LEARNING EXTREME TEMPERATURE REGIMES

Shirin Goshtasbpour, Maxim Samarin & Michele Volpi
Swiss Data Science Center
ETH Zurich and EPFL
Zurich, Switzerland
[firstname].[lastname]@sdsc.ethz.ch

ABSTRACT

Recent changes in climate made previously predictable temperature and weather patterns increasingly unreliable, giving rise to increased volatility and extreme events such as prolonged heat waves, abrupt cold spells, and erratic temperature shifts. This growing unpredictability challenges the capabilities of physics-based climate models, particularly as low-return-rate temperature patterns become more common. In this paper, we present a machine learning approach based on ClimODE to generate projections of future climate scenarios conditional on specific temperature quantiles, treated as classes. Our Uniform Quantile ClimODE (*UQClimODE*) approach presents itself as a promising tool to capture these atypical patterns, to identify localized impacts, and to enable proactive planning for climate adaptation and resilience under different scenarios.

1 INTRODUCTION

In recent decades, frequent heatwaves and persistent high summer temperatures have been observed across the globe, shifting significantly from past climate patterns (Materia et al., 2024; Hao et al., 2022; Fischer et al., 2021). Despite their substantial societal and environmental consequences, analyzing such conditions remains challenging due to limited compound observations of extremes and the high computational cost of simulating such events (Barriopedro et al., 2023). Recent works are focusing on modeling rare events through phenomenological theories of their occurrence (Horton et al., 2016; Messori et al., 2018), incorporating rare event algorithms (Lucarini et al., 2023; Ragone & Bouchet, 2021), large deviation theory (Gálfi et al., 2021; Galfi & Lucarini, 2021; Ragone et al., 2018) and extreme value theory (Philip et al., 2022; Thompson et al., 2022), and ensemble boosting (Fischer et al., 2023) to reproduce similar conditions and simulate meaningful scenarios.
From a statistical perspective, generative models can help simulating, reproducing, and analyzing extreme weather conditions. This capability can inform decision-making, provide stress tests and scenarios for disaster preparedness, infrastructure planning, and climate adaptation and mitigation strategies, ultimately contributing to more effective responses to the growing challenges posed by climate change.
This paper presents the Uniform Quantile ClimODE (*UQClimODE*) generative model, an initial step towards generating specific but realistic time series of extreme weather conditions. It offers a novel statistical framework to study abrupt climate changes and their regional impacts. Due to the scarcity of atypical and rare observations, we leverage ERA5 reanalysis data (Hersbach et al., 2020) to capture underlying climate dynamics and model coherent scenarios in space and time. We further guide the model to generate specific atypical temperature regimes through discriminative training (see Figure 1).

2 METHODOLOGY

Throughout the paper $t \in \mathbb{R}$ denotes time and $\boldsymbol{x} = (h, w) \in \Omega = [-90°, 90°] \times [-180°, 180°]$ denotes the location in latitude-longitude encoding. $\nabla_{\boldsymbol{x}}$, ∇ in short, denotes spatial gradients with respect to $\boldsymbol{x}$, $\dot{u} = \frac{du}{dt}$ is the time derivative of the variable u, $\cdot$ is the inner product on the spatial coordinates and for the spherical vector $\boldsymbol{v}$, $\nabla \cdot \boldsymbol{v} = \text{Tr}(\nabla \boldsymbol{v})$ is the divergence. We assume that the climate variables are described by a K-dimensional spatio-temporal process $u(\boldsymbol{x}, t) = \big(u_1(\boldsymbol{x}, t), ..., u_K(\boldsymbol{x}, t)\big)$ ($u_k : \Omega \times \mathbb{R} \to \mathbb{R}$). The target region, denoted by $\mathcal{X} \subseteq \Omega$ is the region of focus of analysis. We use $u_{\mathcal{X}}(t) = \big(u(\boldsymbol{x}, t)\big)_{\boldsymbol{x} \in \mathcal{X}}$ to denote the climate process *restricted* on the region $\mathcal{X}$. Additionally, let $\mathcal{U}(\mathcal{X})$ represent the space of all possible values of $u_{\mathcal{X}}$.

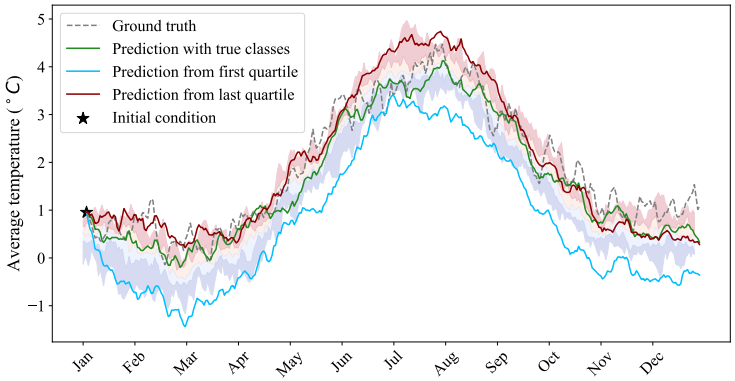

Figure 1: Average global predicted temperature for the most likely 365-day sequences in 2017. The dashed line corresponds to the observed mean, while shaded areas depict the different observed quartile classes. The first ("cold") and last ("hot") predicted quartiles are shown in blue and red, respectively. The star marks the initial condition to which the model simulates the season.

To generate scenarios with specific conditions, we systematically classify the restricted climate process into distinct categories. These categories are determined based on the values of a *query function* applied to the restricted climate process. Each instance of the restricted climate process is then assigned a label corresponding to its respective category.
Formally, the **query function** $q : \mathcal{U}(\mathcal{X}) \to \mathbb{R}$ is any measure of the restricted climate process, such as the average temperature over the region $\mathcal{X}$. We partition the output of the query function into Q distinct **target classes** and assign labels $y \in [Q]$ to instances of restricted climate process $u_{\mathcal{X}}(t)$, based on the value of their query function. Given the query function, the target classes are extracted in a pre-processing step and the labels are paired with the observations $u_{\Omega}(t)$. In this paper, we label each observation using the quartiles of the average temperature over the $\mathcal{X}$ region for each day of the year. Specifically, observations with an average temperature lower than the first quartile of that day's average temperature are assigned label 1 ("cold" regime, referred to as "first quartile"), whereas observations with an average temperature higher than the third quartile are assigned label 4 ("hot" regime, referred to as "last quartile"). These temperature classes are used to generate sequences that capture heat waves and cold spells within the region of focus for analysis while reflecting the average observed dynamics. To model the climate process, we build on the recent success of the ClimODE model (Verma et al., 2024), adapting it to address the specific challenges of our problem.

2.1 ClimODE Model Overview

ClimODE is a physics-based model designed to improve the inductive bias of climate models, while increasing the interpretability of its predictions. It relies on the Partial Differential Equation (PDE) of advection (1 and 2) as a prior to explain the evolution of the K-dimensional climate process $u(\boldsymbol{x}, t)$. Along the advection process, the transport and concentration of the air mass are modeled by auxiliary flow vector fields $\boldsymbol{v}(\boldsymbol{x}, t) = (\boldsymbol{v}_1(\boldsymbol{x}, t), ..., \boldsymbol{v}_K(\boldsymbol{x}, t))$ $(\boldsymbol{v}_k \in \Omega)$. Formally, for $1 \leq k \leq K$,

$$\dot{u}_k(\boldsymbol{x}, t) = -\nabla \cdot (u_k(\boldsymbol{x}, t)\boldsymbol{v}_k(\boldsymbol{x}, t)), \tag{1}$$

$$\dot{\boldsymbol{v}}_k(\boldsymbol{x}, t) = f(u(\boldsymbol{x}, t)), \tag{2}$$

where the function f is an unknown function of the evolution of the flow vector field at time t and location $\boldsymbol{x}$. $f(\cdot)$ is approximated by a function $f_\theta(\cdot, \boldsymbol{v}(\boldsymbol{x}, t), \psi(\boldsymbol{x}, t))$ parametrized by θ and additional context information $\psi(\boldsymbol{x}, t)$. The model accounts for the observation noise over the deterministic advection sequence with a Gaussian likelihood, $\mathcal{N}(u_{\text{obs}}(\boldsymbol{x}, t); u(\boldsymbol{x}, t) + \mu_\theta(\boldsymbol{x}, t), \Sigma_\theta(\boldsymbol{x}, t))$.
The ClimODE model is particularly valuable as it requires a low amount of training data and produces accurate sequential predictions in time, from low spatial resolution variables.

2.2 The proposed Uniform Quantile ClimODE Model

ClimODE is not suitable for generating sequences from predefined target classes, such as different temperature regimes. Since ClimODE is parameterized in terms of an ODE, we need to identify its initial conditions, solve the ODE, and backpropagate through the sequence. Furthermore, ClimODE implements attention blocks inefficiently and processes the observations without considering their

structure. The model struggles to generate temporally stable predictions and diverges when the sequence length extends beyond the one used during training. To improve scalability and produce reliable long-term sequences that align with the desired target classes of the process, we introduce the *UQClimODE* with the following modifications. UQClimODE models the climate process with the advection-inspired Stochastic Differential Equation (SDE),

$$du_k(\boldsymbol{x}, t) = -\nabla \cdot (u_k(\boldsymbol{x}, t)\boldsymbol{v}_k(\boldsymbol{x}, t))\, dt + \mu_\theta(\boldsymbol{x}, y)dt + \sqrt{2\Sigma}dB_t, \tag{3}$$

where dB_t is a Brownian motion (Oksendal, 2013). This difference to ClimODE provides a more natural representation of the process and enables several extensions that enhance the predictions.
Flow parametrization: Instead of the evolution function, we directly parameterize the flow vector field $\boldsymbol{v}_\theta(\boldsymbol{x}, t) = F_\theta(u(\boldsymbol{x}, t), \nabla u(\boldsymbol{x}, t), \psi(\boldsymbol{x}, t))$ and we use one-step-ahead predictions to learn the climate process. This in turn enables the use of teacher forcing (see Appendix A.1) during training, providing stronger gradient signals and stabilizing long-term prediction.
Feature extraction: We encode the day of the year t, the geographical coordinates h and w, and the target temperature class labels y (i.e. quartiles) using trigonometric embeddings,

$$\psi(\boldsymbol{x}, t) = \left[\sin\frac{2\pi h}{180}, \cos\frac{2\pi h}{180}, \sin\frac{2\pi w}{360}, \cos\frac{2\pi w}{360}, \sin\frac{2\pi t}{365}, \cos\frac{2\pi t}{365}, \sin\frac{y}{10000^{0.02}}, \cos\frac{y}{10000^{0.02}}\right].$$

These feature encodings are concatenated to climate variables $u(\boldsymbol{x}, t)$, the spatial gradient ∇u is evaluated with spherical padding of $u(\boldsymbol{x}, t)$ and used as input to the corresponding neural network.
Tokenizing observations: UQClimODE addresses the model scalability issue by leveraging powerful UNets (Ronneberger et al., 2015) for the parameterization of μ_θ and $\boldsymbol{v}_\theta$. UNets are typically used to estimate the drift function in diffusion models (Ho et al., 2020; Song et al., 2020), and for the SDE in Equation 3 they lead to more accurate predictions.
Decomposing seasonality and volatility: The model performs best when the advection flow network $\boldsymbol{v}_\theta$ incorporates all extracted features $\psi(\boldsymbol{x}, t)$, while the mean of the residual pattern μ_θ uses embeddings other than time. This result is intuitive, as temperature volatility, which largely determines the temperature class, is expected to be mostly independent of the seasonal variations. Therefore, the overall mean of the model likelihood is given by

$$u_\theta(\boldsymbol{x}, t + \Delta t, y) = u_{\text{obs}}(\boldsymbol{x}, t) + \Delta t\left(-\nabla \cdot (u_{\text{obs}}(\boldsymbol{x}, t)\boldsymbol{v}_\theta(\boldsymbol{x}, t)) + \mu_\theta(\boldsymbol{x}, y)\right). \tag{4}$$

Training loss: We jointly minimize the negative log-likelihood of the one-step-ahead observations and cross-entropy loss as

$$\mathcal{L}(\theta, \mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\left\{\log\mathcal{N}\left(u_{\text{obs}}^{(i)}; u_\theta^{(i)}, \Sigma\right) + \lambda \log \text{softmax}_{y^{(i)}}\left(L\left(q\left((u_\theta)_{\mathcal{X}}\right)\right)\right)\right\},$$

where $L : \mathbb{R} \to \mathbb{R}^Q$ is a single layer classifier that returns logits of the target classes based on the query function value of the prediction, and λ controls the contribution of cross-entropy.

3 Experiments

Our experimental setup is similar to Verma et al. (2024). To study temperature regimes, we use the temperature variable from ERA5 (`t`) (Hersbach et al., 2020) at a coarsened horizontal resolution of 5.625°, with 1-day aggregates, from 2006 to 2018. `t` is normalized in the range $[0, 1]$ using min-max scaling over the spatial and temporal span. The training data consists of ten years (2006–2015), with 2016 reserved for validation and 2017–2018 used as the test set. Further details can be found in Appendix A.1. We compare UQClimODE to ClimODE and its variants by sequentially predicting the global process for long periods of time, setting the target region $\mathcal{X}$ as the globe, and three subregions which exhibited heatwaves recently, from which we estimate the target class (Fischer et al., 2023; Thompson et al., 2022; Philip et al., 2022; Ragone & Bouchet, 2021).

3.1 Stable Long-Term Generation

As a baseline, we use the original ClimODE with a global prediction model of 9 days, and its variants, DAILY trained only with the daily temperature variable, and NOCONTEXT, which further removes contextual information (orography and information about the surface of the land model).

Table 1: CRPS comparison of predictions with ClimODE variants for global forecasting.

Model	Lead time (days) 1	3	7	14	30	90
ClimODE (24)	**0.010** ± **0.010**	**0.017** ± **0.018**	0.034 ± 0.041	NaN	NaN	NaN
DAILY (24)	0.017 ± 0.017	0.025 ± 0.027	0.034 ± 0.039	0.079 ± 0.072	NaN	NaN
DAILY (6)	0.017 ± 0.021	0.022 ± 0.026	0.027 ± 0.037	0.031 ± 0.041	0.052 ± 0.061	NaN
DAILY (4)	0.019 ± 0.022	0.023 ± 0.026	**0.026** ± **0.032**	**0.029** ± **0.036**	0.041 ± 0.049	NaN
NOCONTEXT (24)	0.022 ± 0.062	0.028 ± 0.064	0.035 ± 0.069	0.052 ± 0.076	0.108 ± 0.169	NaN
NOCONTEXT (6)	0.033 ± 0.076	0.036 ± 0.077	0.039 ± 0.079	0.047 ± 0.081	0.077 ± 0.089	NaN
UQClimODE	0.015 ± 0.01	0.024 ± 0.027	0.031 ± 0.034	0.034 ± 0.037	**0.035** ± **0.037**	**0.034** ± **0.036**

These variants are introduced to provide a fair comparison with UQClimODE, which is trained similarly to NOCONTEXT. As the performance of the variants is highly dependent on the number of integration steps per 24 hours, we specify this value for each model in the reported results. We evaluate the model performance using latitude-weighted RMSE, the Anomaly Correlation Coefficient (ACC), and the Continuous Ranked Probability Score (CRPS) (see Appendix A.1).

Table 1 shows the CRPS of sequential predictions with different lead times up to 90 days. We observe that the CLIMODE model outperforms its variants for short-term (1-3 day period), while training with daily signals (DAILY) and removing context information decreases the performance considerably. All ClimODE variants struggle to remain stable for longer prediction horizons, with CLIMODE failing and generating NaNs after 7 days. The integration resolution is critical, where better results in short-term prediction are possible with 24 integration steps per day and the CRPS values deteriorate with more resolution for longer lead times.

For UQClimODE, the 1-3 day prediction performance is similar to the best DAILY ClimODE variant, although UQClimODE does not rely on contextual information. Notably, with the proposed modifications, UQClimODE retains its performance for up to 90 days with only a slight deterioration in time. We observed a similar pattern for RMSE and ACC (see Appendix A.2). The model is able to remain in a small prediction error range for sequences as long as 365 days.

3.2 STEERING TEMPERATURE REGIMES IN DESIRED REGIONS

Here, we evaluate the ability of UQClimODE to generate sequential predictions from a specific target class in different regions of the planet. For this purpose, we train the model when the target classes are determined by the average daily temperature value over the entire world, the Pacific Northwest, Europe, and the Chicago area, respectively. We generate the most likely sequences of 90 and 360 days in each region, conditioning the generations on the coldest and the warmest classes (first and last quartile) of 2017 and 2018, respectively, and plot the average daily temperature. The global predictions are illustrated in Figure 1. Additional results on average global and regional temperatures, as well as exemplary temperature maps, are provided in Appendix A.3.

4 CONCLUSIONS AND IMPACT

In this work, we introduced UQClimODE, a class-conditional extension of ClimODE, designed to generate conditional forecasts for a climate process. In this study, we focus on predicting the temperature patterns within different temperature regimes. We model these dynamics through an advection-based SDE, which enables the model to better capture temperature volatility and long-term dependencies. We extend ClimODE by modifying its parametrization and optimization to improve stability and accuracy of long-term predictions. These enhancements address key limitations of the original approach, including scalability challenges and issues with model mismatch.

Our experimental evaluation demonstrates that the proposed model outperforms ClimODE baselines across global and regional forecasting tasks in long-term prediction. We further showcase the model's ability to capture extreme and atypical weather events like heat waves and cold spells, which are critical for climate impact analysis and adaptation planning.

This model represents a first attempt to generate classes of extreme weather conditions reliably. It provides a valuable tool for studying abrupt climate changes and identifying regional impact patterns. Moreover, it addresses the challenges posed by the lack of extensive datasets available for this type of analysis, offering a framework to better understand and simulate extreme climate scenarios.

ACKNOWLEDGMENTS: The authors are part of SPEED2ZERO, a Joint Initiative co-financed by the ETH Board.

REFERENCES

David Barriopedro, Ricardo F. García-Herrera, Carlos Ordóñez, Diego G. Miralles, and Sancho Salcedo-Sanz. Heat waves: Physical understanding and scientific challenges. *Reviews of Geophysics*, 61(2):e2022RG000780, 2023.

Erich M Fischer, Sebastian Sippel, and Reto Knutti. Increasing probability of record-shattering climate extremes. *Nature Climate Change*, 11(8):689–695, 2021.

Erich M Fischer, Urs Beyerle, Luna Bloin-Wibe, Clauia Gessner, Vincent Humphrey, Flavio Lehner, Angeline G Pendergrass, Sebastian Sippel, Joel Zeder, and Reto Knutti. Storylines for unprecedented heatwaves based on ensemble boosting. *Nature Communications*, 14(1):4643, 2023.

Vera Melinda Galfi and Valerio Lucarini. Fingerprinting heatwaves and cold spells and assessing their response to climate change using large deviation theory. *Physical review letters*, 127(5): 058701, 2021.

Vera Melinda Gálfi, Valerio Lucarini, Francesco Ragone, and Jeroen Wouters. Applications of large deviation theory in geophysical fluid dynamics and climate science. *La Rivista del Nuovo Cimento*, 44(6):291–363, 2021.

Zengchao Hao, Fanghua Hao, Youlong Xia, Sifang Feng, Cheng Sun, Xuan Zhang, Yongshuo Fu, Ying Hao, Yu Zhang, and Yu Meng. Compound droughts and hot extremes: Characteristics, drivers, changes, and impacts. *Earth-Science Reviews*, 235:104241, 2022.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Radley M Horton, Justin S Mankin, Corey Lesk, Ethan Coffel, and Colin Raymond. A review of recent advances in research on extreme heat events. *Current Climate Change Reports*, 2:242–259, 2016.

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.

Valerio Lucarini, Vera Melinda Galfi, Jacopo Riboldi, and Gabriele Messori. Typicality of the 2021 western north america summer heatwave. *Environmental Research Letters*, 18(1):015004, 2023.

Stefano Materia, Lluís Palma García, Chiem van Straaten, Sungmin O, Antonios Mamalakis, Leone Cavicchia, Dim Coumou, Paolo de Luca, Marlene Kretschmer, and Markus Donat. Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 15(6):e914, 2024.

Gabriele Messori, Rodrigo Caballero, Freddy Bouchet, Davide Faranda, Richard Grotjahn, Nili Harnik, Steve Jewson, Joaquim G Pinto, Gwendal Rivière, Tim Woollings, et al. An interdisciplinary approach to the study of extreme weather events: large-scale atmospheric controls and insights from dynamical systems theory and statistical mechanics. *Bulletin of the American Meteorological Society*, 99(5):ES81–ES85, 2018.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Sjoukje Y Philip, Sarah F Kew, Geert Jan Van Oldenborgh, Faron S Anslow, Sonia I Seneviratne, Robert Vautard, Dim Coumou, Kristie L Ebi, Julie Arrighi, Roop Singh, et al. Rapid attribution analysis of the extraordinary heat wave on the pacific coast of the us and canada in june 2021. *Earth System Dynamics*, 13(4):1689–1713, 2022.

Francesco Ragone and Freddy Bouchet. Rare event algorithm study of extreme warm summers and heatwaves over europe. *Geophysical Research Letters*, 48(12):e2020GL091197, 2021.

Francesco Ragone, Jeroen Wouters, and Freddy Bouchet. Computation of extreme heat waves in climate models using a large deviation algorithm. *Proceedings of the National Academy of Sciences*, 115(1):24–29, 2018. doi: 10.1073/pnas.1712645115.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Vikki Thompson, Alan T Kennedy-Asser, Emily Vosper, YT Eunice Lo, Chris Huntingford, Oliver Andrews, Matthew Collins, Gabrielle C Hegerl, and Dann Mitchell. The 2021 western north america heat wave among the most extreme events ever recorded globally. *Science Advances*, 8 (18):eabm6860, 2022.

Yogesh Verma, Markus Heinonen, and Vikas Garg. ClimODE: Climate and weather forecasting with physics-informed neural ODEs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xuY33XhEGR.

A Appendix

A.1 Experiments

Here, we provide more details on the dataset, model, and training procedure. Our final goal is to provide the climate specialist with sample heat wave and cold spell patterns in various regions of interest. To show the proof-of-concept application of the class-conditional generative models to this particular task we use a similar setup as the one in Verma et al. (2024), relying on a coarsened ERA5 dataset. We alter some of the configurations and focus on three regions which were prone to climate extremes in recent years (Fischer et al., 2023; Thompson et al., 2022; Philip et al., 2022; Ragone & Bouchet, 2021). Due to the low resolution, the target regions have been specified as *the entire world*, *Pacific Northwest* ($34°$**N**-$64°$**N** and $81°$**W**-$160°$**W**), *Chicago* ($28°$**N**-$52°$**N** and $74°$**W**-$100°$**W**), and *Europe* ($35°$**N**-$75°$**N** and $10°$**W**-$25°$**E**).

We use the ERA5 dataset (Hersbach et al., 2020) at a $5.625°$ resolution with 1-day intervals. We only use the temperature variable (`t`), normalizing it to the range $[0, 1]$ using min-max scaling over the data. The training data consists of ten years (2006–2015), with 2016 used for validation and 2017–2018 for testing. As the original ClimODE model used 6-hour intervals and 4 other variables (`t2m`, `z`, `u10`, `v10`), along the temperature and the context information from orography and land surface model, we train variations of ClimODE to serve as the baseline for a fair comparison with UQClimODE. In the DAILY variant we train the model with 1-day interval of the temperature variable and in the NOCONTEXT version we further remove the context embedding from the data.

UQClimODE models $\boldsymbol{v}_\theta$ with a UNet (Ronneberger et al., 2015) with 11 channels for the embedded features, 3 residual blocks with 0.2 dropout value. Each block is repeated (1, 2, 3, 4) interleaved with (16, 8) pixel token attention blocks in between. A smaller UNet with similar residual blocks that are repeated (1, 2) times is used to model μ_θ. For training we additionally use a one-layer classifier with piecewise linear non-linearity to get logits for the target classes and we minimize the combined negative log-likelihood and cross-entropy loss

$$\mathcal{L}(\theta, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \log \mathcal{N}\left(u_{\text{obs}}^{(i)}; u_\theta^{(i)}, \Sigma\right) + \lambda \log \text{softmax}_{y^{(i)}} \left(L\left(q\left(\left(u_\theta\right)_{\mathcal{X}}\right)\right)\right) \right\}.$$

Although UQClimODE is a sequential generative model, instead of training on previously predicted steps in the sequence, we input the ground truth from training dataset to the model during training, otherwise known as teacher forcing. This method is known to accelerate the training of sequential models (Lamb et al., 2016). We train the model for 100 epochs with $\lambda = 0.1$ and learning rate of 0.001. Σ is set to 0.01 through cross validation in $0.1, 0.05, 0.01, 0.005, 0.001$.

We evaluate the quality of the predicted temperature maps with latitude-weighted RMSE

$$\text{RMSE} = \frac{1}{N}\sum_{t}^{N}\sqrt{\frac{1}{HW}\sum_{h}\sum_{w}\alpha(h)(u_{\text{obs}} - u_\theta)^2}, \tag{5}$$

where $\alpha(h) = H\cos(h)/\sum_{h'}\cos(h')$, Anomaly Correlation Coefficient (ACC)

$$\text{ACC} = \frac{\sum \alpha(h)(u_{\text{obs}} - \bar{u}_{\text{obs}})(u_\theta - \bar{u}_\theta)}{\sqrt{\sum \alpha(h)(u_{\text{obs}} - \bar{u}_{\text{obs}})^2 \sum \alpha(h)(u_\theta - \bar{u}_\theta)^2}}, \tag{6}$$

for the averages $\bar{u}_{\text{obs}} = 1/N\sum_i u_{\text{obs}}$ and $\bar{u}_\theta = 1/N\sum_i u_\theta$ and Continuous Ranked Probability Score (CRPS).

A.2 Long-Term Generation Results

Tables 1, 2, and 3, show the results for CRPS, weighted RMSE, and ACC of sequential predictions with different lead times up to 90 days, in the setup explained in Section 3.1.

Table 2: RMSE comparison of predictions with ClimODE variants for global forecasting.

Model	Lead time (days)					
	1	3	7	14	30	90
ClimODE (24)	**1.58 ± 0.09**	**2.58 ± 0.16**	4.77 ± 0.37	NaN	NaN	NaN
daily (24)	2.04 ± 0.10	3.15 ± 0.21	4.50 ± 0.34	8.71 ± 0.76	NaN	NaN
daily (6)	2.55 ± 0.21	3.23 ± 0.21	**3.74 ± 0.31**	4.05 ± 0.37	6.10 ± 1.15	NaN
daily (4)	2.81 ± 0.41	3.32 ± 0.37	**3.74 ± 0.40**	**3.96 ± 0.42**	5.33 ± 1.05	NaN
nocontext (24)	3.33 ± 0.34	4.03 ± 0.31	4.90 ± 0.29	6.83 ± 0.72	13.21 ± 5.03	NaN
nocontext (6)	4.22 ± 0.24	4.61 ± 0.25	4.97 ± 0.20	5.85 ± 0.37	9.27 ± 1.08	NaN
UQClimODE	2.06 ± 0.11	3.26 ± 0.24	4.10 ± 0.36	4.38 ± 0.39	**4.38 ± 0.41**	**4.33 ± 0.31**

Table 3: ACC comparison of predictions with ClimODE variants for global forecasting.

Model	Lead time (days)					
	1	3	7	14	30	90
ClimODE (24)	**0.95 ± 0.02**	**0.86 ± 0.06**	0.52 ± 0.21	NaN	NaN	NaN
daily (24)	0.91 ± 0.04	0.76 ± 0.13	0.53 ± 0.25	0.15 ± 0.39	NaN	NaN
daily (6)	0.85 ± 0.08	0.73 ± 0.15	0.61 ± 0.23	0.54 ± 0.29	0.26 ± 0.39	NaN
daily (4)	0.82 ± 0.09	0.72 ± 0.15	**0.63 ± 0.22**	**0.57 ± 0.27**	0.33 ± 0.41	NaN
nocontext (24)	0.77 ± 0.10	0.61 ± 0.19	0.45 ± 0.24	0.25 ± 0.32	0.07 ± 0.41	NaN
nocontext (6)	0.69 ± 0.13	0.56 ± 0.20	0.47 ± 0.22	0.37 ± 0.22	0.22 ± 0.25	NaN
UQClimODE	0.91 ± 0.05	0.76 ± 0.13	0.61 ± 0.21	0.54 ± 0.24	**0.50 ± 0.24**	**0.49 ± 0.23**

A.3 Results of Steering Temperature Regimes

All of our results highlight that UQClimODE can successfully generate global and regional temperature maps in the specified temperature class, in particular the first ("cold") and last ("hot") quartiles. We provide additional visualizations of our results in the following. In all plots, blue and red lines indicate results for the first and last quartile, while dashed lines correspond to the true temperature class.

Global results: Figure 2 provides global temperature maps for five consecutive days in 2017. We show the ground truth map as well as the most likely predictions for the first ("cold") and last ("hot") quartiles. Notice, for instance, the difference between the cold and hot regimes and the ground truth in Europe. Figures 3 and 4 provide the average global temperature curves of several generated sequences for 2017 and 2018. The initial conditions after 30 consecutive days were used to start the generation. In Figures 5 and 6, we extend the period for which we want to generate to 90 days. We take initial conditions after this time period and show the results for the global average temperature of the most likely 90-day sequences. Figures 1 (in the main text) and 7 provide the result if we take

the initial conditions at the beginning of 2017 and 2018, respectively, and generate the most likely sequences for the full year. In Figures 5, 6, 1, 7, the shaded areas depict the different (ground truth) quartile classes.

Regional results for Europe, the Pacific Northwest, and the Chicago region: Similar to the global setting, we provide regional results for regions which exhibited heat waves in recent years. As before, we take the initial conditions at the beginning of 2017 and 2018, respectively, and show the most likely sequences for the full year in Figures 8 and 9 for Europe, 12 and 13 for the Pacific Northwest. We provide several generated sequences with initial conditions after 30 consecutive days for 2017 in Figures 10 for Europe, 14 for the Pacific Northwest, and 17 for the Chicago region. For the latter region, we provide the most likely results for the 90-day sequences for 2017 and 2018 in Figures 15 and 16.

Figure 11 shows the prediction of the most likely temperature maps for 5 consecutive days in the region of Pacific Northwest, with 10-day leads and under the cold and hot temperature regimes. It illustrates that heat waves are most likely going to impact the south and central part of the map while cold temperatures are most likely bounded to the northern areas.

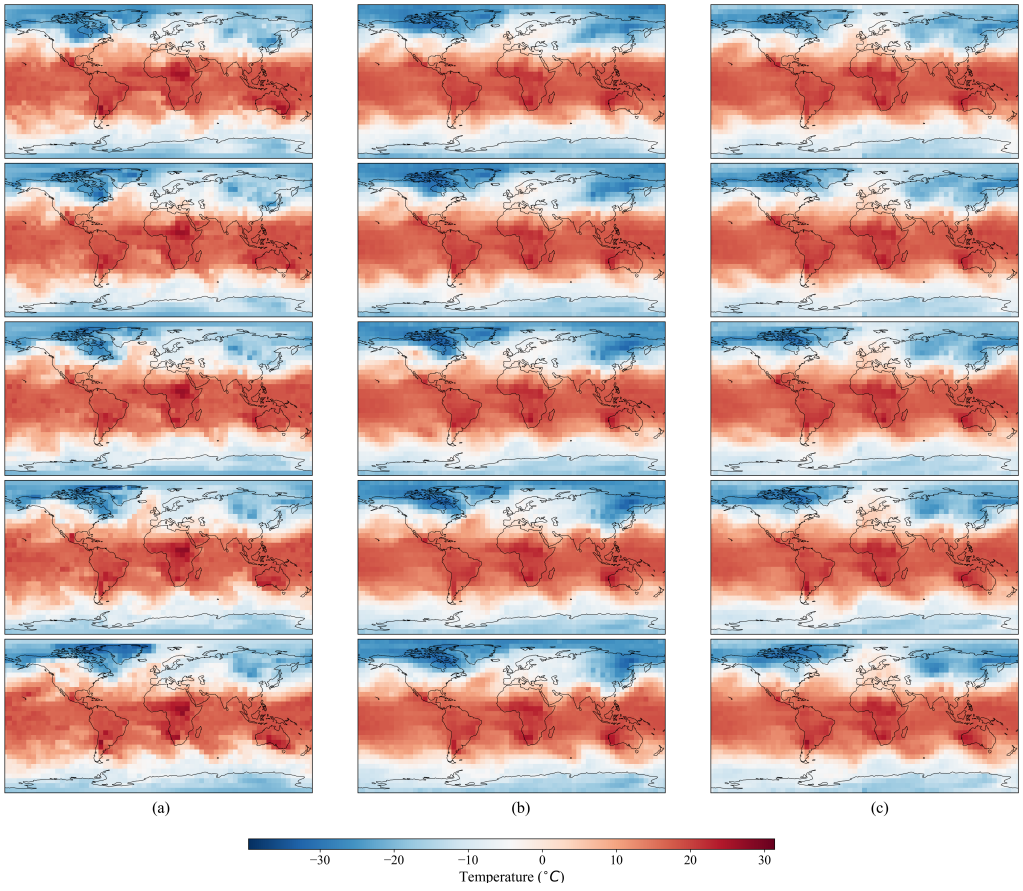

Figure 2: Prediction of the most likely temperature maps with 10-day lead for 5 consecutive days (globally for 2017). (a) Ground truth, (b) prediction from first quartile ("cold") temperature regime, (c) prediction from last quartile ("hot") temperature regime.

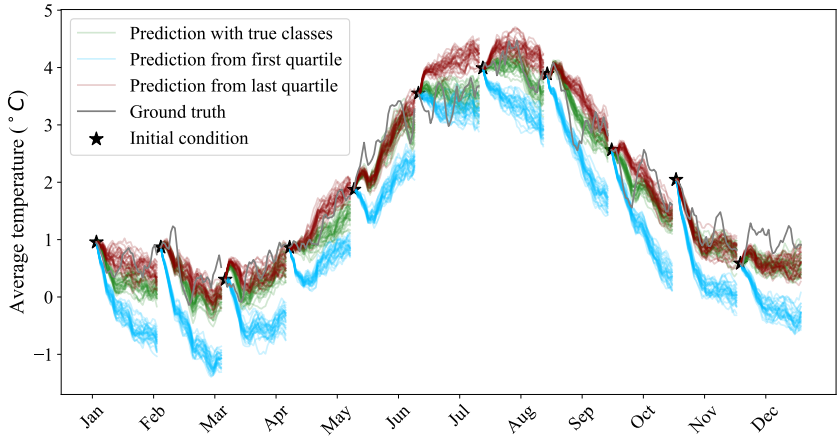

Figure 3: Average global temperature of 30 generated 30-day sequences conditioned on target temperature regimes (2017).

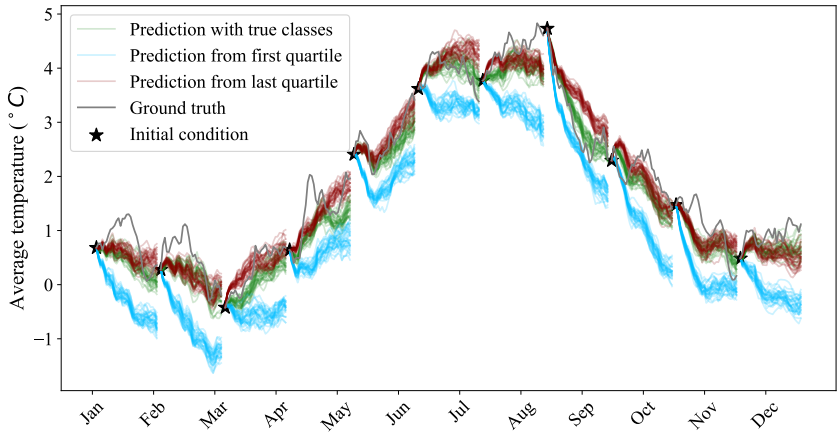

Figure 4: Average global temperature of generated 30-day sequences conditioned on target temperature regimes (2018).

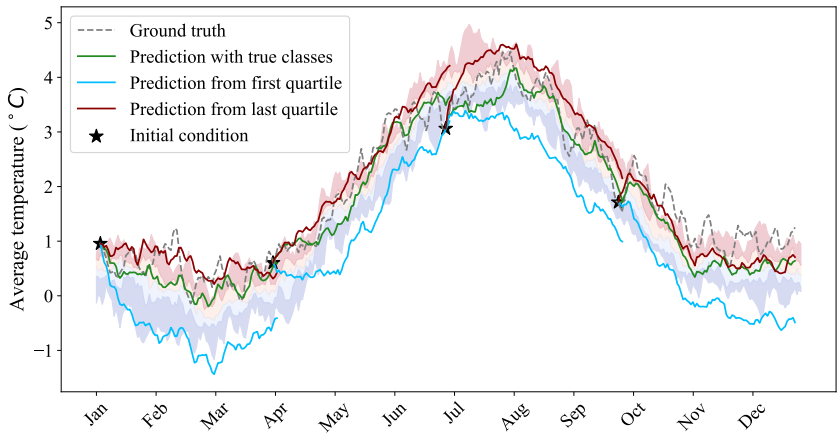

Figure 5: Average global predicted temperature for the most likely 90-day sequences conditioned on target temperature regimes, target classes are displayed in shades (2017).

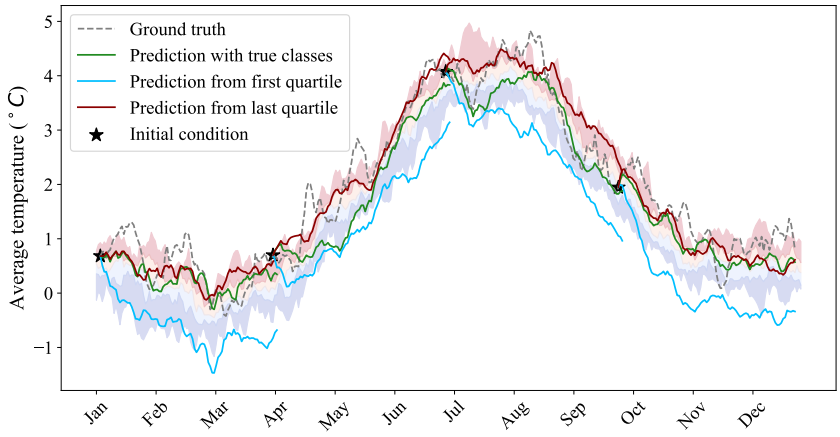

Figure 6: Average global predicted temperature for the most likely 90-day sequences conditioned on target temperature regimes, target classes are displayed in shades (2018).

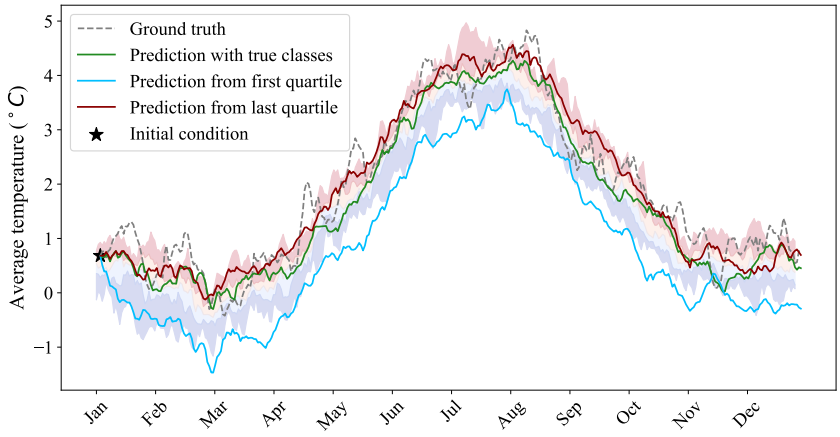

Figure 7: Average global predicted temperature for the most likely 365-day sequences conditioned on target temperature regimes, target classes are displayed in shades (2018).

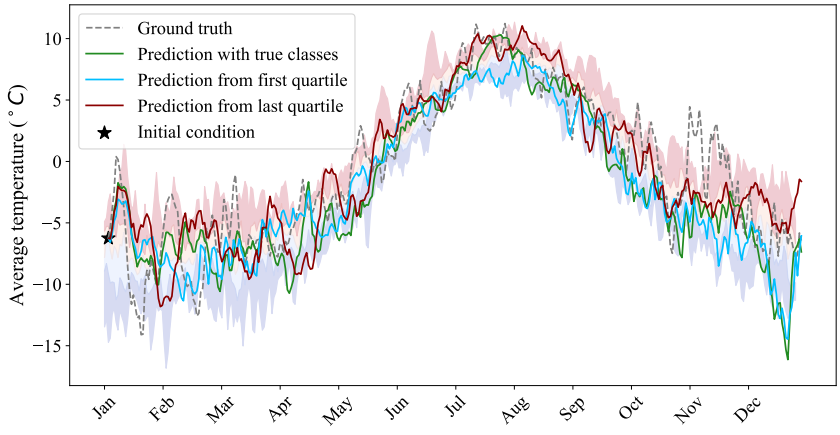

Figure 8: Average predicted temperature over Europe for the most likely 365-day sequences conditioned on target temperature regimes (2017).

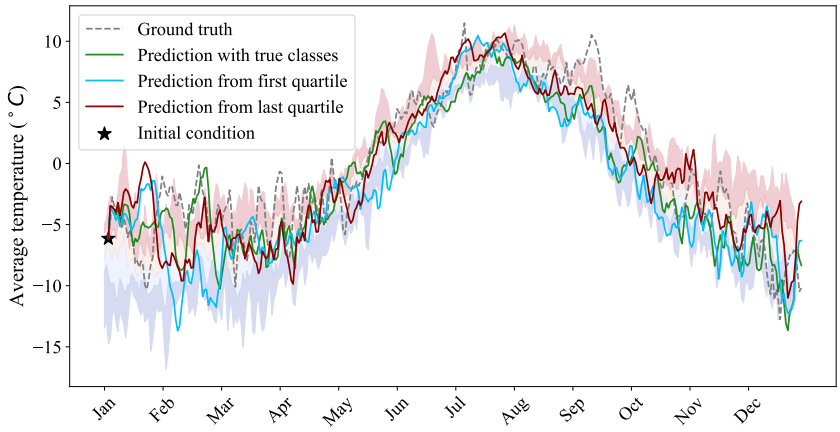

Figure 9: Average predicted temperature over Europe for the most likely 365-day sequences conditioned on target temperature regimes (2018).

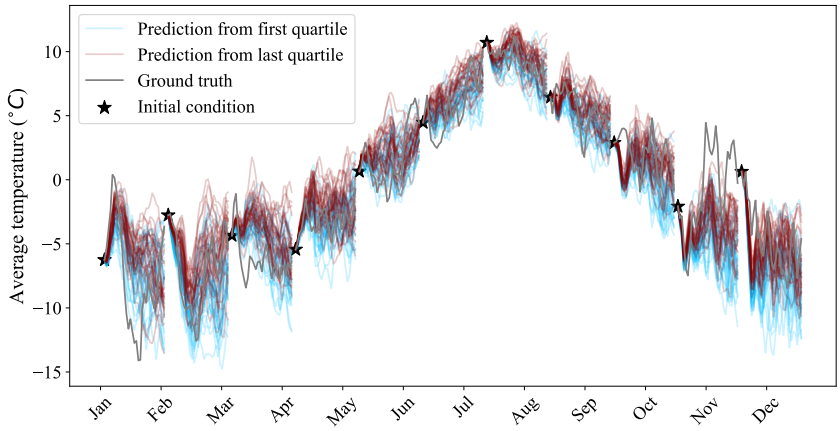

Figure 10: Average temperature over Europe of 30 generated 30-day sequences conditioned on target temperature regimes (2017).

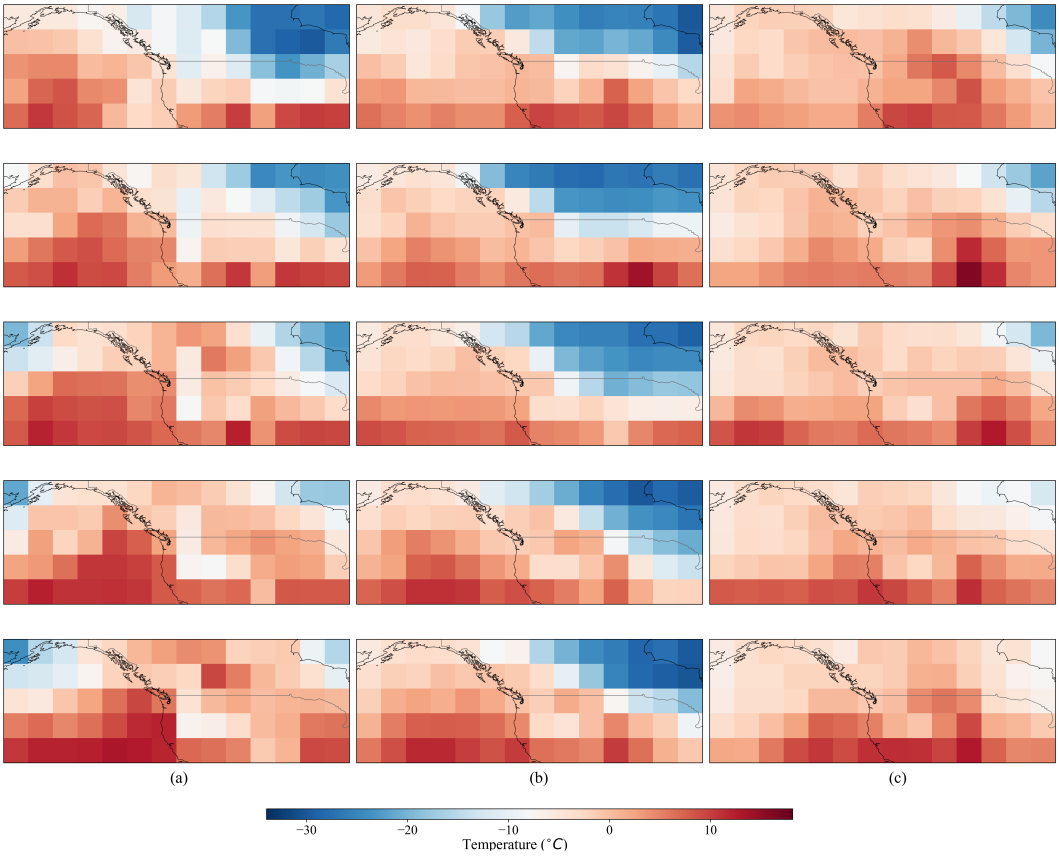

Figure 11: Prediction of the most likely temperature maps with 10-day leads for 5 consecutive days in the Pacific Northwest region (2017). (a) Ground truth, (b) prediction from first quartile ("cold") temperature regime, (c) prediction from last quartile ("hot") temperature regime.

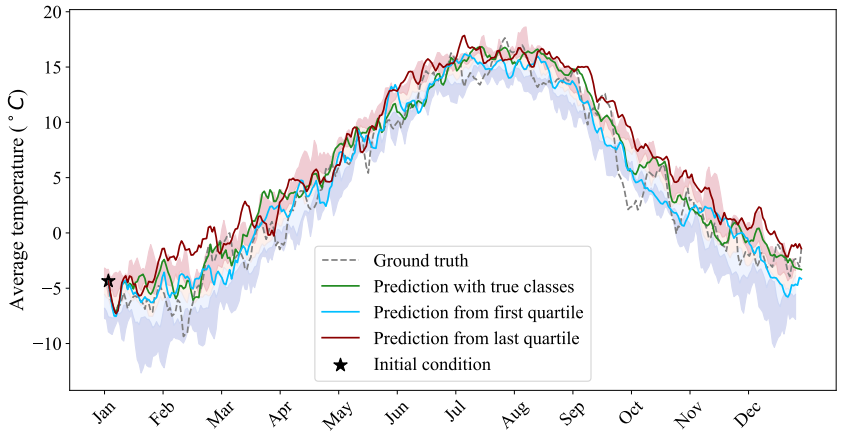

Figure 12: Average predicted temperature over Pacific Northwest for the most likely 365-day sequences conditioned on target temperature regimes (2017).

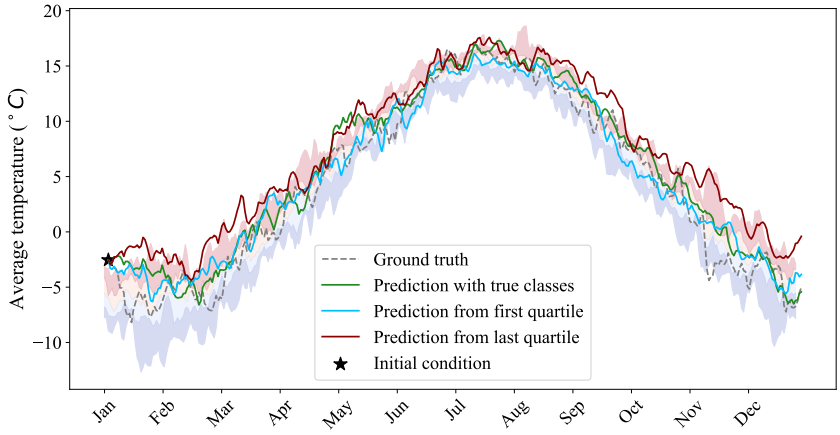

Figure 13: Average predicted temperature over Pacific Northwest for the most likely 365-day sequences conditioned on target temperature regimes (2018).

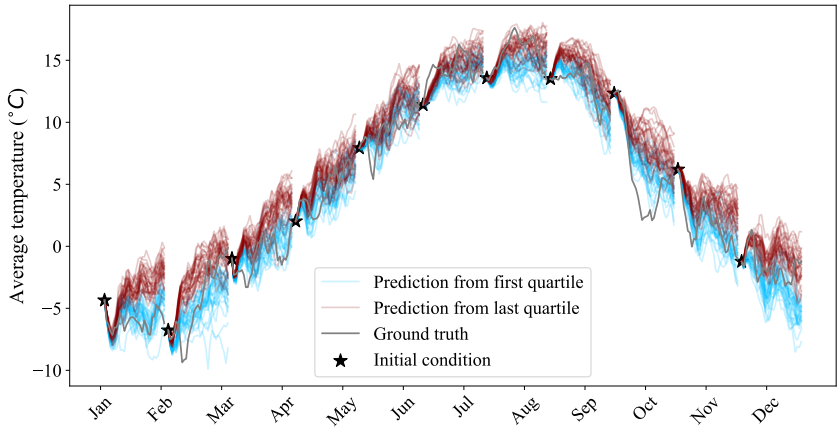

Figure 14: Average temperature over Pacific Northwest of 30 generated 30-day sequences conditioned on target temperature regimes (2017).

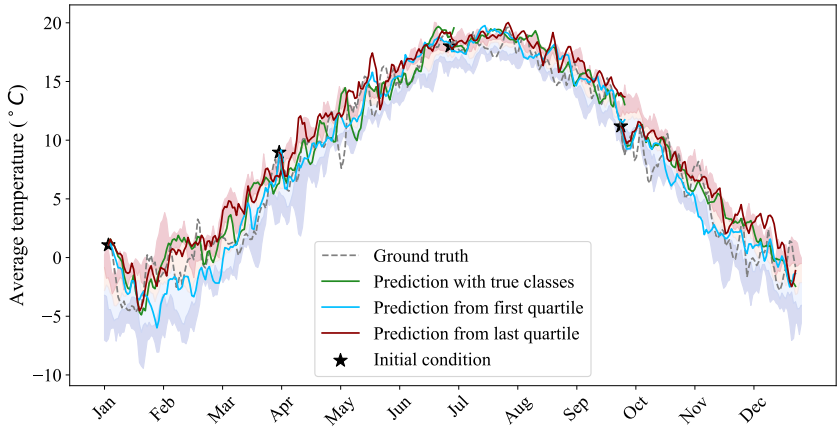

Figure 15: Average predicted temperature over the Chicago region for the most likely 90-day sequences conditioned on target temperature regimes (2017).

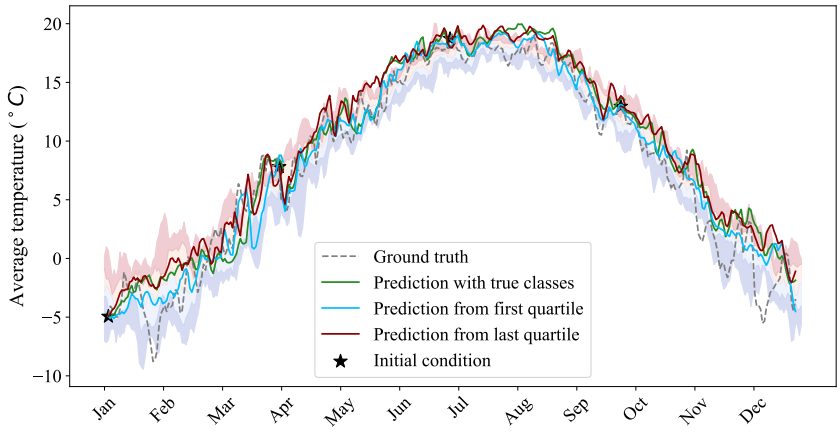

Figure 16: Average predicted temperature over the Chicago region for the most likely 90-day sequences conditioned on target temperature regimes (2018).

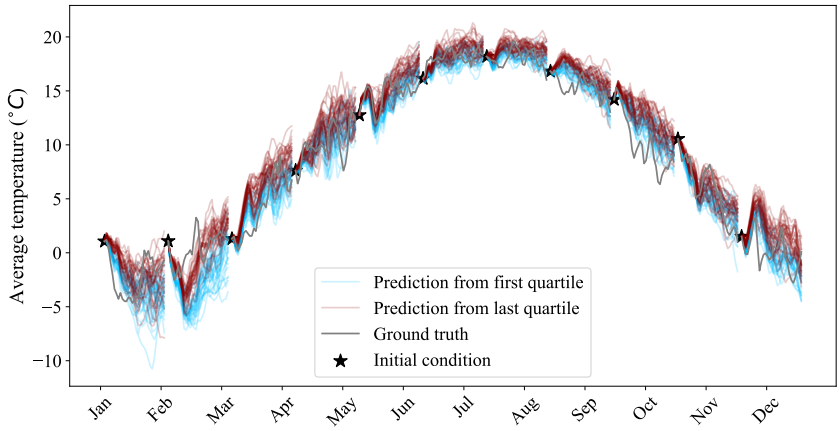

Figure 17: Average temperature over the Chicago region of 30 generated 30-day sequences conditioned on target temperature regimes (2017).