 2015). At these scales clouds can organise into complex patterns which have a direct effect on the Earth's energy budget because their organisation can determine cloud amount, radiative properties, and vertical structure, which combined control the complex balance between reflecting incoming solar radiation (cooling effect), or trapping outgoing longwave radiation (warming effect) of the cloud field. The evolution of complex mesoscale cloud structures encodes the underlying physics.

Existing methods to constrain cloud feedbacks include the usage of numerical models and observational datasets. Classic numerical global climate models (GCMs) that are used in large model intercomparison studies like the Coupled Model Intercomparison Project (CMIP) (Eyring et al., 2016) model the atmosphere at coarse resolutions of around 25-100 kilometers which means they do not explicitly resolve mesoscale cloud features and associated cloud feedbacks. This contributes to the huge uncertainties in cloud feedback estimates from numerical models (Stevens & Bony, 2013; Schneider et al., 2017). Global storm resolving models that resolve the atmosphere at km-scale resolutions are starting to emerge but these models are computationally so expensive to run that they cannot yet be used to make century scale predictions (Stevens et al., 2019; Segura et al., 2025).

Another approach relies on cloud controlling factor (CCF) analysis. CCFs are physics-based hand designed features that are computed from full 3D atmospheric profiles and that have been shown to correlate well with cloud cover and cloud radiative effects (Stevens & Brenguier, 2009; Wood & Bretherton, 2006; Klein et al., 2018; Ceppi & Nowack, 2021; Myers et al., 2021). One application of CCFs is to inform numerical experiments with large eddy simulations (Zhang et al., 2012; Bretherton, 2015; Bretherton & Blossey, 2017). Additionally, CCFs can help estimate cloud radiative effects directly from observational data which can lead to tighter constraints on cloud feedbacks compared to relying on numerical models (Klein et al., 2018; Ceppi & Nowack, 2021; Myers et al., 2021). But a lot of these methods rely on coarse scale observational products, therefore ignoring

[1]AOPP, Department of Physics, University of Oxford. Correspondence to: Tim Reichelt <tim.reichelt@physics.ox.ac.uk>.

Accepted at ICML 2026 AI for Science Workshop. *Copyright 2026 by the author(s).*

most of the information content in high-resolution data that encapsulates the underlying physics.

However, we have an extensive observational record of mesoscale cloud systems from satellite missions. For example, the Moderate Resolution Imaging Spectroradiometer (MODIS) satellites have been capturing spectrally resolved $\sim$1 km scale resolution data for the last 20 years, providing a valuable resource for studying mesoscale cloud patterns and their response to increased CO_2 emissions. Data at this resolution, resolving mesoscale cloud structures, is currently underutilized in our efforts to constrain cloud feedbacks.

In this work, we present CloudFlow, a conditional flow matching model that generates high-resolution cloud fields from coarse scale atmospheric conditions. CloudFlow takes as input atmospheric profiles of temperature, humidity, and wind from reanalysis data, and generates MODIS calibrated radiances and cloud properties. We believe that CloudFlow will be a useful tool to study how large scale atmospheric conditions influence mesoscale cloud patterns and help to constrain cloud feedbacks, the dominant uncertainty in climate predictions.

2. CloudFlow

Our goal is to generate satellite data which we model as a multi-channel image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ with C output channels. Additionally, we assume that we have access to environmental conditions $\mathbf{c} = \{\mathbf{c}^A, \mathbf{c}^S\}$ which break down into atmospheric profiles $\mathbf{c}^A \in \mathbb{R}^{N \times L \times V_A}$, with L levels and V_A variables, and a land/sea mask $\mathbf{c}^S \in \mathbb{R}^{N \times V_S}$. Crucially, the satellite data and the environmental conditions do not lie on a shared grid, the environmental conditions are generally much coarser resolution than the satellite data, i.e. $N \ll H\,W$. For modeling purposes, we assume both the satellite data and the environmental conditions are generated from some data generating distribution $\mathbf{x}, \mathbf{c} \sim p_{\text{data}}(\mathbf{x}, \mathbf{c})$.

Due to their success in image modeling domains (Esser et al., 2024), we use a flow matching formalism (Lipman et al., 2024; Li & He, 2025) to learn a conditional generative model that generates the satellite data conditioned on environmental conditions. As is standard in flow matching models, during training we generate interpolations between the training data and random noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, as follows

$$\mathbf{z}_t = t\mathbf{x} + (1-t)\epsilon \tag{1}$$

with t sampled uniformly between 0 and 1.

Under the formalism introduced by Li & He (2025), we follow the **x-prediction** and **v-loss** setup in which we train a neural network $\mathbf{x}_\theta(\mathbf{z}_t, \mathbf{c}, t)$ that learns to directly predict denoised data from noisy quantities $\mathbf{z}_t$. This is done by optimizing the loss

$$\mathcal{L} = \mathbb{E}_{\mathbf{x},\mathbf{c},\epsilon,t}\left[\mathbf{v}_\theta - \mathbf{v}\right] \tag{2}$$

where $\mathbf{v}_\theta := (\mathbf{x}_\theta(\mathbf{z}_t, \mathbf{c}, t) - \mathbf{z}_t)/(1-t)$ and $\mathbf{v} := \mathbf{x} - \epsilon$.

To parameterize $\mathbf{x}_\theta(\mathbf{z}_t, \mathbf{c}, t)$, we use the same U-Net backbone as the CorrDiff model (Mardani et al., 2025) due to its success in downscaling coarse scale numerical model data to high-resolution fields.

2.1. Preprocessing of Environmental Conditions

Internally, the network $\mathbf{x}_\theta(\mathbf{z}_t, \mathbf{c}, t)$ transforms the environmental conditions $\mathbf{c}$ into a latent representation $\mathbf{h} \in \mathbb{R}^{N \times D}$. This preprocessing operates on each column of environmental conditions $\mathbf{c}_i^A$ and $\mathbf{c}_i^S$ independently. The environmental conditions are passed to the U-Net backbone by matching each pixel of satellite data to the closest column of environmental conditions so that the latent representation of environmental conditions is effectively upsampled to a $H \times W$ resolution.

We consider two separate architectures to do the preprocessing, the first being a simple linear projection of the atmospheric conditions and the second being a cross-attention mechanism that is inspired by the pressure-level encoding of the Aurora model (Bodnar et al., 2025).

2.1.1. Linear transform

The linear preprocessing architecture simply flattens the atmospheric profiles $\mathbf{c}_i^A$ into a LV_A length vector before passing it through a feedforward neural network, resulting in the transformation:

$$\mathbf{h}_i = \text{Concat}(\text{MLP}(\text{Flatten}(\mathbf{c}_i^A)), \mathbf{c}_i^S). \tag{3}$$

2.1.2. Attention transform

For the attention-based preprocessing we first encode all variables at each pressure level into a latent representation

$$\mathbf{z}_{i,j} = \text{MLP}(\mathbf{c}_{i,j}^A) + \mathbf{p}_j \tag{4}$$

with $\mathbf{p}_j$ denoting a pressure-level Fourier encoding for the pressure level j. Each of these encodings is then passed through a cross-attention mechanism

$$\text{Att}(\mathbf{z}_i) = \text{softmax}\left(\frac{Q(\mathbf{z}_i W_k)^\top}{\sqrt{d_a}}\right)\mathbf{z}_i W_v \tag{5}$$

with trainable parameters $Q \in \mathbb{R}^{D_q \times D_a}$, $W_k \in \mathbb{R}^{D_a \times D_a}$, and $W_v \in \mathbb{R}^{D_a \times D_a}$. We use multiple attention heads that are then projected into a shared representation space with a projection matrix $W_o \in \mathbb{R}^{D_a \times D}$, as follows

$$\text{MultiHead}(\mathbf{z}_i) = \text{Concat}(\text{Att}_1(\mathbf{z}_i), \ldots, \text{Att}_M(\mathbf{z}_i))W_o \tag{6}$$

The final latent representation for the environmental conditions is then given by passing the learned encodings through

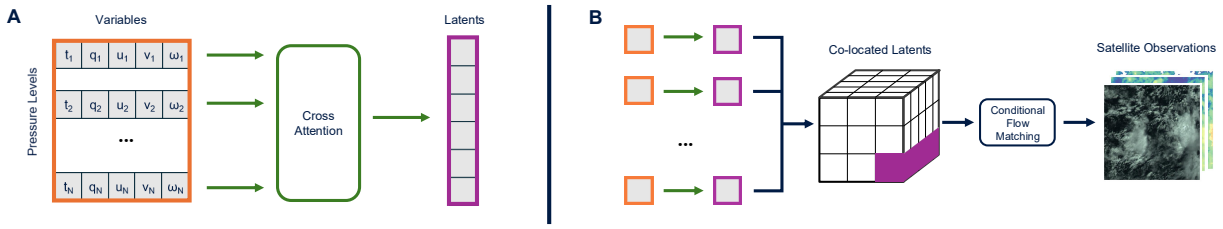

Figure 1. CloudFlow model components. **A** Attention encoder: the environmental conditions in the form of ERA5 atmospheric profiles are passed through a cross-attention encoder to produce a fixed dimensional latent. **B** Conditional flow matching: the attention encoder is applied independently to each ERA5 grid cell. The ERA5 grid cells are then co-located with the coordinates from the satellite observations before being passed in as conditioning information for the flow matching training. The whole setup is trained end-to-end with the attention encoder weights being optimized on the flow matching objective.

the multi-head attention mechanism

$$\mathbf{h}_i^A = \text{MultiHead}(\mathbf{z}_i) \tag{7}$$

These atmospheric latent embeddings are then concatenated with the land/sea mask for the final embedding: $\mathbf{h}_i = \text{Concat}(\mathbf{h}_i^A, \mathbf{c}_i^S)$.

Compared to the linear transform, the attention-based mechanism introduces an inductive bias by grouping variables at the same pressure level together and by allowing different attention heads and queries to attend to different parts of the atmospheric profiles.

3. Related Work

Cloud Controlling Factors (CCFs). The cloud controlling factor approach has been successful in constraining cloud feedbacks using observational data (Stevens & Brenguier, 2009; Klein et al., 2018; Ceppi & Nowack, 2021; Myers et al., 2021; Cesana & Del Genio, 2021). CCFs are manually defined physics-informed features derived from atmospheric profiles that are used to predict the cloud radiative effect in observational data (sometimes, large-eddy simulations are used instead of observational data as well (Zhang et al., 2012)). Some CCFs, like convergence and divergence of wind fields, do not depend on a single profile but instead explicitly account for the full 3D (latitude, longitude, vertical) organisational structure of environmental conditions. Common choices of cloud controlling factors include: estimated inversion strength (EIS) (Wood & Bretherton, 2006), surface temperature, vertical velocity at 500 hPa, upper tropospheric relative humidity (RH), and 700-hPa RH. A common workflow of using CCFs to constrain cloud feedbacks with observational data proceeds as follows: 1. Compute historical statistics of CCFs on atmospheric profiles from reanalysis datasets; 2. Based on the computed CCFs, fit a linear statistical model to predict cloud radiative effects in observational datasets such as CERES (Loeb et al., 2018); 3. Compute potential future values of CCFs from atmospheric profiles taken from GCM predictions; 4. Use the fitted linear model and the CCFs derived from GCM predictions to constrain cloud feedbacks. While this approach has been especially successful in constraining the feedback for low-level marine clouds, it has proven harder to find appropriate CCFs for high-clouds (Wilson Kemsley et al., 2024). The preprocessing step of the CloudFlow model can be interpreted as a fully data driven method to learn cloud controlling factors. Rather than using manually defined CCFs that discard most of the information from the atmospheric profiles, CloudFlow can learn which aspects of the atmospheric profiles are relevant to generate the MODIS observed cloud structures.

Generative Models in Weather and Climate. In recent years, data-driven generative models have shown the ability to increase forecast skill in weather forecasting compared to physics-based numerical models (Price et al., 2025; Bodnar et al., 2025; Couairon et al., 2024; Alet et al., 2025; Kochkov et al., 2024). However, these models are unsuitable for long-term climate predictions due to instabilities in long autoregressive rollouts. Moreover, their training data is insuffient to constrain cloud feedbacks, limiting their generalisability to future climate. Models such as ACE (Watt-Meyer et al., 2023; 2025) aim to address these issue and show initial promise but their reliability for climate projections is yet to be proven and they rely on highly uncertain GCMs to learn cloud feedbacks. In any case, these models operate at the coarse scale of reanalysis data and therefore do not explicitly resolve clouds which limits their ability to constrain the physics underlying cloud feedbacks. The cBottle class of models (Brenowitz et al., 2025) proposes an alternative mechanism to do climate projections that is not based on autoregressive rollouts and that is able to generate output fields at kilometre scale resolution. However, cBottle still inherits the biases of the underlying numerical physics-based models on which it is trained on. This is why instead we use observational satellite data as our generation target to avoid inheriting any biases from numerical models.

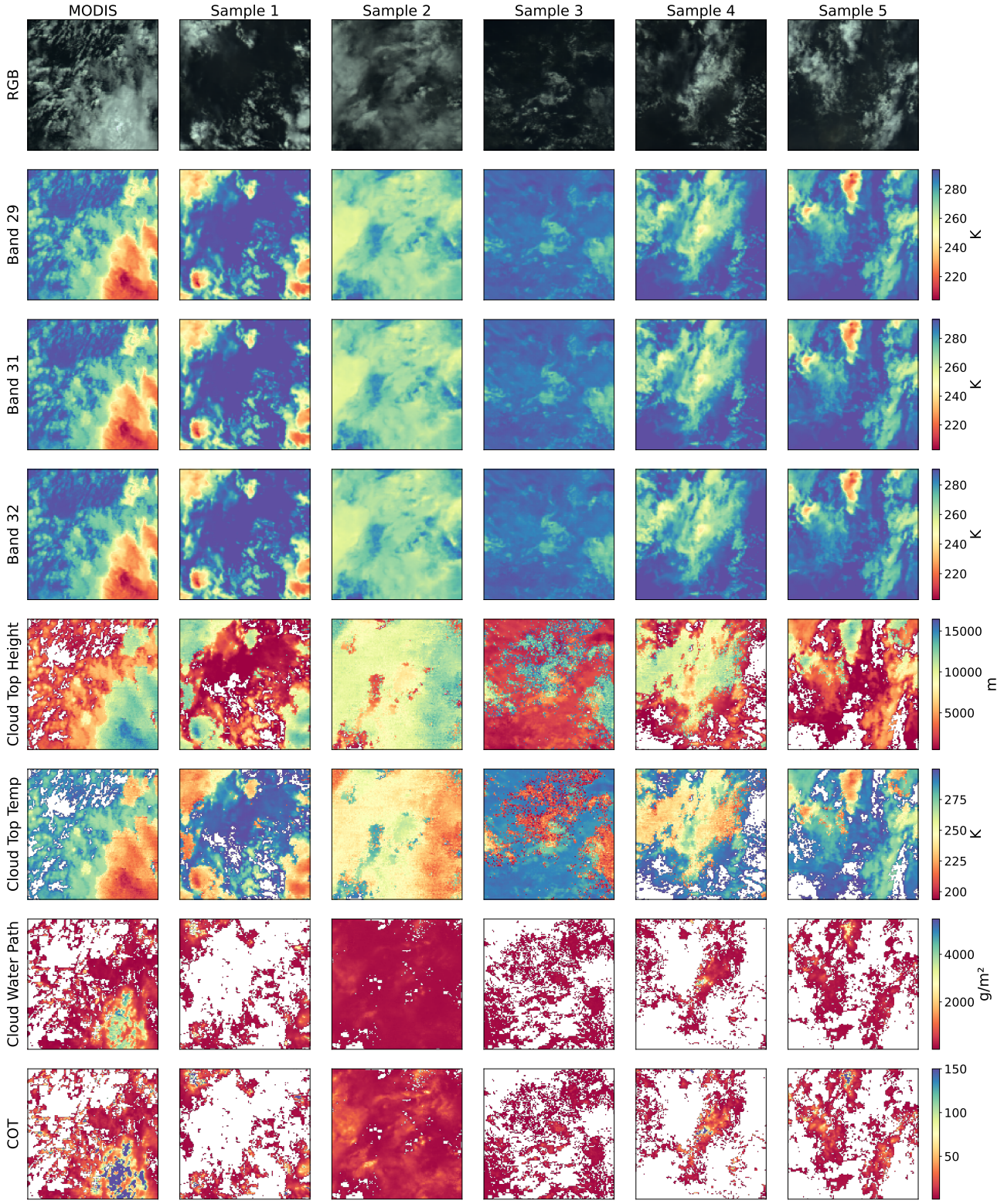

Figure 2. Sample generations for the CloudFlow (Att) model. First column contains the true MODIS data and the remaining columns show samples from our model.

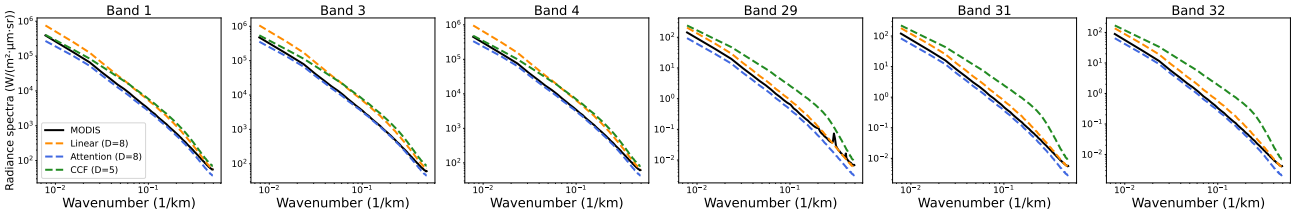

Figure 3. Comparison of the power spectra of MODIS calibrated radiances and model generations averaged over all test tiles. Spectra for the cloud properties are excluded because they often contain a large proportion of missing values. Note that the bumps in the Band 29 spectra are due to known striping artifacts in the MODIS sensor.

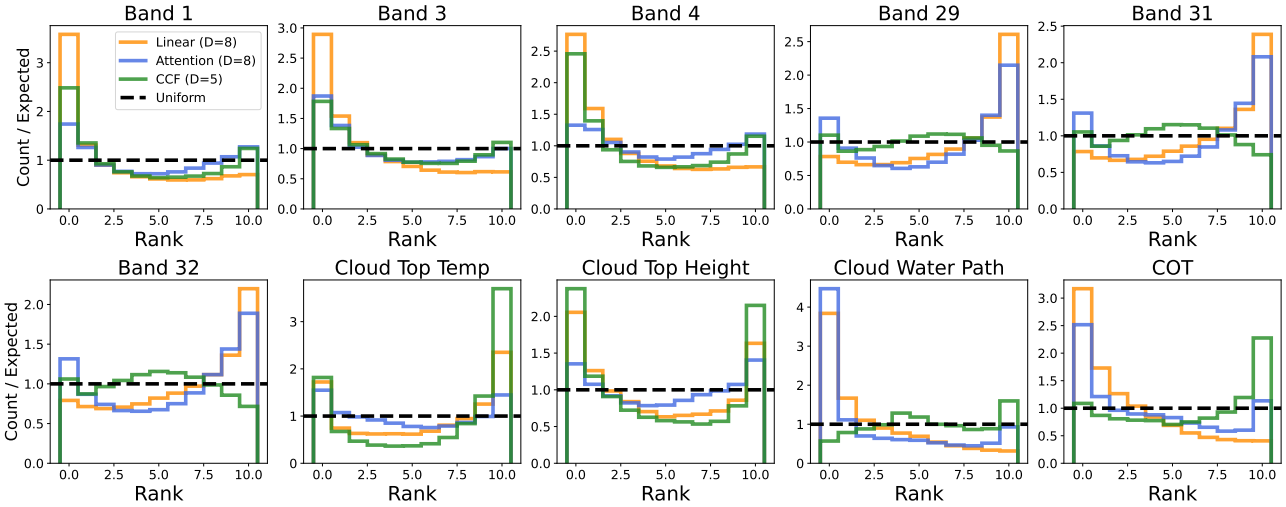

Figure 4. Rank histogram calibration plots. To produce these plots, first for a given ground-truth MODIS tile and pixel location the model generated values are sorted from smallest to largest. Then over all tiles and pixels the we aggregate what rank of ensemble member is closest to the true observations. A perfectly calibrated model would result in a uniform distribution.

4. Experimental Results

Data. We train our model on multiple channels of MODIS satellite data, namely we generate the raw calibrated radiances of the bands 1, 3, 4, 29, 31, 32.[1] Additionally, we also directly generate corresponding MODIS level 2 retrieved cloud properties of cloud top temperature (CTT), cloud top height (CTH), cloud water path (CWP), and cloud optical thickness (COT). As environmental conditions, we use atmospheric profiles of temperature, specific humidity, as well as the vertical, zonal, and meridional wind components taken from ERA5 reanalysis data (Hersbach et al., 2020) using all 37 available pressure levels. The only explicit surface level variable we consider is a binary land/sea mask. We deliberately do not include any location information such as latitude and longitude metadata to prevent the model overfitting onto the climatologies of geographical regions. Overall, the MODIS satellite data is available at ∼1 km resolution while the ERA5 data is at roughly 25 km resolution. For training, we extract 40,000 MODIS tiles of size 128×128 pixels collocated with ERA5 data from the years 2003, 2006, 2008, and 2011. We use 1,000 tiles MODIS tiles from 2018 as a test set for evaluation.

Model hyperparameters. Overall, we train two versions of the CloudFlow model described in Section 2, one using the linear transform and another using the attention based transform which we will refer to as **CloudFlow (Linear)** and **CloudFlow (Att)**, respectively. Both of these models compress the atmospheric profiles into latents of dimensionality $D = 8$. For training the models we use the Adam optimizer with learning rate 0.0002 and batch size 32. All models are optimized for 185,000 iterations.

Baseline. As a baseline model we consider using a flow matching model that only takes in the cloud controlling factors used in Ceppi & Nowack (2021) rather than the full atmospheric profiles. Namely, these factors are: estimated inversion strength (EIS), surface temperature, vertical velocity at 500 hPa, upper tropospheric relative humidity

[1] Bands 1 (620-670nm), 4 (545-565nm), and 3 (459-479nm) correspond to red, green, and blue visible light, whereas channels 29 (8.4-8.7μm), 31 (10.78-11.28μm), and 32 (11.77-12.27μm) measure infrared radiation that are important to capture cloud properties.

Table 1. Mean absolute error (MAE, lower is better) for estimating the tile-wide mean per channel. The best model per channel is highlighted in bold. Additionally, models that are not significantly worse according to the Wilcoxon signed-rank test ($p \geq 0.05$) are also highlighted in bold. Units: B$\{1, 3, 4\}$ ($\mathrm{W\,m^{-2}\,\mu m^{-1}\,sr^{-1}}$), B$\{29, 31, 32\}$ (K), CTT (K), CTH (m), CWP ($\mathrm{g\,m^{-2}}$).

Model	B1	B3	B4	B29	B31	B32	CTT	CTH	CWP	COT
Linear (D=8)	47.60	53.85	50.67	7.07	7.76	7.69	16.32	2473.32	85.62	5.76
Attention (D=8)	**31.16**	**35.00**	**32.62**	**4.89**	**5.41**	**5.43**	**13.92**	**2116.99**	92.64	**4.20**
CCF (D=5)	40.38	44.76	42.98	6.67	7.30	7.43	16.56	2415.23	**58.13**	**4.29**

Table 2. Continuous Ranked Probability Score (CRPS, lower is better) for estimating the tile-wide mean per channel. The best model per channel is highlighted in bold. Additionally, models that are not significantly worse according to the Wilcoxon signed-rank test ($p \geq 0.05$) are also highlighted in bold. Units: B$\{1, 3, 4\}$ ($\mathrm{W\,m^{-2}\,\mu m^{-1}\,sr^{-1}}$), B$\{29, 31, 32\}$ (K), CTT (K), CTH (m), CWP ($\mathrm{g\,m^{-2}}$).

Model	B1	B3	B4	B29	B31	B32	CTT	CTH	CWP	COT
Linear (D=8)	34.63	38.79	36.67	5.07	5.55	5.49	11.29	1709.95	60.03	4.00
Attention (D=8)	**23.16**	**25.56**	**24.25**	**3.69**	**4.06**	**4.06**	**9.76**	**1530.69**	69.39	**2.97**
CCF (D=5)	29.84	33.12	31.69	5.02	5.49	5.61	12.33	1800.68	**42.48**	**3.11**

(RH), 700-hPa RH. We will denote this model as **CloudFlow (CCF)** and it can be interpreted as replacing the learnt atmospheric profile processing with a hand-designed feature transform that results in latent representations $\mathbf{h}$ with $D = 5$. To ensure a fair comparison, these CCFs are computed from the exact same ERA5 profiles that are used to train the other CloudFlow models and we re-use the same optimization hyperparameters.

Sampling. To generate samples from the trained models we use an Euler integration scheme with 30 time steps. For each test data point we generate 10 ensemble members.

Evaluation. In general, because the specified environmental conditions are so coarse and the chaotic nature of this nonlinear system, we do not expect CloudFlow to be able to perfectly reconstruct the original MODIS image with pixel-wise accuracy. Therefore, we do not use pixel-wise error metrics to evaluate and compare the trained models. Instead, we validate whether the models are able to generate images with realistic spower spectra (Figure 3), are well calibrated in their probablistic predictions (Figure 4) and whether the models predict the correct tile-level means for each channel (Tables 1 and 2).

Example generations of the CloudFlow (Att) model are shown in Figure 2 along with the corresponding ground-truth MODIS observations. The generations show that it is able to generate realistic looking samples.Importantly, the generations for individual samples are consistent across channel, especially the longwave bands (bands 29, 31, 32) are consistent with the cloud property retrievals, e.g. cooler brightness temperatures in the longwave bands coincide with higher cloud top height retrievals.

To consider more quantitative metrics, Figure 3 shows that CloudFlow generates realistic power spectra that match well with the MODIS spectra even for high wavenumbers. This is consistent with results from the weather forecasting community that demonstrated that generative models have the ability to have "sharp" predictions with realistic spectra (Price et al., 2025). Based on manual inspection of the generations of the CCF model, the spectral bias can be explained by the fact that the CCF model tends to have a bias to generate cloudy scenes, even if the true MODIS image does not contain any clouds.

The rank histograms in Figure 4 show that the CloudFlow (Att) model is generally well calibrated across the different output bands, although it has a tendency to be underdispersive. It especially struggles with the Cloud Water Path (CWP) generations, often over-estimating the true value. The CCF model is surprisingly well calibrated for this output channel given that CWP is a column integrated quantity and the CCF model does not get access to the full atmospheric profiles.

Tables 1 and 2 demonstrate that CloudFlow (Att) is able to better predict per-channel averages compared to CloudFlow (Linear) or the CCF model for most of the output bands. For the level 1 MODIS bands CloudFlow with an attention encoder is consistenly the best model. One thing to note is that the MAE for the mean estimate in cloud top height (CTH) is around 2 km. To put this into context, the MODIS derived CTH is itself an approximation because it relies solely on passive sensing. When comparing MODIS CTH values to more reliably CTH estimates derived from active sensors the MODIS retrieval error has been shown to average around 1-2 km for high-level cirrus clouds and 100-500 m for low-level clouds (Menzel & Strabala, 1997). Hence, the errors in the CloudFlow model predictions are roughly on the same order of magnitude than some of the MODIS retrieval errors.

In general, the baseline model using the hand-defined physics-informed CCFs provides a surprisingly strong baseline, often outperforming the simple linear encoder and sometimes even the attention based encoder. This demonstrates that the CCFs are already successfully compressing a lot of the information content contained in the full atmospheric profiles and that it is a non-trivial task to find features that provide better predictive capabilities of the cloud fields. Notably, the estimated inversion strength (EIS) and relative humidity quantities are quantities that require non-linear transforms of the atmospheric profile data and therefore these features cannot be reproduced exactly by the linear encoder model.

5. Conclusion

Our limited understanding of how atmospheric environmental conditions influence cloud organisational patterns constitutes one of the biggest sources of uncertainties in climate predictions. We presented CloudFlow, a flow matching model that resolves mesoscale cloud structures conditioned on atmospheric environmental conditions of temperature, humidity, and wind. Our experiments demonstrated that CloudFlow is able to generate realistic mesoscale structures. Its generations have power spectra that match observations, its predictions are well calibrated for most output channels (with a tendency to be slightly underdispersive), and the features learned by CloudFlow's preprocessing step lead to lower reconstruction errors compared to hand-designed CCFs.

The natural next step is to use CloudFlow to tackle meaningful research questions in climate science. One promising avenue is using CloudFlow to study and explain existing taxonomies of cloud structures. For example, Stevens et al. (2020) categorised clouds in the trade wind regions into four visually distinct classes called "sugar", "gravel", "flower", and "fish". Traditionally, studies used large-eddy simulations (LES) to investigate how environmental conditions impact these organisational patterns (Jansson et al., 2023). However, running an LES even for small domains often takes hours on a supercomputer whereas sampling from CloudFlow can be done in seconds on a single A100 NVIDIA GPU. CloudFlow therefore provides a cheaper, more interactive model to study cloud organisational patterns with the potential to constrain cloud feedbacks globally.

Additionally, it will be insightful to use tools from mechanistic interpretability (Bereska & Gavves, 2024) to study how CloudFlow uses the atmospheric profile information and whether the features learned by the model can be mapped onto existing cloud controlling factors or whether they constitute new, previously unconsidered factors. Tools from mechanistic interpretability have recently been successfully applied to study the internal mechanisms of data-driven weather forecast models (MacMillan & Ouellette, 2025) and we believe they can also help to understand the internals of the CloudFlow models.

Overall, CloudFlow extends the types of models available to climate researcher which we hope will contribute towards new lines of evidence for constraining cloud feedbacks that provides complementary strength compared to existing modelling approaches (Sherwood et al., 2020; Arias et al., 2021).

Software and Data

The code for CloudFlow is openly available at `https://github.com/treigerm/cloudflow`.

Acknowledgements

Tim Reichelt received funding from ARIA and DSIT and Pillar VC under the Encode: AI for Science Fellowship. Philip Stier acknowledges funding from the EU's Horizon Europe program under grant agreement number 10113184 and also acknowledges funding from UK Research and Innovation (UKRI).

Impact Statement

This paper presents work whose goal is to leverage tools from machine learning to advance the field of climate sciences. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alet, F., Price, I., El-Kadi, A., Masters, D., Markou, S., Andersson, T. R., Stott, J., Lam, R., Willson, M., Sanchez-Gonzalez, A., et al. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025.

Arias, P., Bellouin, N., Coppola, E., Jones, R., Krinner, G., Marotzke, J., Naik, V., Palmer, M., Plattner, G.-K., Rogelj, J., et al. Climate change 2021: the physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change; technical summary. 2021.

Bereska, L. and Gavves, E. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Bony, S., Stevens, B., Frierson, D. M., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T. G., Sherwood, S. C., Siebesma, A. P., Sobel, A. H., et al. Clouds, circulation and climate sensitivity. *Nature Geoscience*, 8(4): 261–268, 2015.

Brenowitz, N. D., Ge, T., Subramaniam, A., Manshausen, P., Gupta, A., Hall, D. M., Mardani, M., Vahdat, A., Kashinath, K., and Pritchard, M. S. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere. *arXiv preprint arXiv:2505.06474*, 2025.

Bretherton, C. and Blossey, P. Understanding mesoscale aggregation of shallow cumulus convection using large-eddy simulation. *Journal of Advances in Modeling Earth Systems*, 9(8):2798–2821, 2017.

Bretherton, C. S. Insights into low-latitude cloud feedbacks from high-resolution models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2054), 2015.

Ceppi, P. and Nowack, P. Observational evidence that cloud feedback amplifies global warming. *Proceedings of the National Academy of Sciences*, 118(30):e2026290118, 2021.

Cesana, G. V. and Del Genio, A. D. Observational constraint on cloud feedbacks suggests moderate climate sensitivity. *Nature Climate Change*, 11(3):213–218, 2021.

Couairon, G., Singh, R., Charantonis, A., Lessig, C., and Monteleoni, C. Archesweather & archesweathergen: a deterministic and generative model for efficient ml weather forecasting. *arXiv preprint arXiv:2412.12971*, 2024.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730): 1999–2049, 2020.

Jansson, F., Janssens, M., Grönqvist, J. H., Siebesma, A. P., Glassmeier, F., Attema, J., Azizi, V., Satoh, M., Sato, Y., Schulz, H., et al. Cloud botany: Shallow cumulus clouds in an ensemble of idealized large-domain large-eddy simulations of the trades. *Journal of Advances in Modeling Earth Systems*, 15(11):e2023MS003796, 2023.

Klein, S. A., Hall, A., Norris, J. R., and Pincus, R. Low-cloud feedbacks from cloud-controlling factors: A review. *Shallow clouds, water vapor, circulation, and climate sensitivity*, pp. 135–157, 2018.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Li, T. and He, K. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R. T., Lopez-Paz, D., Ben-Hamu, H., and Gat, I. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., Liang, L., Mitrescu, C., Rose, F. G., and Kato, S. Clouds and the earth's radiant energy system (ceres) energy balanced and filled (ebaf) top-of-atmosphere (toa) edition-4.0 data product. *Journal of climate*, 31(2):895–918, 2018.

MacMillan, T. and Ouellette, N. T. Towards mechanistic understanding in a data-driven weather model: internal activations reveal interpretable physical features. *arXiv preprint arXiv:2512.24440*, 2025.

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M. A., Ge, T., Subramaniam, A., et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025.

Menzel, W. P. and Strabala, K. *Cloud top properties and cloud phase algorithm theoretical basis document*. University of Wisconsin–Madison, 1997.

Myers, T. A., Scott, R. C., Zelinka, M. D., Klein, S. A., Norris, J. R., and Caldwell, P. M. Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity. *Nature Climate Change*, 11(6):501–507, 2021.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

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

Segura, H., Pedruzo-Bagazgoitia, X., Weiss, P., Müller, S. K., Rackow, T., Lee, J., Dolores-Tesillos, E., Benedict, I., Aengenheyster, M., Aguridan, R., et al. nextgems: entering the era of kilometer-scale earth system modeling. *Geoscientific Model Development*, 18(20):7735–7761, 2025.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., et al. An assessment of earth's climate sensitivity using multiple lines of evidence. *Reviews of geophysics*, 58(4):e2019RG000678, 2020.

Stevens, B. and Bony, S. What are climate models missing? *science*, 340(6136):1053–1054, 2013.

Stevens, B. and Brenguier, J.-L. Cloud-controlling factors: Low clouds. In Heintzenberg, J. and Charlson, R. J. (eds.), *Clouds in the Perturbed Climate System: Their Relationship to Energy Balance, Atmospheric Dynamics, and Precipitation*. The MIT Press, 02 2009. ISBN 9780262012874. doi: 10.7551/mitpress/9780262012874.003.0008. URL https://doi.org/10.7551/mitpress/9780262012874.003.0008.

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M., Klocke, D., et al. Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, 6(1): 61, 2019.

Stevens, B., Bony, S., Brogniez, H., Hentgen, L., Hohenegger, C., Kiemle, C., L'Ecuyer, T. S., Naumann, A. K., Schulz, H., Siebesma, P. A., et al. Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds. *Quarterly Journal of the Royal Meteorological Society*, 146(726):141–152, 2020.

Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S. K., Henn, B., Duncan, J., Brenowitz, N. D., Kashinath, K., Pritchard, M. S., Bonev, B., et al. Ace: A fast, skillful learned global atmospheric model for climate prediction. *arXiv preprint arXiv:2310.02074*, 2023.

Watt-Meyer, O., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., Wu, E., Harris, L., and Bretherton, C. S. Ace2: accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, 8(1):205, 2025.

Wilson Kemsley, S., Ceppi, P., Andersen, H., Cermak, J., Stier, P., and Nowack, P. A systematic evaluation of high-cloud controlling factors. *Atmospheric Chemistry and Physics*, 24(14):8295–8316, 2024.

Wood, R. and Bretherton, C. S. On the relationship between stratiform low cloud cover and lower-tropospheric stability. *Journal of climate*, 19(24):6425–6432, 2006.

Zhang, M., Bretherton, C. S., Blossey, P. N., Bony, S., Brient, F., and Golaz, J.-C. The cgils experimental design to investigate low cloud feedbacks in general circulation models by using single-column and large-eddy simulation models. *Journal of Advances in Modeling Earth Systems*, 4(4), 2012.