Learning to infer weather states using partial observations

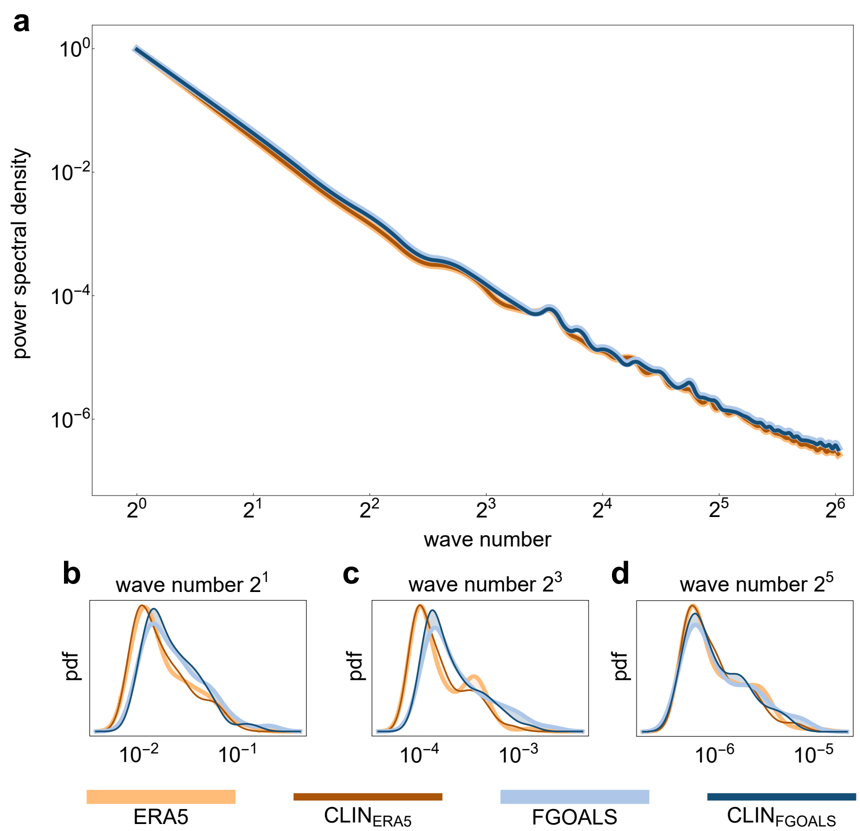
Jie Chao[1], Baoxiang Pan[2], Quanliang Chen[3], Shangshang Yang[4], Jingnan Wang[5], congyi nai[6], Yue Zheng[7], Xichen Li[8], Huiling Yuan[9], Xi Chen[2], Bo Lu[10], and Ziniu Xiao[2]

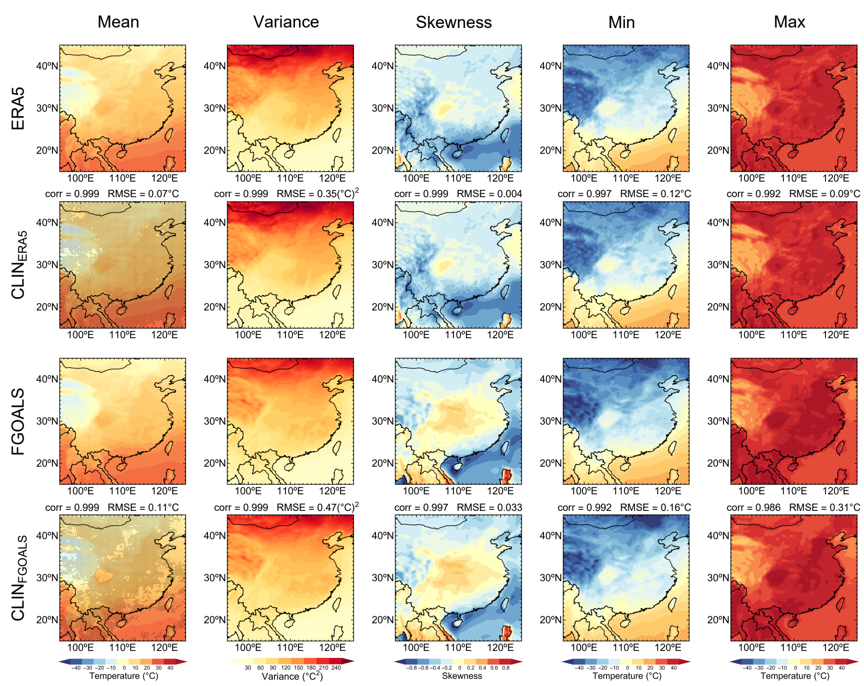
[1]Chengdu University of Information Technology
[2]Institute of Atmospheric Physics, Chinese Academy of Sciences
[3]Plateau Atmosphere and Environment Key Laboratory of Sichuan Province, College of Atmospheric Science,Chengdu University of Information Technology
[4]Key Laboratory of Mesoscale Severe Weather Ministry of Education/School of Atmospheric Sciences, Nanjing University
[5]College of Computer, National University of Defense Technology
[6]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
[7]ClusterTech Limited, Hong Kong
[8]International Center for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences
[9]Key Laboratory of Mesoscale Severe Weather/Ministry of Education, and School of Atmospheric Sciences, Nanjing University, Nanjing, China
[10]National Climate Center, China Meterological Administration

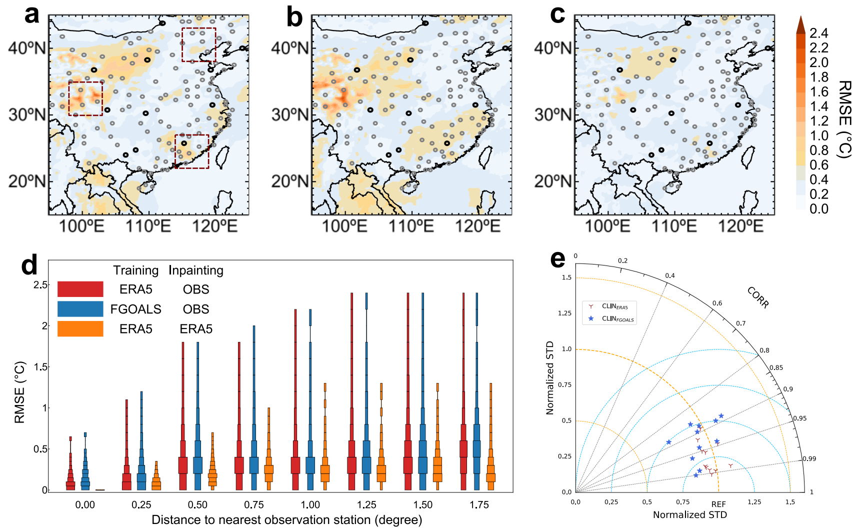
April 19, 2024

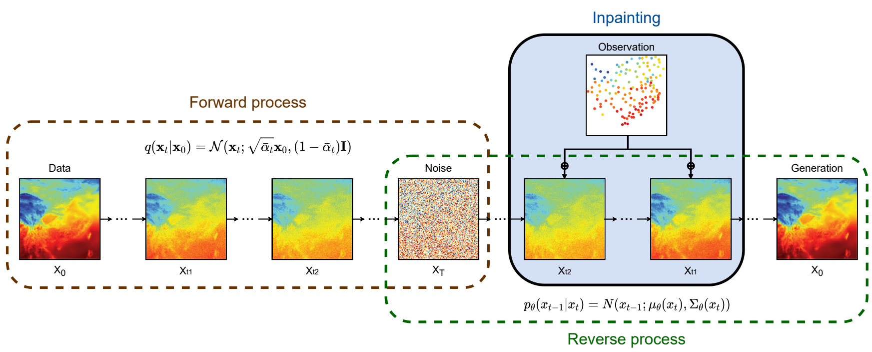
Abstract

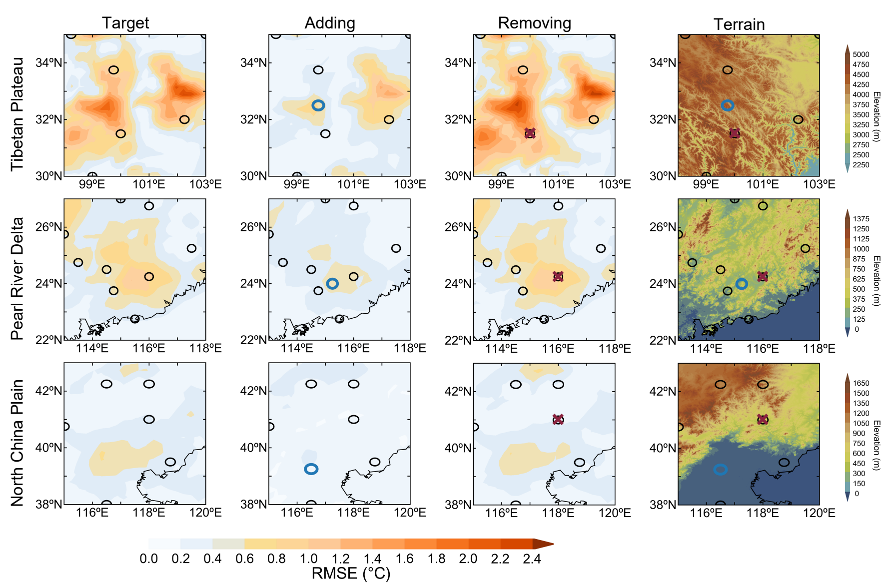
Accurate state estimation of the high-dimensional, chaotic Earth atmosphere marks a Sisyphean task, yet is indispensable for initiating weather forecast and gauging climate variability. While much effort is devoted to assimilating observations and forecasts to infer weather state, the inherent low-dimensional statistical structure in atmospheric circulation, shaped by geophysical laws and geographic boundaries, is underutilized as informative prior for state inference, or as reference for assessing representative of existing observations and planning new ones. We realize these potential by learning climatological distribution from climate reanalysis/simulation, using deep generative model. For a case study of estimating 2 m temperature spatial patterns, the learned distribution faithfully reproduces climatology statistics. A combination of the learned climatological prior with few station observations yields strong posterior of spatial pattern estimates, which are spatially coherent, faithful and adaptive to observation constraints, and uncertainty-aware. This allows us to evaluate each observation's value in reducing state estimation uncertainty, and guide optimal observation network design by pinpointing the most informative sites. Our study showcases how generative models can extract and utilize information produced in the chaotic evolution of climate system.


a
power spectral density
wave number
b
wave number 2^1
c
wave number 2^3
d
wave number 2^5
pdf
ERA5
CLIN$_{ERA5}$
FGOALS
CLIN$_{FGOALS}$


Mean
Variance
Skewness
Min
Max
ERA5
CLIN$_{ERA5}$
FGOALS
CLIN$_{FGOALS}$
corr = 0.999 RMSE = 0.07°C
corr = 0.999 RMSE = 0.35(°C)^2
corr = 0.999 RMSE = 0.004
corr = 0.997 RMSE = 0.12°C
corr = 0.992 RMSE = 0.09°C
corr = 0.999 RMSE = 0.11°C
corr = 0.999 RMSE = 0.47(°C)^2
corr = 0.997 RMSE = 0.033
corr = 0.992 RMSE = 0.16°C
corr = 0.986 RMSE = 0.31°C
Temperature (°C)
Variance (°C^2)
Skewness
Temperature (°C)
Temperature (°C)


a
b
c
40ºN
30ºN
20ºN
100ºE
110ºE
120ºE
RMSE (°C)
d
Training
Inpainting
ERA5
OBS
FGOALS
OBS
ERA5
ERA5
RMSE (°C)
Distance to nearest observation station (degree)
e
CORR
Normalized STD
REF


Inpainting
Observation
Forward process
$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$
Data
Noise
Generation
X_0
X_{t1}
X_{t2}
X_T
X_{t2}
X_{t1}
X_0
$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t), \Sigma_\theta(x_t))$
Reverse process


Target
Adding
Removing
Terrain
Tibetan Plateau
Pearl River Delta
North China Plain
Elevation (m)
RMSE (°C)

Learning to infer weather states using partial observations

Jie Chao[1,2], Baoxiang Pan[2], Quanliang Chen[1], Shangshang Yang[2,3], Jingnan Wang[2,4], Congyi Nai[2,5], Yue Zheng[6], Xichen Li[2], Huiling Yuan[3], Xi Chen[2], Bo Lu[7], Ziniu Xiao[2]

[1]School of Atmospheric Sciences, Chengdu University of Information Technology, Sichuan, China
[2]Institute of Atmospheric Physics, Chinese Academy of Science, Beijing, China
[3]Key Laboratory of Mesoscale Severe Weather, Ministry of Education, and School of Atmospheric Sciences, Nanjing University, Jiangsu, China
[4]College of Computer, National University of Defense Technology, Hunan, China
[5]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[6]Clustertech LTD, Hong Kong, China
[7]National Climate Center, China Meteorological Administration, Beijing, China

Key Points:

- Deep generative model enables accurate spatial interpolation of weather variables from sparse observations.
- The model generates probabilistic weather estimates with reliable uncertainty quantification by combining learned priors and observations.
- The model quantifies the value of observations for reducing uncertainty, guiding optimal observation network design.

Corresponding author: Baoxiang Pan, panbaoxiang@lasg.iap.ac.cn

multi-scale background information using forecasting models requires operational run of large ensemble high-resolution numerical simulations, which is prohibitively expensive and burdensome (Toth et al., 2003; Palmer, 2017).

Is there extra information source for inferring the state of the high-dimensional, chaotic Earth atmosphere? It turns out that, the inherent low-dimensional statistical structure in atmospheric circulation, shaped by the underlying geophysical laws and quasi-static geographic boundaries, can serve as an informative prior for state inference. The Earth climate system, like any other chaotic system, is an information producer: it gradually reveals the characteristic structure of its phase space at ever- ner scales (Gilpin, 2024). By identifying and parameterizing this characteristic structure, we can potentially bypass the curse of high dimensionality, and make more e cient use of limited observations for the state inference task.

Some pioneering works have explored this direction, leveraging the inherent structure of climate data to ll in missing observations and rebuild historical climate records. For instance, Kadow et al. (2020) developed a partial convolution method to reconstruct historical global temperature patterns based on partial observations and climate simulation. Kanngie er and Fiedler (2024) applied a similar methodology to restore the spatial extent of dust plumes in cloud-masked satellite images. Most of these practices consider deterministic models, which are designed for speci c \reconstruction" problem con gurations, yielding deterministic results regardless of whether observations can adequately constrain the estimation uncertainty. As a result, these methodologies generalize poorly to state inference tasks where the number or layout of observations change, fail to reproduce extremes or apply for scenarios where only limited observations are available.

A solution to these dilemmas is to shift from deterministic model to probabilistic model (B. Pan et al., 2021). Speci cally, we prefer to build a probabilistic model that explicitly represents the inherent statistical structure of the atmosphere as revealed by climate observations or simulations. Thereafter, we hope to e ectively and e ciently combine the learned climatological prior with incomplete observations, so as to obtain strong posterior of spatial pattern estimates. This problem setup poses two stringent requirements on the underlying probabilistic model. First, the model must faithfully approximate the high-dimensional climatological distribution as generated by the chaotic evolution of climate dynamics. Second, the model must enable exible probabilistic inference, allowing us to e ciently obtain posterior atmospheric state estimates given arbitrary observational constraints.

To ful ll these requirements, we resort to generative machine learning, in particular, probabilistic di usion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song, Sohl-Dickstein, et al., 2020; Kingma et al., 2021). Probabilistic di usion models learn to approximate complex, high-dimensional probability distributions in an iterative manner, achieving unprecedented tting capacity and controlling exibility (B. Pan et al., 2023; Nai et al., 2024). To demonstrate the idea, we consider a case example of inferring the spatial pattern of 2 m temperature based on sparse observations from operational meteorology stations. We learn probabilistic di usion models to approximate the climatological distribution of 2 m temperature spatial patterns from climate reanalysis or simulation data. After carefully assessing the model's ability to reproduce climatology, we develop tools to \inpaint" arbitrary observation constraints into the sample generation process, yielding probabilistic 2 m temperature spatial pattern estimates. Finally, we ap-

Deficiencies in observation render it an ill-posed task to estimate the state of the high-dimensional Earth atmosphere, calling for strong prior to achieve feasible solution. Forecasts from previous time steps are frequently applied to serve this mission, carrying information from previous step observations to the current step via a process-based model (Wang et al., 2000). As a result, the state estimation accuracy depends on an intricate interplay among model biases, background uncertainty, and observation error, which cannot be effectively disentangled or controlled (Law et al., 2015). Moreover, to provide multi-scale background information using forecasting models requires operational run of large ensemble high-resolution numerical simulations, which is prohibitively expensive and burdensome (Toth et al., 2003; Palmer, 2017).

Is there extra information source for inferring the state of the high-dimensional, chaotic Earth atmosphere? It turns out that, the inherent low-dimensional statistical structure in atmospheric circulation, shaped by the underlying geophysical laws and quasi-static geographic boundaries, can serve as an informative prior for state inference. The Earth climate system, like any other chaotic system, is an information producer: it gradually reveals the characteristic structure of its phase space at ever-finer scales (Gilpin, 2024). By identifying and parameterizing this characteristic structure, we can potentially bypass the curse of high dimensionality, and make more efficient use of limited observations for the state inference task.

Some pioneering works have explored this direction, leveraging the inherent structure of climate data to fill in missing observations and rebuild historical climate records. For instance, Kadow et al. (2020) developed a partial convolution method to reconstruct historical global temperature patterns based on partial observations and climate simulation. Kanngießer and Fiedler (2024) applied a similar methodology to restore the spatial extent of dust plumes in cloud-masked satellite images. Most of these practices consider deterministic models, which are designed for specific “reconstruction” problem configurations, yielding deterministic results regardless of whether observations can adequately constrain the estimation uncertainty. As a result, these methodologies generalize poorly to state inference tasks where the number or layout of observations change, fail to reproduce extremes or apply for scenarios where only limited observations are available.

A solution to these dilemmas is to shift from deterministic model to probabilistic model (B. Pan et al., 2021). Specifically, we prefer to build a probabilistic model that explicitly represents the inherent statistical structure of the atmosphere as revealed by climate observations or simulations. Thereafter, we hope to effectively and efficiently combine the learned climatological prior with incomplete observations, so as to obtain strong posterior of spatial pattern estimates. This problem setup poses two stringent requirements on the underlying probabilistic model. First, the model must faithfully approximate the high-dimensional climatological distribution as generated by the chaotic evolution of climate dynamics. Second, the model must enable flexible probabilistic inference, allowing us to efficiently obtain posterior atmospheric state estimates given arbitrary observational constraints.

To fulfill these requirements, we resort to generative machine learning, in particular, probabilistic diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song, Sohl-Dickstein, et al., 2020; Kingma et al., 2021). Probabilistic diffusion models learn to approximate complex, high-dimensional probability distributions in an iterative manner, achieving unprecedented fitting capacity and controlling flexibility (B. Pan et al., 2023; Nai et al., 2024). To demonstrate the idea, we consider a case example of inferring the spatial pattern of 2 m temperature based on sparse observations from operational meteorology stations. We learn probabilistic diffusion models to approximate the climatological distribution of 2 m temperature spatial patterns from climate reanalysis or simulation data. After carefully assessing the model’s ability to reproduce climatology, we develop tools to “inpaint” arbitrary observation constraints into the sample generation process, yielding probabilistic 2 m temperature spatial pattern estimates. Finally, we apply this methodology to evaluate each observation's value in reducing state estimation uncertainty, and guide optimal observation network design by pinpointing the most informative sites.

2 Methodology

2.1 Data and problem setup

We consider the task of inferring the spatial pattern of 2 m temperature over East Asia ($15°\text{N} - 45°\text{N}, 95°\text{E} - 125°\text{E}$), using station observations covering $\sim 1\%$ grids of the considered region. To achieve this, we learn climatological distribution of 2 m temperature spatial pattern using climate reanalysis or simulation data. The reanalysis data are hourly, 0.25° 2 m temperature data from the fifth-generation global climate and weather reanalysis (ERA5) developed at European Centre for Medium-Range Weather Forecasts (Hersbach et al., 2020, ECMWF). The simulation data are 3-hourly, 0.25° 2 m temperature historical simulation from the Flexible Global Ocean-Atmosphere-Land System Model version f3-H (Bao et al., 2020, FGOALS-f3-H), which participates in the sixth phase of the Coupled Model Intercomparison Project (Eyring et al., 2016, CMIP6). The station observation data are obtained from the Chinese National Climatic Data Center (X. Pan et al., 2021).

Formally, we denote the spatial pattern of 2 m temperature for the target region as $\mathbf{x}$, which is a 120×120 dimensional random variable here. Our objective is to approximate the distribution of $\mathbf{x}$, based on large number of samples from climate reanalysis or simulation:

$$p_{\theta^*} = \arg\max_{p_\theta} \sum \log p_\theta(\mathbf{x}) \tag{1}$$

Here p_θ is parameterized probability density function approximator, θ^* is the optimal parameter, optimized by maximizing the overall likelihood of p_θ assigned to the training samples.

Given p_{θ^*} and sparse observations, we need to provide probabilistic estimates of 2 m temperature spatial patterns, i.e., $p_{\theta^*}(\mathbf{x}|\mathbf{x}\odot\mathbf{m})$. Here, $\odot$ is dot product, $\mathbf{m}$ is observation mask, with value 1/0 denoting the existence/absence of observations for each geogrid. $p_{\theta^*}(\mathbf{x}|\mathbf{x}\odot\mathbf{m})$ should yield samples that are spatially coherent and faithful to observational constraints. Also, $p_{\theta^*}(\mathbf{x}|\mathbf{x}\odot\mathbf{m})$ should offer accurate uncertainty quantification. For instance, geogrids close to observation stations should typically have low state estimate uncertainties, while distant ones have high uncertainties. Finally, we prefer $p_{\theta^*}(\mathbf{x}|\mathbf{x}\odot\mathbf{m})$ to be adaptive to changes in observation configurations, such as the abortion or inclusion of observation stations, or rearrangement of station network layout. Below we illustrate how to achieve these requirements using the proposed methodology.

2.2 Learning climatology with probabilistic diffusion model

We elucidate how to learn climatological distribution of the target random variable using probabilistic diffusion model, thereafter leverage this learned prior for the inference task (Sec. 2.3). For clarity, we only cover key steps necessary for establishing our methodology. Details can be found in the literature referenced through the description.

To approximate a target distribution using probabilistic diffusion model, we train a series of deep neural networks that can be chained to establish bijective mapping between the target distribution and a prior distribution (Sohl-Dickstein et al., 2015; Ho et al., 2020). Specifically, we define the following Gaussian process:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{(1-\beta_t)}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

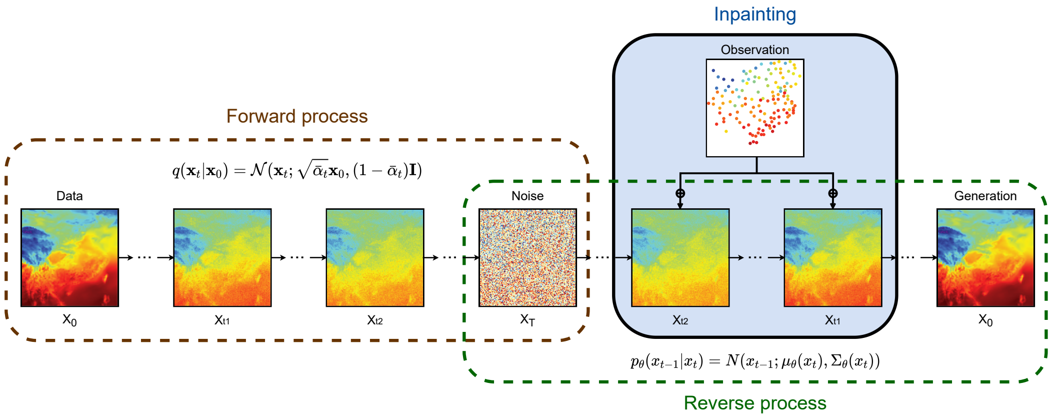


Figure 1. Overview of the Climate Inpainting (CLIN) methodology. A pre-defined forward Gaussian process (left) turns distribution of target climate variable into a prior distribution, i.e., standard Gaussian. A learned reverse Gaussian process (right) turns the prior distribution into the distribution of the target climate variable. We "inpaint" sparse observations throughout the reverse Gaussian process (right top), so as to obtain spatial pattern estimates of the target variable.

Here $p(\mathbf{x}_0) = p(\mathbf{x})$, which is the target distribution; $p(\mathbf{x}_T)$ is the prior distribution; we bridge $\mathbf{x}_0$ and $\mathbf{x}_T$ using $\mathbf{x}_{t\in[1,T]}$, which are latent variables with increasing noise level; $\mathcal{N}$ is Gaussian distribution; $\mathbf{I}$ is identity matrix; β_t is diffusion coefficient, which is pre-defined so that, give large enough T, $p(\mathbf{x}_T|\mathbf{x}_0)$ is drawn close to $p(\mathbf{x}_T)$, which is $\mathbf{x}_0$ agnostic. This setup offers analytical solution for $p(\mathbf{x}_{t+\tau}|\mathbf{x}_t), \forall\tau \in [0, T-t], t \in [0, T]$, facilitating convenient inference as detailed in Sec. 2.3.

To achieve generative modeling, we reverse Eq. 2 using the following variation distributions:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta, \mathbf{\Sigma}_\theta) \tag{3}$$

Here $\mathbf{\Sigma}_\theta$ is represented as an interpolation between its analytical lower and upper bound (Dhariwal & Nichol, 2021); μ_θ can be optimized by maximizing the variational lower bound (ELBO) on the log-likelihood of the training samples (Sohl-Dickstein et al., 2015; Kingma et al., 2021). In practice, we represent μ_θ as function of neural network parameterization for $\nabla p(\mathbf{x}_t|\mathbf{x}_0)$, which is known as the *score function* (Song, Garg, et al., 2020; Song, Sohl-Dickstein, et al., 2020). This simplifies the ELBO objective function to the following form:

$$L = \mathbb{E}_{t\in[1,T],\mathbf{x}_0\sim p(\mathbf{x}_0)}||\nabla p(\mathbf{x}_t|\mathbf{x}_0) - \epsilon_\theta||^2 \tag{4}$$

Here ϵ_θ is a neural network parameterization for $\nabla p(\mathbf{x}_t|\mathbf{x}_0)$. Given the trained score estimates, we can derive $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta, \Sigma_\theta)$ and sample it, starting with $p(\mathbf{x}_T)$, ending with $p(\mathbf{x}_0)$.

2.3 CLIN: inferring weather states using partial observations

We combine the learned climatology prior with station observations to infer the posterior probability distribution of the target variable, using a *repainting* methodology (Lugmayr et al., 2022; Zhang et al., 2023). Specifically, given a pre-trained diffusion model that sequentially applies $p_{\theta^*}(\mathbf{x}_t|\mathbf{x}_{t+1}) = \mathcal{N}(\mathbf{x}_t; \mu_{\theta^*}, \Sigma_{\theta^*})$ to transform $p(\mathbf{x}_T)$ to $p(\mathbf{x}_0)$, within a pre-selected time window of Ω, for grid points where we have observations, we replace values of $\mathbf{x}_t$ with observations noisified to time step t, by sampling $p(\mathbf{x}_t \odot \mathbf{m}|\mathbf{x}_0 \odot \mathbf{m})$. This replacement does not consider the generated parts of $\mathbf{x}_t$, therefore, the observations could not explicitly constrain the variability of unobserved parts.

To address this issue, for any $t \in \Omega$, after the replacement, instead of progressing to $t-1$ directly, we rewind to time step $t-\tau$ by sampling $p(\mathbf{x}_{t-\tau}|\mathbf{x}_t)$. We thereafter repeat the denoising steps from $t-\tau$ to t for k rounds, and carry out observation replacement for $\mathbf{x}_t$ at each round. This allows us to jointly modify both observed and unobserved regions throughout the denoising steps, yielding generated samples that are spatially coherent, faithful and adaptive to observation constraints, and uncertainty-aware. This methodology is referred to as *inpainting*, we hence name our methodology as CLIN, short for Climate Inpainting. A formal algorithm description is given below. Details for data processing, neural network architecture, hyperparameters for training and inference, are given in Supporting Information.

Algorithm 1 CLIN

Require: trained diffusion model p_{θ^*}, observations $\mathbf{x}_0 \odot \mathbf{m}$, repainting time step set Ω, rewinding step τ, rewinding round K

Ensure: observation constrained, spatially coherent sample $\mathbf{x}_0$

```
Initialize x_T ~ N(0, I)
for t = T − 1, ..., 1 do
    x_t ~ p_θ*(x_t | x_{t+1})                          ▷ Reverse sampling
    if t ∈ Ω then:
        for k = 1, ..., K do
            x_t^obser ~ p(x_t ⊙ m | x_0 ⊙ m)
            x_t ← x_t ⊙ (I − m) + x_t^obser            ▷ Condition on observations
            x_{t+τ} ~ p(x_{t+τ} | x_t)                 ▷ Rewind in time by τ steps
            for i = t + τ − 1, ..., t do
                x_i ~ p_θ*(x_i | x_{i+1})              ▷ Reverse sampling within a rewinding round
            end for
        end for
    end if
end for
return x_0
```

3 Results

The accuracy for state estimation depends on 1) how well we can approximate the climatological distribution, and 2) based on a learned climatological prior, how well we can combine it with limited observations to obtain probabilistic state estimates. Below we assess model's performance for these two aspects (Sec. 3.1 and 3.2). We further employ the model to quantify the extent to which observations reduce uncertainty in state estimation, offering insights for optimal observation design (Sec. 3.3).

3.1 Climatology

We compare grid-scale and field-scale statistics of 10,000 reference/generated samples to evaluate how well the probabilistic diffusion models reproduce their training data's climatology. Two models trained with climate reanalysis (ERA5) and historical climate simulation (FGOALS) data, hereafter referred to as $\text{CLIN}_{\text{ERA5}}$ and $\text{CLIN}_{\text{FGOALS}}$, are deployed and evaluated.

The grid-scale assessment considers the mean, variance, skewness, minimum, and maximum of climatological distribution at each grid (Fig. 2). These statistics from ERA5 (Fig. 2 Row 1) and FGOALS (Fig. 2 Row 3) generally agree well, due to shared constraints from geophysical laws and geographic boundaries. The key spatial patterns are the latitudinal gradient, the influence of topography (e.g., the Tibetan Plateau), and the land-sea contrast, which are most evident in the mean, minimum and maximum maps. The variance and skewness maps reveal more regional variations. A notable discrepancy is that, compared to ERA5, FGOALS tends to hold larger skewness for most of the land regions in Southern China and Philippine Island, implying a more frequent present of high 2 m temperature for these regions.

$\mathrm{CLIN}_{\mathrm{ERA5}}$ (Fig. 2 Row 2) and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ (Fig. 2 Row 4) can well reproduce the considered statistics of their training data, achieving high spatial correlation coefficient (~ 0.99) and low root mean squared error ($\sim 0.1°\mathrm{C}$) in matching these statistics. Besides reproducing the large scale patterns, both models accurately capture high frequency local variations influenced by complex topography, such as for mountainous regions and coastal areas. Also, the climatology difference between ERA5 and FGOALS are well reproduced by the corresponding CLIN models.

We further carry out grid-wise Kolmogorov-Smirnov tests to assess whether the generated and referential samples are likely to have come from the same underlying distribution: 96/76% grid points (stippled grids in Fig. 2) within the considered region pass a 95% confidence interval test for the $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ model. These results suggest that the CLIN model can well reproduce climatological distribution of its training data at grid scale.

We hereafter compare the referential and generated distributions using field-scale statistics. We first examine the linear spatial structure of the 2 m temperature spatial patterns using a principal component analysis (Supporting Information Fig. S2): we decompose the spatial pattern of the target random variable into a set of orthogonal modes that capture the maximum amount of variance, and compare the spatial modes (Empirical Orthogonal Functions, EOFs), as well as the variance explained by these modes. For EAR5, the first to third leading principal components explained 90/2.7/2.0% of the total variance. While for $\mathrm{CLIN}_{\mathrm{ERA5}}$, the first to third leading principal components explained 91/2.6/1.5% of the total variance, which closely matches results for the ERA5 referential data. More importantly, we obtain spatial correlation coefficient of 0.994/0.990/0.986 between the first to third EOF of EAR5 and $\mathrm{CLIN}_{\mathrm{ERA5}}$. While the spatial modes of FGOALS differs considerably with ERA5, $\mathrm{CLIN}_{\mathrm{FGOALS}}$ closely matches FGOALS: the first to third leading principal components explained 83.6/5.1/2.2% or 83.9/4.9/2.1% of the total variance for FGOALS or $\mathrm{CLIN}_{\mathrm{FGOALS}}$. The spatial correlation coefficient between the first to third EOF of FGOALS and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ are 0.999/0.997/0.994. These results suggest that the CLIN model can well reproduce the linear spatial mode of the considered climatological distribution.

Lastly, we examine the distribution of spatial variability across different spatial scales in the referential/generated dataset: we carry out 2D Fourier transform on the referential/generated samples, and draw the radial averaged squared magnitude of the complex Fourier coefficients as function of wave numbers (Fig. 3). The radially averaged power spectrum density of the considered referential and generated data samples follow a similar power-law scaling, suggesting that the CLIN model can well reproduce the spatial variability across scales.

To sum up, the analysis of both grid-scale and field-scale statistics demonstrates that the CLIN methodology accurately reproduces the essential characteristics and patterns of the climatological distribution present in the training data. We can thereafter leverage this learned climatological prior for the state inference task.

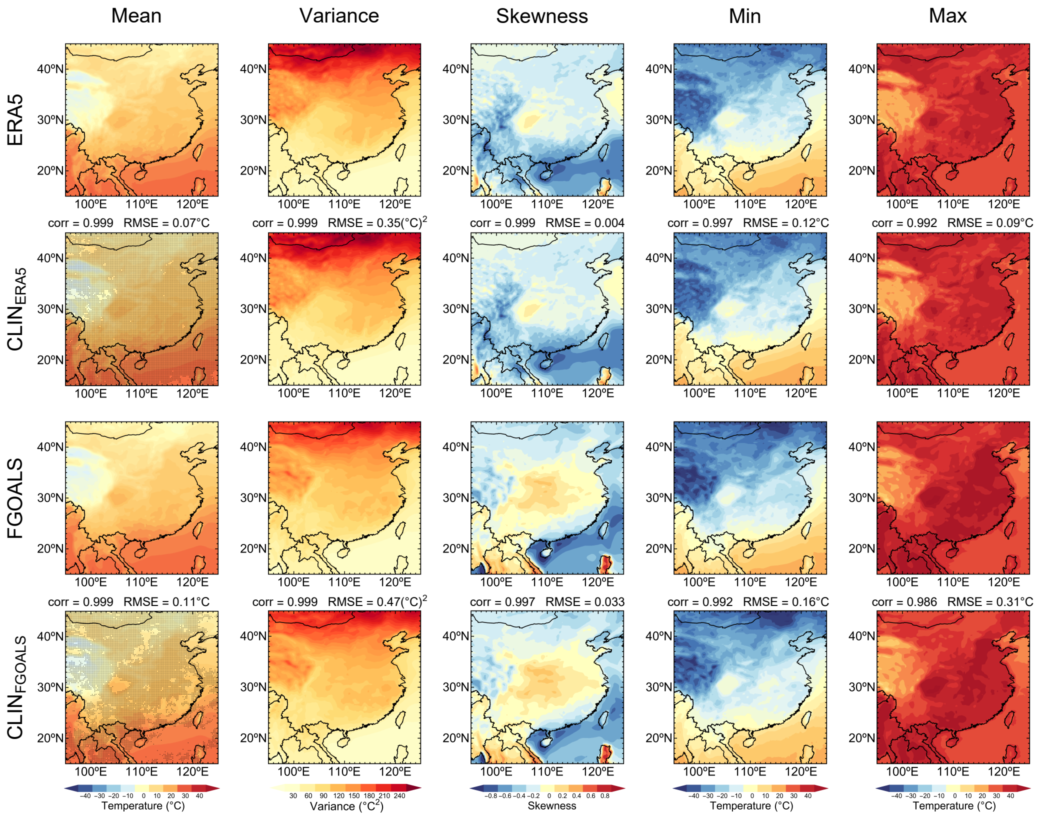


Figure 2. Grid-scale comparison of climatological statistics for climate reanalysis (ERA5, Row 1), climate simulation (FGOALS, Row 3), and probabilistic diffusion models trained using these datasets ($\mathrm{CLIN}_{\mathrm{ERA5}}$, Row 2; and $\mathrm{CLIN}_{\mathrm{FGOALS}}$, Row 4). The considered statistics are mean, variance, skewness, minimum, and maximum. The spatial correlation coefficient (corr) and root mean squared error (RMSE) between the referential dataset statistics and generated dataset statistics are labeled. Stipples denote grids that pass the Kolmogorov-Smirnov test at 95% confidence interval.

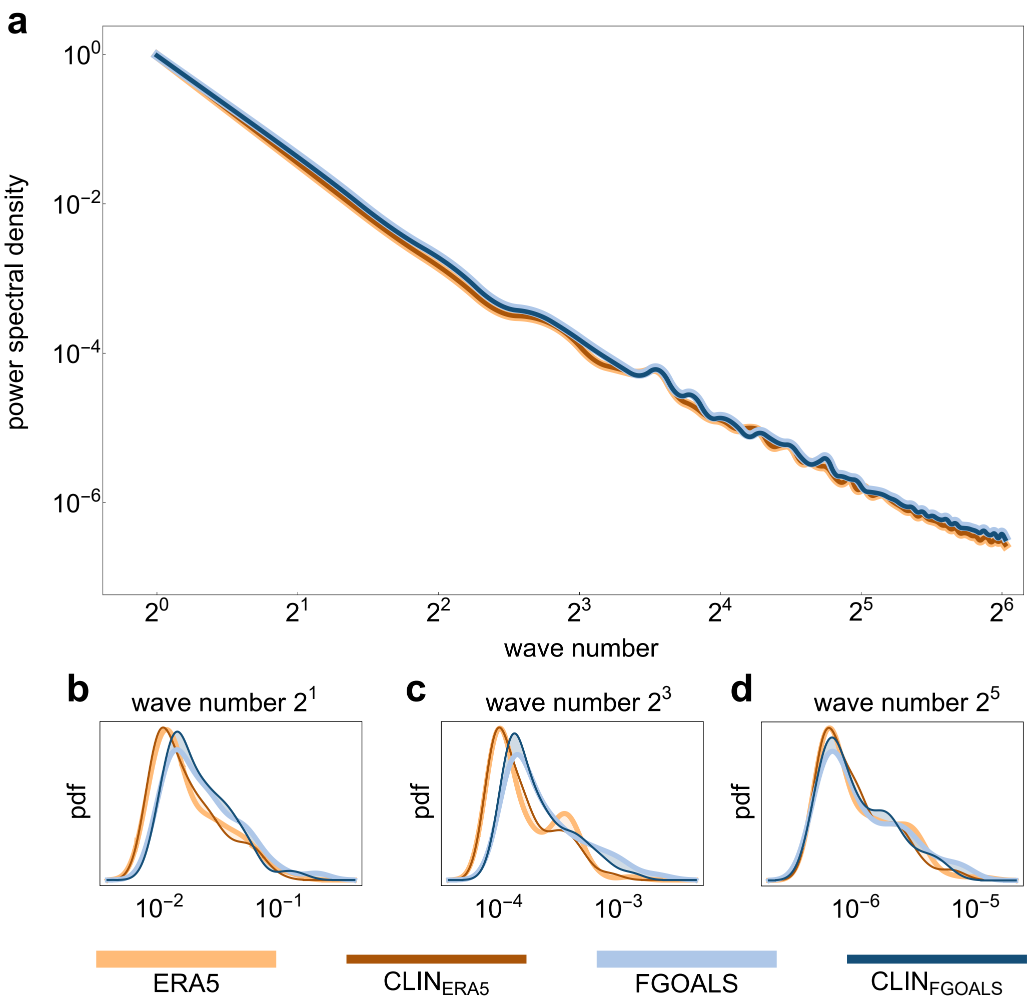


Figure 3. Radial averaged power spectrum density as function of wave number for 2 m temperature spatial pattern. **a**: results for ERA5, FGOALS, CLIN$_{\text{ERA5}}$, and CLIN$_{\text{FGOALS}}$ averaged over 100 ensemble members. **b-d**: probability distribution of power spectrum density at wave number 2^1, 2^3, 2^5 for ERA5, FGOALS, CLIN$_{\text{ERA5}}$, and CLIN$_{\text{FGOALS}}$.

3.2 Inferring weather states using partial observations

Given a learned climatological prior, we assess how well we can combine it with partial observations to obtain probabilistic estimate of the 2 m temperature spatial patterns. The climatological priors are probabilistic diffusion models trained using climate reanalysis (ERA5) and climate simulation (FGOALS) data. The observations are from 131 operational meteorological stations across China. We randomly select 120 of these stations to inpaint into the generation process, and leave the rest 11 stations for test. For regions without station observations, we consider ERA5 data as benchmark. Below we report case example results (Sec. 3.2.1) and a 1-year round skill assessment (Sec. 3.2.2) .

3.2.1 Case study

We consider four case examples covering different hours of a day and different seasons (Fig. 4). To make probabilistic inference of spatial patterns using partial observations, we gradually inpaint station observations into the generation process of $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$, creating 100 ensemble members for each model and each case. We report the ERA5 spatial pattern (Fig. 4 Row 1), the ensemble mean (Fig. 4 Row 2 and 5), the standard deviation of the ensemble (Fig. 4 Row 3 and 6), the mean squared error between ERA5 and the ensemble members (Fig. 4 Row 4 and 7) for $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$.

Both the repainted $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ ensemble mean results closely match the ERA5 spatial pattern, regarding latitudinal gradient, influence of topography, and the land-sea contrast, yielding spatial correlation coefficient of $0.980\pm0.02/0.977\pm0.02$ for the four considered case examples. These results suggest that the proposed methodology allows effectively propagation of information from limited ($\sim 1\%$) observed locations to a broad range of unobserved parts.

Next, we test if the CLIN methodology offers reliable uncertainty quantification (Fig, 4 Row 3 and 6). A larger ensemble variance indicates greater uncertainty in the estimate, while a smaller variance suggests more confidence in the estimate. As is expected, geogrids close to observation stations tend to have low ensemble variance, while distant ones may have relatively higher ensemble variance. The information constraint from observations may be blocked by topography, such as for Tibetan Plateau and Tian Shan Mountains. While for plain regions, we can expect a larger extension of observation constraints. We further examine the relationship between the spread of the ensemble members and their estimation skill, by computing the correlation between ensemble variance and ensembles' mean squared error score. The high spread skill correlation for $\mathrm{CLIN}_{\mathrm{ERA5}}$ (0.90 ± 0.08) and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ (0.94 ± 0.04) suggest that ensemble spread is a good predictor of model's estimation skill. This means that the CLIN model can capture the underlying uncertainties and provide reliable estimates of spatial estimation confidence.

To sum up, the case studies confirm that the CLIN methodology can make successful probabilistic inference of 2 m temperature spatial patterns using limited observations. The results are spatially coherent, well-constrained by observations, and offer reliable uncertainty quantification. It is worth noting that there are unneglectable mismatches between station observations and ERA5/FGOALS, regarding either climatological statistics or values. These mismatches introduce domain shift error, which is frequently encountered as we deploy a machine learning model in real-world scenarios where the data distribution differs from the training data. Below we dissect this error source by inpainting with different data sources in a 1-year round evaluation.

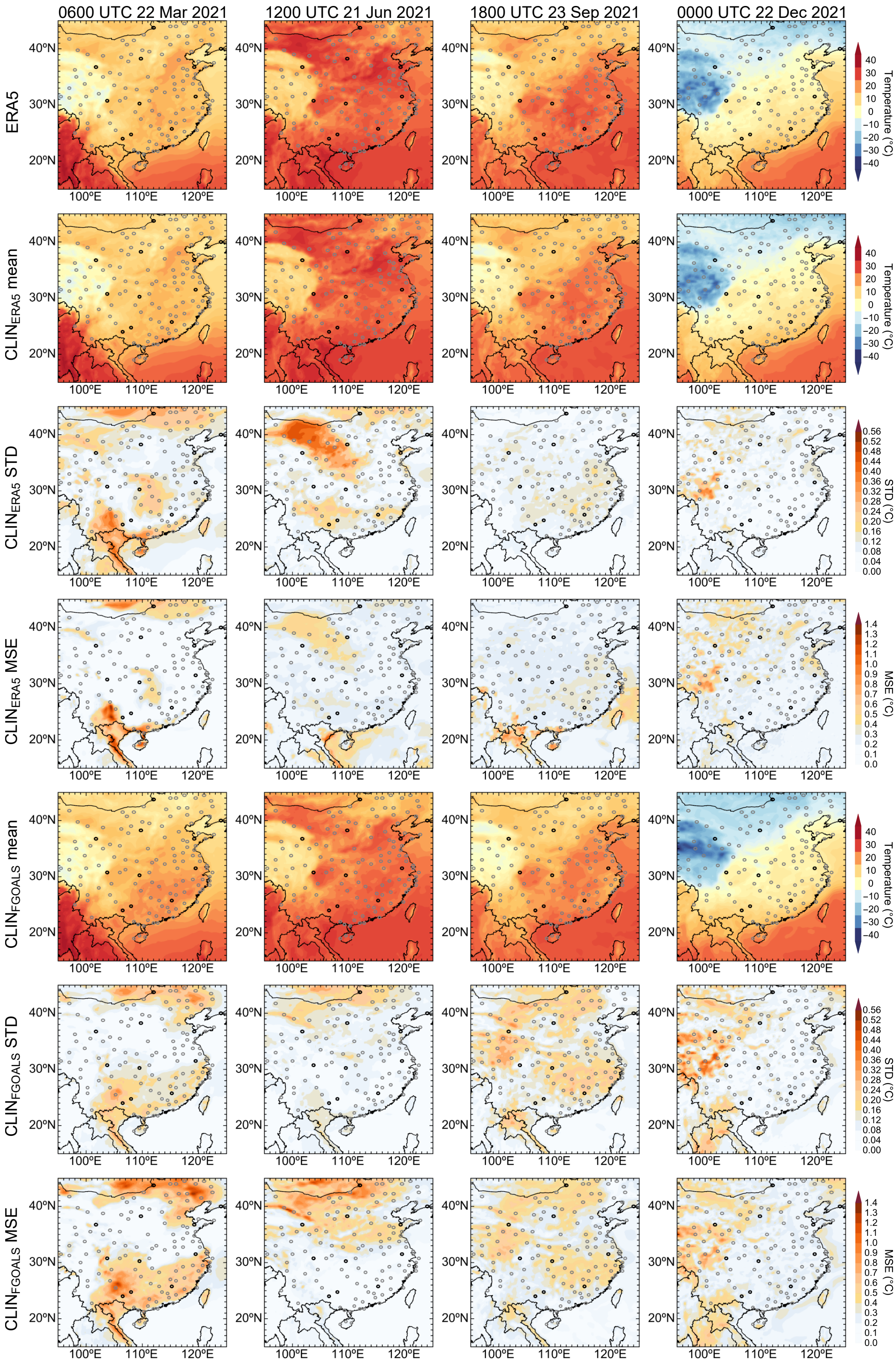


Figure 4. Case examples for probabilistic inference for 2 m temperature spatial pattern using partial observations. For $\mathrm{CLIN_{ERA5}}$ and $\mathrm{CLIN_{FGOALS}}$, 100 ensemble members are created by repainting observations. The ERA5 spatial pattern (Row 1), the ensemble mean (Row 2 and 5), the standard deviation of the ensemble (Row 3 and 6), the mean squared error between ERA5 and the ensemble members (Row 4 and 7) for $\mathrm{CLIN_{ERA5}}$ and $\mathrm{CLIN_{FGOALS}}$ are plotted.

3.2.2 Skill evaluation

We conduct a year-long evaluation of the models' performance in inferring spatial patterns, using data from Year 2021, which are not included in the models' training process. We compare ERA5 with $\mathrm{CLIN_{ERA5}}$ and $\mathrm{CLIN_{FGOALS}}$, both inpainted using station observations, and present the spatial distribution of their RMSE in Fig. 5a and Fig. 5b. To further investigate different uncertainty sources in the state inference task, we also consider inpainting $\mathrm{CLIN_{ERA5}}$ using ERA5 data at the observation stations. The RMSE between this inpainted $\mathrm{CLIN_{ERA5}}$ and the ERA5 whole-field data is shown in Fig. 5c.

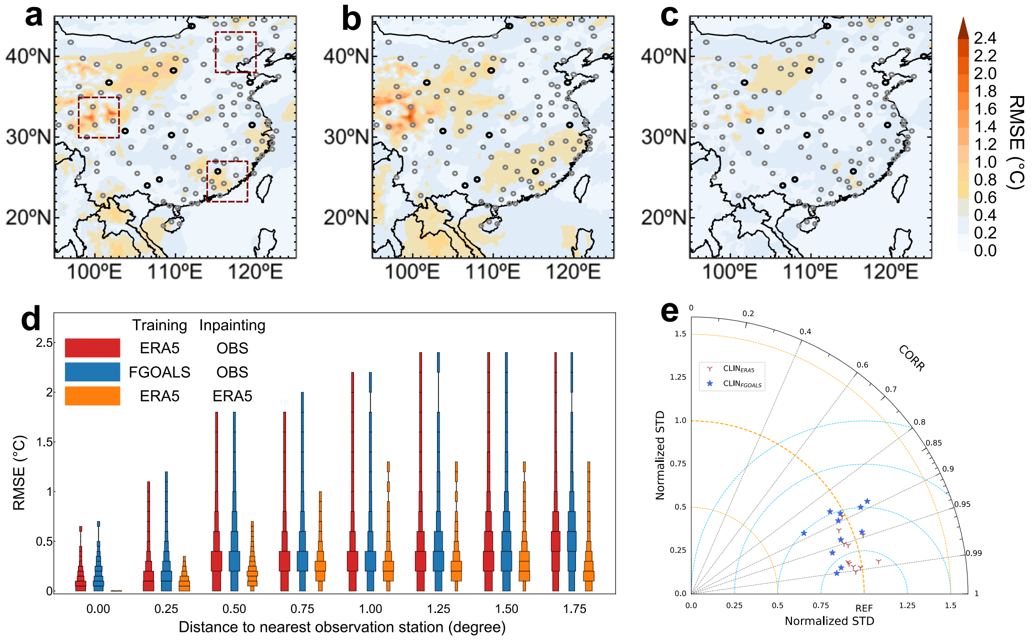


Figure 5. Skill evaluation for CLIN models to estimate spatial pattern of 2 m temperature using data for Year 2021. **a**: root mean squared error (RMSE) between ERA5 reanalysis and $\mathrm{CLIN_{ERA5}}$ inpainted using station observations; **b**: RMSE between ERA5 reanalysis and $\mathrm{CLIN_{FGOALS}}$ inpainted using station observations; **c**: RMSE between ERA5 reanalysis and $\mathrm{CLIN_{ERA5}}$ inpainted using ERA5 data at station observations; **d**: distribution of RMSE as function of grid's distance to nearest observation station for the three considered methods; **e**: Taylor diagram comparing the left-out station observations with $\mathrm{CLIN_{ERA5}}$ (orange) and $\mathrm{CLIN_{FGOALS}}$ (blue) results. Both $\mathrm{CLIN_{ERA5}}$ and $\mathrm{CLIN_{FGOALS}}$ are constrained by 120 station observations here. We delineate three representative regions to evaluate the value of observations in Sec. 3.3

The RMSE between ERA5 and observation inpainted $\mathrm{CLIN_{ERA5}}/\mathrm{CLIN_{FGOALS}}$ is 0.25 ± 0.21°C/0.31 ± 0.20°C, suggesting that the CLIN methodology enables accurate spatial pattern estimates. Both models exhibit low uncertainty in plain terrain regions or over the ocean, despite that no ocean observations were applied. This suggests that the learned climatological prior effectively captures the spatial patterns and variability in these regions, allowing the models to make confident estimates using limited and far-away observational constraints. On the other hand, both models exhibit higher uncertainty in regions with complex terrain, such as the Tibetan Plateau and the mountainous areas of Southeast China. Additionally, land areas with complicated terrain but lacking observational constraints, such as Southeast Asia, also show large uncertainty in the model estimates.

The uncertainty in state inference comes from the following three sources (Tab. 1). The first is domain shift error, which is due to distribution mismatch among data applied for model training, data applied for inpainting, and data applied for skill evaluation. The second is model error, which is due to the approximation/optimization/statistical error in applying probabilistic diffusion model to fit climatological prior, or due to errors in inpainting. These two types of uncertainties are *epistemic*, as they could be reduced by gathering more data, improving the model, or incorporating knowledge about data distribution differences. The third source of uncertainty is intrinsic/aleatoric, which is due to existence of multiple plausible spatial patterns given partial observational constraints, reflecting the inherent randomness in the system being modeled.

To disentangle these uncertainty sources, we consider the following comparisons.

1. We compare the RMSE of $\mathrm{CLIN}_{\mathrm{ERA5}}$ (Fig. 5a) and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ (Fig. 5b). $\mathrm{CLIN}_{\mathrm{ERA5}}$ achieves an overall lower RMSE, which can be attributed to a relieved domain shift error from the following two aspects: a. compared to FGOALS, ERA5 better matches the "true" climatology as partially revealed by the scattered observations; b. we consider ERA5 data as "ground truth" for evaluating model performance, which gives advantage to CLIN model trained using ERA5 data.
2. We compare $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using observation data (Fig. 5a) and $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using scattered ERA5 data (Fig. 5c). The latter achieve significantly lower RMSE ($0.19 \pm 0.12°\mathrm{C}$), suggesting a relatively low model error and a relatively low intrinsic uncertainty of the considered task. The difference between these cases highlights the domain shift error as the observation distribution differs from ERA5.
3. We compare the performance of $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ in predicting the observations at test stations that are excluded during repainting (Fig. 5e). For these test stations, both $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ results show high correlation coefficient (0.87-0.99) and low root mean squared error (0.2-0.4°C) with the observations , with $\mathrm{CLIN}_{\mathrm{ERA5}}$ performing slightly better than $\mathrm{CLIN}_{\mathrm{FGOALS}}$; $\mathrm{CLIN}_{\mathrm{ERA5}}$ holds a normalized standard deviation close to 1, which closely matches the observations, while $\mathrm{CLIN}_{\mathrm{FGOALS}}$ holds a normalized standard deviation slightly less than 1, suggesting a smaller temporal variability.

Table 1. Uncertainty sources for state inference using partial observations

Uncertainty source	Type	Illustration
Domain shift	Epistemic	Distribution mismatch among data applied for model training, data applied for inpainting, and data applied for skill evaluation.
Model error	Epistemic	1. Approximation/optimization/statistical error in fitting climatological prior. 2. Error in constraining the prior with observations.
Intrinsic uncertainty	Aleatoric	Existence of multiple plausible spatial patterns given observational constraints.

Finally, we quantify the spatial extension of observational constrains by showing models' RMSE skill as function of grid's distance to nearest observation station (Fig. 5d). We consider $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using observation data and ERA5 data, as well as $\mathrm{CLIN}_{\mathrm{FGOALS}}$ inpainted using observation data. For all these cases, models' performances at an arbitrary grid depends closely on the grid's proximity to observations. Meanwhile, there is large variation of models' RMSE skills for grids that are at least 1° away from any observation stations. Below we further investigate the value of individual observations in constraining the variability of its nearby spatial patterns, and offer guidelines for better observation planning.

3.3 On the value of observations

We apply the CLIN methodology to quantify the value of observations in constraining state estimation uncertainty, using three representative regions delineated in Fig. 5a. To achieve this, we add or remove observational stations and evaluate the impact on the estimation error (Fig. 6). Here, the first column shows the RMSE spatial pattern for the original $\mathrm{CLIN_{ERA5}}$ model estimates in each target region; the second column (Adding) demonstrates the impact of adding an observation station in a high-error area; the third column (Removing) illustrates the effect of removing an existing observation station; the fourth column (Terrain) provides a topographical context for each target region.

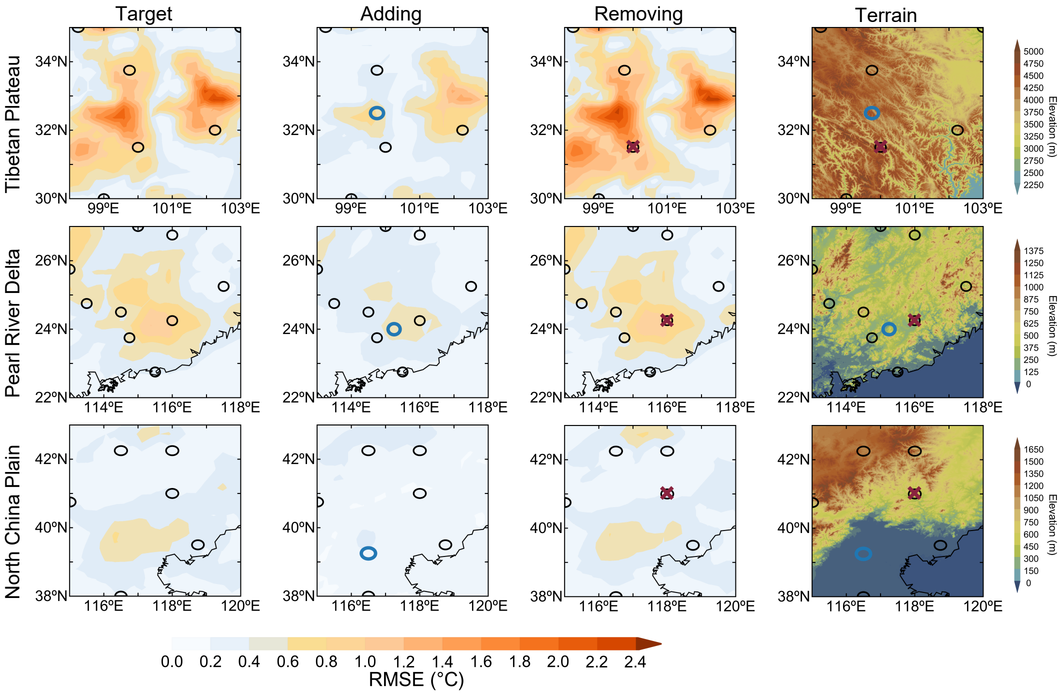


Figure 6. Evaluation of CLIN in reconstructing 2m temperature spatial pattern using different observation setups. Column 1: RMSE between $\mathrm{CLIN_{ERA5}}$ inpainted using observation data and ERA5 for three selected regions delineated in Fig. 5. Column 2: RMSE after including a pseudo new observation. This new observation data is from ERA5. Column 3: RMSE after including a pseudo new observation. Column 4: elevation map of the considered regions. The results are based on a year-long (Year 2021) evaluation.

For the case of Tibetan Plateau (first row), where the terrain is highly complex, with average elevations exceeding 4500 meters, we obtain a relatively high RMSE given existing observation constrains, particularly in the central and eastern parts of the region. Adding a station in the high-error area significantly reduces the RMSE for a broad range of the considered region, this impact is more pronounced here as compared to the other two cases, highlighting the importance of observational constraints in areas with complex terrain. Removing a station results in a noticeable increase in RMSE in the surrounding areas. Similarly, the effect of station removal is more evident compared to the other two cases, suggesting that the model heavily relies on the limited observational data to constrain its estimates in this complex terrain. The loss of a station in a critical location can greatly impact the model's ability to capture the local temperature patterns.

For the case of Peal River Delta (second row), the terrain is characterized by a mix of lowlands and hilly regions, with elevations ranging from 0 to 1000 meters. The original RMSE is low overall, with some higher values in the central and northwest mountain regions. Adding a station in the high-error area effectively reduces the RMSE. Meanwhile, removing a station leads to a hardly noticeable increase in RMSE in the surrounding areas.

For the case of North China Plain (third row), the northern part is featured by mountainous terrains exceeding 1000 meters, and the southern part has flat topography and homogeneous terrain. Adding a station in the central of southern plain area reduces the RMSE significantly, as existing observations are either from the northern mountain areas, or is too far away. Same as previous case, removing a station has minimal impact on the RMSE distribution.

To sum up, we discusses the application of the CLIN methodology to evaluate the impact of observational data on state estimation uncertainty across three diverse regions. It emphasizes the importance of strategic addition and removal of observational stations in improving estimation accuracy, particularly in areas with complex terrain. The findings highlight how existing observation constraints influence RMSE distribution, with significant reductions observed when stations are added in high-error areas. Conversely, removal of stations leads to increased RMSE, underscoring the model's reliance on limited observational data. Overall, we provide valuable insights for optimizing the design of observation networks, leading to a reduction in uncertainties and biases in weather and climate analysis.

4 Conclusion

Accurate state estimation of Earth atmosphere marks a daunting task due to its high-dimensionality and chaotic nature. We demonstrated the potential of deep generative models, specifically probabilistic diffusion models, in learning the inherent low-dimensional statistical structure of atmospheric circulation from climate reanalysis and simulation data. By leveraging this learned climatological prior, we developed a methodology named CLIN (Climate Inpainting) to effectively infer weather states from partial observations.

For the case study of estimating 2 m temperature spatial patterns, the learned climatological prior accurately reproduced the essential characteristics and patterns of the training data at both grid-scale and field-scale. This learned prior effectively captured multi-scale climate patterns, providing regularization and stability to the state estimation task.

Combining the learned climatological prior with station observations, CLIN yielded strong posterior estimates of 2 m temperature spatial patterns. The estimates were spatially coherent, well-constrained by observations, and provided reliable uncertainty quantification. Regions near observation stations exhibited low ensemble variance, indicating high confidence in the estimates, while distant regions showed relatively higher ensemble variance. The high spread-skill correlation confirmed that the ensemble spread was a good predictor of the model's estimation skill.

Moreover, CLIN allowed us to quantify the value of each observation station in reducing state estimation uncertainty. By adding or removing stations and evaluating the impact on the estimation error, we demonstrated the potential of this approach in guiding the design of optimal observation networks.

Our study showcases the power of deep generative models in extracting and utilizing the information produced by the chaotic evolution of the climate system. The proposed CLIN methodology opens up new opportunities for data-driven weather state estimation, potentially complementing traditional data assimilation approaches.

Future work could focus on extending CLIN to handle indirect observations (i.e., remote sensing) and multiple interdependent variables, incorporating temporal dynamics, and adapting to long-term climate trends. Addressing the computational demands and data requirements of diffusion models is another important direction for making this approach more practical and accessible.

In conclusion, this study demonstrates the immense potential of deep generative models in advancing climate data exploration and tackling complex inference tasks in atmospheric sciences. By learning the intrinsic statistical structure of the climate system, these models can effectively bridge the gap between sparse observations and complete weather state estimates, paving the way for more accurate and efficient climate monitoring and prediction.

5 Data Availability

The ERA5 reanalysis data are obtained from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessible at https://cds.climate.copernicus.eu/.

The FGOALS model data are obtained from the Coupled Model Intercomparison Project Phase 6 (CMIP6), hosted by the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory (LLNL), accessible at https://pcmdi.llnl.gov/CMIP6/.

The observational data are freely available for download from the following website: http://www.ncdc.noaa.gov/oa/ncdc.html. The site information used in this study was obtained from the China Meteorological Data Network, hosted by the China National Meteorological Science Data Center (NMDC), accessible at http://data.cma.cn/.

6 Open Research

Model configuration, analysis scripts, data files used for this study will be publicly available upon accept of the work.

Acknowledgments

This research is supported by National Key R&D Program of China (Grant NoS. 2023YFC3007700 and 2023YFC3007705). We appreciate the insightful discussions with Dr. Niklas Boers from Technical University of Munich, and with Dr. Bin Wang and Dr. Juanjuan Liu from Chinese Academy of Science.

References

Bao, Q., Liu, Y., Wu, G., He, B., Li, J., Wang, L., ... others (2020). Cas fgoals-f3-h and cas fgoals-f3-l outputs for the high-resolution model intercomparison project simulation of cmip6. *Atmospheric and Oceanic Science Letters*, *13*(6), 576–581.

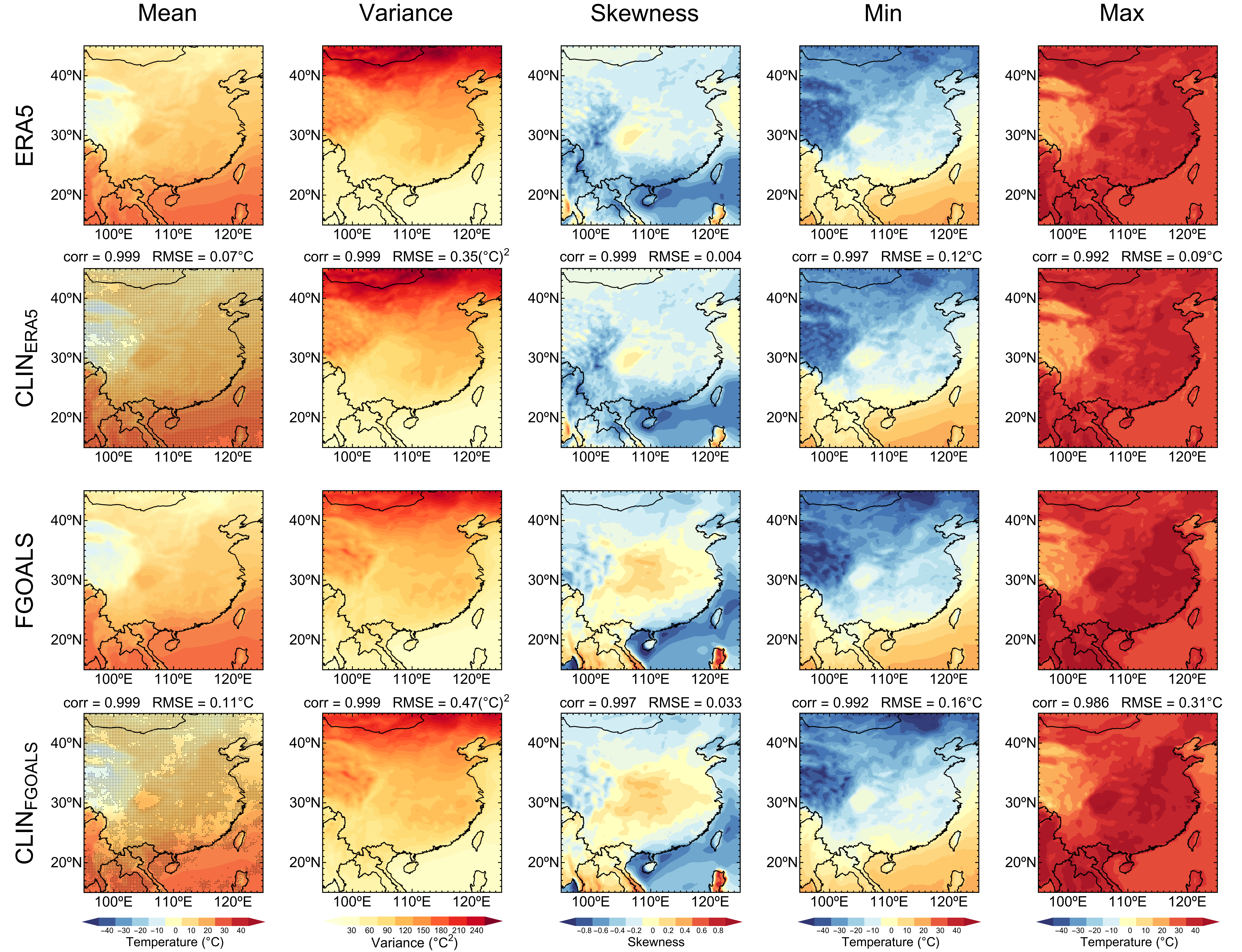
Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(5), e535.

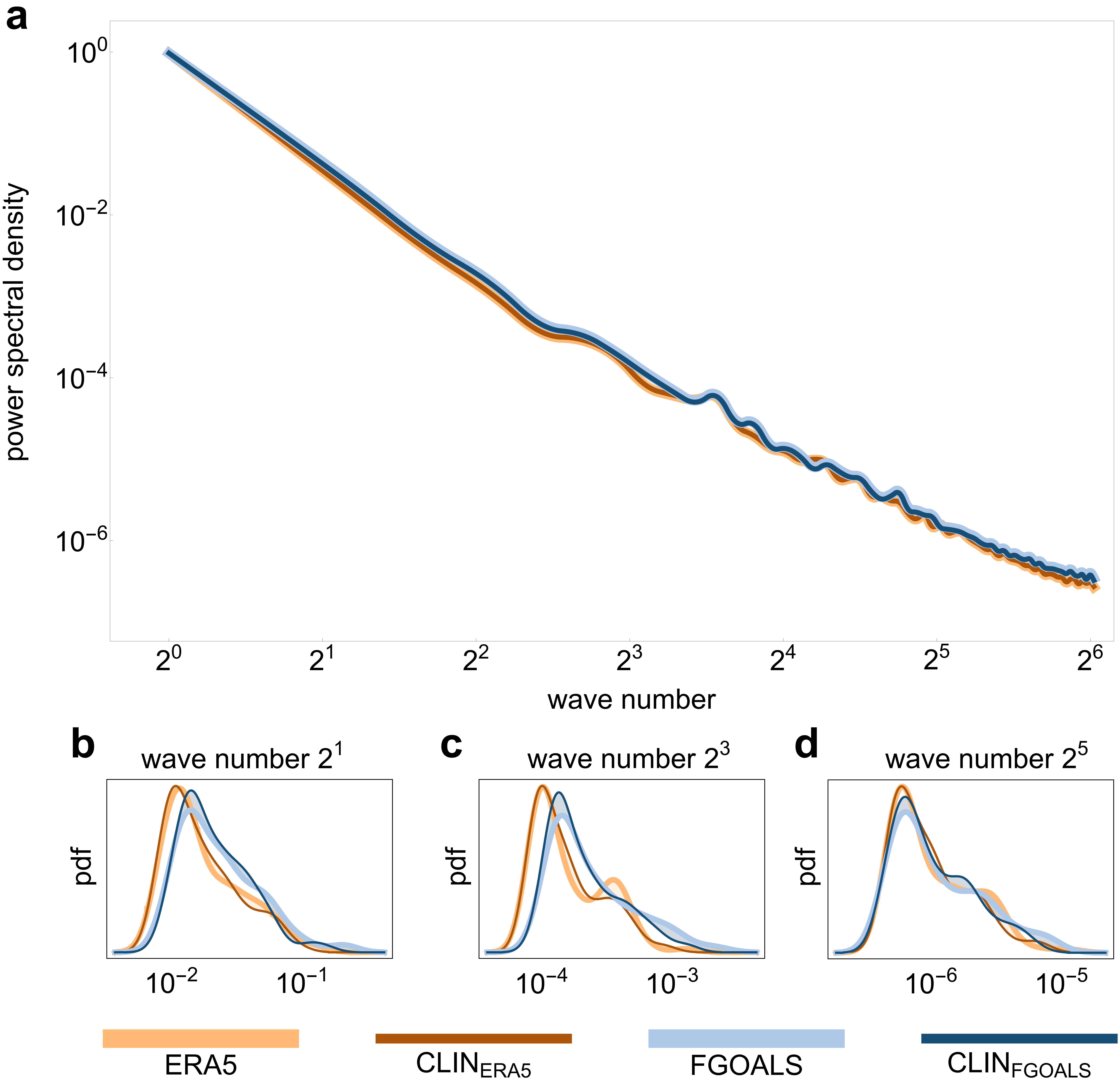
Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, *34*, 8780–8794.

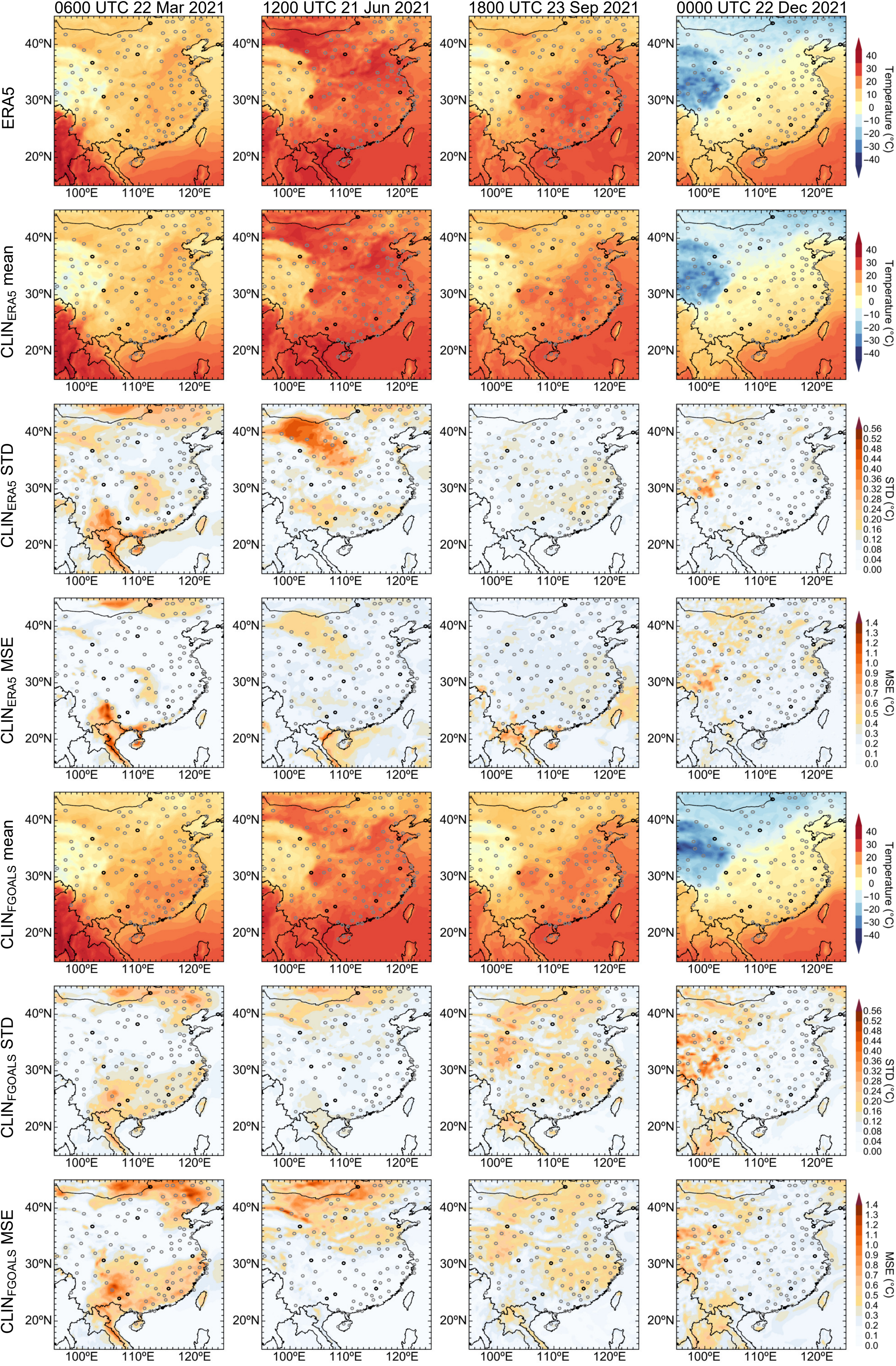
Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958.
Ghil, M. (2020). Hilbert problems for the climate sciences in the 21st century–20 years later. *Nonlinear Processes in Geophysics*, *27*(3), 429–451.
Gilpin, W. (2024). Generative learning for nonlinear dynamics. *Nature Reviews Physics*, 1–13.
Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., . . . others (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.
Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.
Holton, J. R., & Hakim, G. J. (2012). *An introduction to dynamic meteorology.* Academic press.
Kadow, C., Hall, D. M., & Ulbrich, U. (2020). Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, *13*(6), 408–413.
Kanngießer, F., & Fiedler, S. (2024). "seeing" beneath the clouds—machine-learning-based reconstruction of north african dust plumes. *AGU Advances*, *5*(1), e2023AV001042.
Kingma, D., Salimans, T., Poole, B., & Ho, J. (2021). Variational diffusion models. *Advances in neural information processing systems*, *34*, 21696–21707.
Law, K., Stuart, A., & Zygalakis, K. (2015). Data assimilation. *Cham, Switzerland: Springer*, *214*, 52.
Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11461–11471).
Nai, C., Pan, B., Chen, X., Tang, Q., Ni, G., Duan, Q., . . . Liu, X. (2024). Reliable precipitation nowcasting using probabilistic diffusion models. *Environmental Research Letters*.
Palmer, T. (2017). The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *Journal of Advances in Modeling Earth Systems*, *9*(2), 730–734.
Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., . . . Ma, H.-Y. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509.
Pan, B., Wang, L.-Y., Zhang, F., Duan, Q., Li, X., Pan, X., . . . others (2023). Probabilistic diffusion model for stochastic parameterization–a case example of numerical precipitation estimation. *Authorea Preprints*.
Pan, X., Guo, X., Li, X., Niu, X., Yang, X., Feng, M., . . . others (2021). National tibetan plateau data center: promoting earth system science on the third pole. *Bulletin of the American Meteorological Society*, *102*(11), E2062–E2078.
Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, *44*(24), 12–396.
Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256–2265).
Song, Y., Garg, S., Shi, J., & Ermon, S. (2020). Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in artificial intelligence* (pp. 574–584).
Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

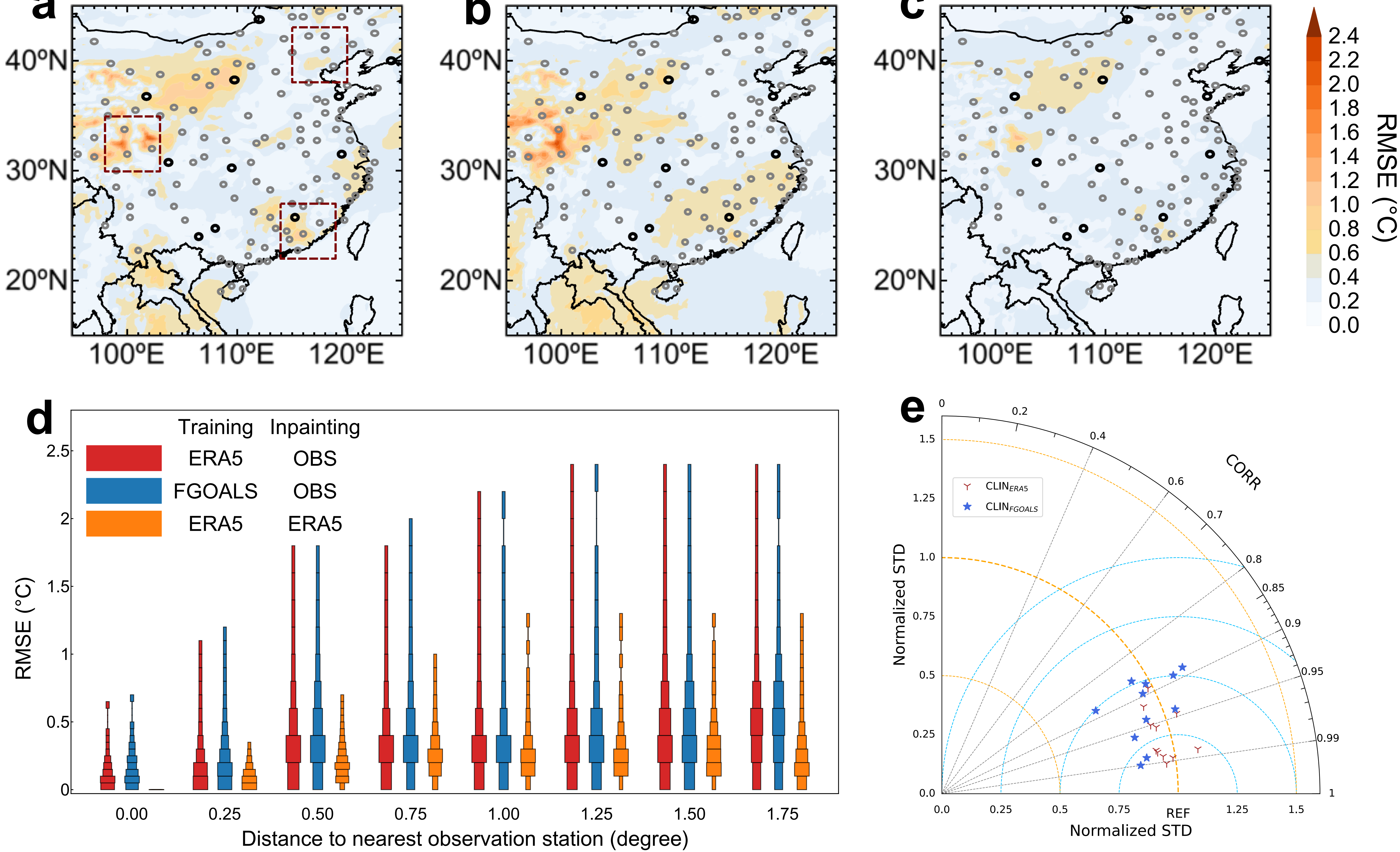
Toth, Z., Talagrand, O., Candille, G., & Zhu, Y. (2003). Probability and ensemble forecasts. *Forecast verification: A practitioner's guide in atmospheric science*, *137*, 163.
Wang, B., Zou, X., & Zhu, J. (2000). Data assimilation and its applications. *Proceedings of the National Academy of Sciences*, *97*(21), 11143–11144.
Zhang, G., Ji, J., Zhang, Y., Yu, M., Jaakkola, T. S., & Chang, S. (2023). Towards coherent image inpainting using denoising diffusion implicit models.

Inpainting
Observation
Forward process
$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$
Data
Noise
Generation
X_0
X_{t1}
X_{t2}
X_T
X_{t2}
X_{t1}
X_0
$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t), \Sigma_\theta(x_t))$
Reverse process

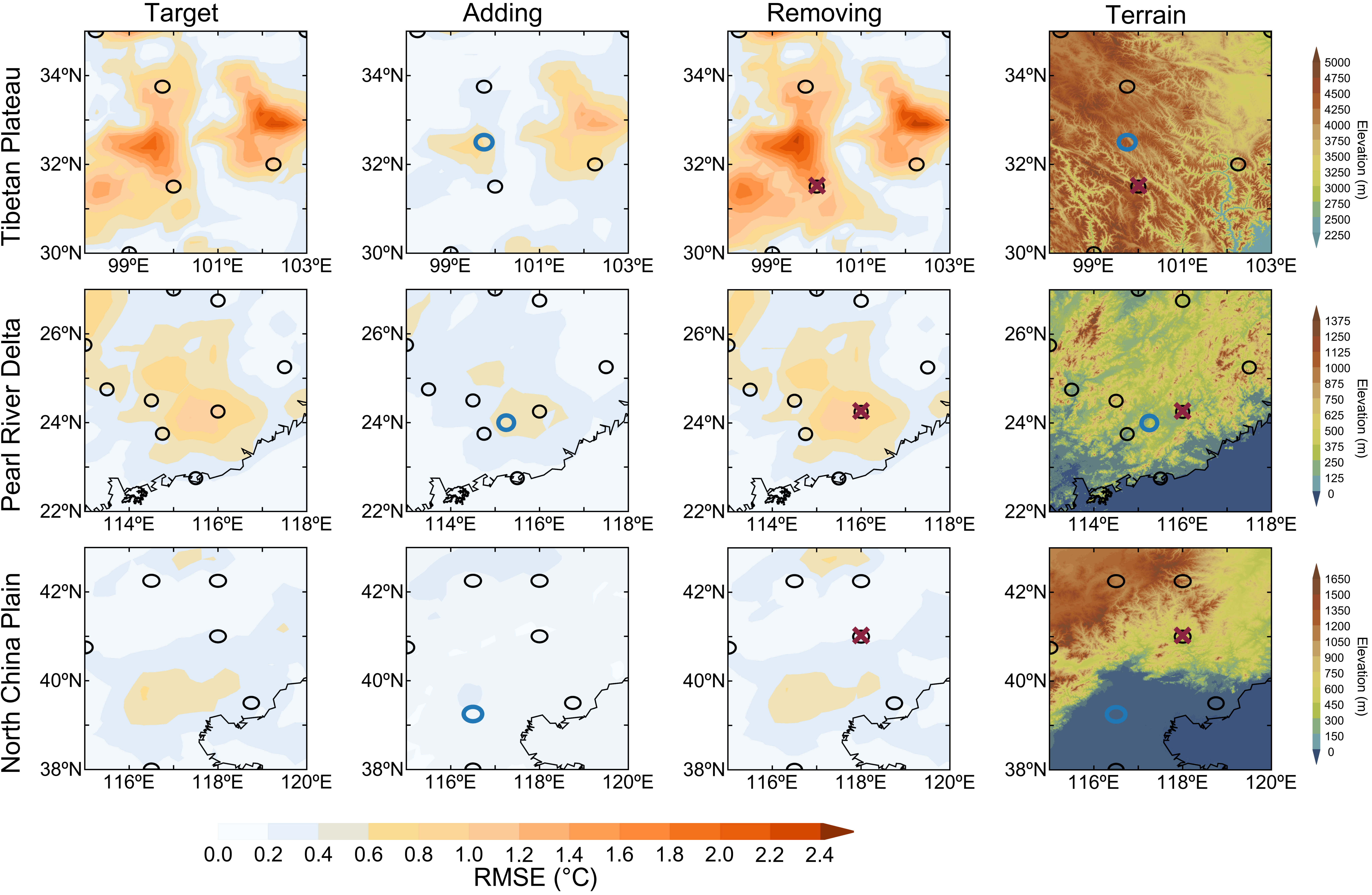

Mean
Variance
Skewness
Min
Max
ERA5
CLIN$_{ERA5}$
FGOALS
CLIN$_{FGOALS}$
40ºN
30ºN
20ºN
100ºE
110ºE
120ºE
corr = 0.999 RMSE = 0.07°C
corr = 0.999 RMSE = 0.35(°C)^2
corr = 0.999 RMSE = 0.004
corr = 0.997 RMSE = 0.12°C
corr = 0.992 RMSE = 0.09°C
corr = 0.999 RMSE = 0.11°C
corr = 0.999 RMSE = 0.47(°C)^2
corr = 0.997 RMSE = 0.033
corr = 0.992 RMSE = 0.16°C
corr = 0.986 RMSE = 0.31°C
−40 −30 −20 −10 0 10 20 30 40
Temperature (°C)
30 60 90 120 150 180 210 240
Variance (°C^2)
−0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8
Skewness


a
power spectral density
10^0
10^{-2}
10^{-4}
10^{-6}
2^0
2^1
2^2
2^3
2^4
2^5
2^6
wave number
b
wave number 2^1
pdf
10^{-2}
10^{-1}
c
wave number 2^3
pdf
10^{-4}
10^{-3}
d
wave number 2^5
pdf
10^{-6}
10^{-5}
ERA5
$CLIN_{ERA5}$
FGOALS
$CLIN_{FGOALS}$


0600 UTC 22 Mar 2021
1200 UTC 21 Jun 2021
1800 UTC 23 Sep 2021
0000 UTC 22 Dec 2021
ERA5
$CLIN_{ERA5}$ mean
$CLIN_{ERA5}$ STD
$CLIN_{ERA5}$ MSE
$CLIN_{FGOALS}$ mean
$CLIN_{FGOALS}$ STD
$CLIN_{FGOALS}$ MSE
Temperature (°C)
STD (°C)
MSE (°C)
40°N
30°N
20°N
100°E
110°E
120°E



a
b
c
40°N
30°N
20°N
100°E
110°E
120°E
RMSE (°C)
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
d
Training
Inpainting
ERA5
OBS
FGOALS
OBS
ERA5
ERA5
RMSE (°C)
Distance to nearest observation station (degree)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
e
CORR
CLIN$_{ERA5}$
CLIN$_{FGOALS}$
Normalized STD
REF


Target
Adding
Removing
Terrain
Tibetan Plateau
Pearl River Delta
North China Plain
34ºN
32ºN
30ºN
99ºE
101ºE
103ºE
26ºN
24ºN
22ºN
114ºE
116ºE
118ºE
42ºN
40ºN
38ºN
116ºE
118ºE
120ºE
Elevation (m)
5000
4750
4500
4250
4000
3750
3500
3250
3000
2750
2500
2250
1375
1250
1125
1000
875
750
625
500
375
250
125
0
1650
1500
1350
1200
1050
900
750
600
450
300
150
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
RMSE (°C)

Learning to infer weather states using partial observations

Jie Chao[1,2], Baoxiang Pan[2], Quanliang Chen[1], Shangshang Yang[2,3], Jingnan Wang[2,4], Congyi Nai[2,5], Yue Zheng[6], Xichen Li[2], Huiling Yuan[3], Xi Chen[2], Bo Lu[7], Ziniu Xiao[2]

[1]School of Atmospheric Sciences, Chengdu University of Information Technology, Sichuan, China
[2]Institute of Atmospheric Physics, Chinese Academy of Science, Beijing, China
[3]Key Laboratory of Mesoscale Severe Weather, Ministry of Education, and School of Atmospheric Sciences, Nanjing University, Jiangsu, China
[4]College of Computer, National University of Defense Technology, Hunan, China
[5]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[6]Clustertech LTD, Hong Kong, China
[7]National Climate Center, China Meteorological Administration, Beijing, China

Key Points:

- Deep generative model enables accurate spatial interpolation of weather variables from sparse observations.
- The model generates probabilistic weather estimates with reliable uncertainty quantification by combining learned priors and observations.
- The model quantifies the value of observations for reducing uncertainty, guiding optimal observation network design.

Corresponding author: Baoxiang Pan, panbaoxiang@lasg.iap.ac.cn

Abstract

Accurate state estimation of the high-dimensional, chaotic Earth atmosphere marks a Sisyphean task, yet is indispensable for initiating weather forecast and gauging climate variability. While much effort is devoted to assimilating observations and forecasts to infer weather state, the inherent low-dimensional statistical structure in atmospheric circulation, shaped by geophysical laws and geographic boundaries, is underutilized as informative prior for state inference, or as reference for assessing representative of existing observations and planning new ones. We realize these potential by learning climatological distribution from climate reanalysis/simulation, using deep generative model. For a case study of estimating 2 m temperature spatial patterns, the learned distribution faithfully reproduces climatology statistics. A combination of the learned climatological prior with few station observations yields strong posterior of spatial pattern estimates, which are spatially coherent, faithful and adaptive to observation constraints, and uncertainty-aware. This allows us to evaluate each observation's value in reducing state estimation uncertainty, and guide optimal observation network design by pinpointing the most informative sites. Our study showcases how generative models can extract and utilize information produced in the chaotic evolution of climate system.

Plain Language Summary

Accurate estimation of weather conditions across a large area is crucial but challenging due to the complex and chaotic nature of the atmosphere. Traditional methods rely on combining observations with forecasts, which can be computationally expensive and sensitive to model biases. We propose a new approach called Climate Inpainting (CLIN) that learns the inherent spatial patterns of the atmosphere from climate data using machine learning techniques. CLIN can effectively combine the learned patterns with limited observations to reconstruct complete spatial maps of weather variables, such as temperature. We demonstrate that CLIN can accurately reproduce the key spatial features and variability of temperature over East Asia. Moreover, CLIN can quantify the uncertainty in the estimated weather maps and evaluate the importance of each observation site in reducing the overall uncertainty. This information can guide the optimal design of weather station networks. Our approach showcases the potential of machine learning in utilizing the rich information contained in climate data to improve weather estimation and observation planning.

1 Introduction

The state of the Earth atmosphere, which concerns a broad range of socioeconomic sectors and the overall environment, is characterized by the spatial distribution of a specific set of physical properties, including temperature, pressure, wind speed and direction, density, concentration of water of different phases, composition of aerosol, greenhouse gas, etc (Holton & Hakim, 2012). To determine the atmosphere state at 50 km grid resolution requires estimating the value for all the above-mentioned physical properties at around $\sim 10^7$ grids (Schneider et al., 2017). Doubling the resolution increases the total number of grids by a factor of 8. This high dimensionality poses a daunting challenge for monitoring the atmosphere (Ghil, 2020).

Current operational forecasting centers routinely update their atmosphere state estimates by combining multi-source observations and previous forecasts, so as to reboot weather forecast and gauge climate variability (Carrassi et al., 2018). Ground based observations offer direct meteorological measurements, yet come with limited spatial coverage and high maintenance cost. Remote sensing offers broader spatial coverage, yet is indirect and error prone, requiring careful calibration based on ground-based observations.

Deficiencies in observation render it an ill-posed task to estimate the state of the high-dimensional Earth atmosphere, calling for strong prior to achieve feasible solution. Forecasts from previous time steps are frequently applied to serve this mission, carrying information from previous step observations to the current step via a process-based model (Wang et al., 2000). As a result, the state estimation accuracy depends on an intricate interplay among model biases, background uncertainty, and observation error, which cannot be effectively disentangled or controlled (Law et al., 2015). Moreover, to provide multi-scale background information using forecasting models requires operational run of large ensemble high-resolution numerical simulations, which is prohibitively expensive and burdensome (Toth et al., 2003; Palmer, 2017).

Is there extra information source for inferring the state of the high-dimensional, chaotic Earth atmosphere? It turns out that, the inherent low-dimensional statistical structure in atmospheric circulation, shaped by the underlying geophysical laws and quasi-static geographic boundaries, can serve as an informative prior for state inference. The Earth climate system, like any other chaotic system, is an information producer: it gradually reveals the characteristic structure of its phase space at ever-finer scales (Gilpin, 2024). By identifying and parameterizing this characteristic structure, we can potentially bypass the curse of high dimensionality, and make more efficient use of limited observations for the state inference task.

Some pioneering works have explored this direction, leveraging the inherent structure of climate data to fill in missing observations and rebuild historical climate records. For instance, Kadow et al. (2020) developed a partial convolution method to reconstruct historical global temperature patterns based on partial observations and climate simulation. Kanngießer and Fiedler (2024) applied a similar methodology to restore the spatial extent of dust plumes in cloud-masked satellite images. Most of these practices consider deterministic models, which are designed for specific “reconstruction” problem configurations, yielding deterministic results regardless of whether observations can adequately constrain the estimation uncertainty. As a result, these methodologies generalize poorly to state inference tasks where the number or layout of observations change, fail to reproduce extremes or apply for scenarios where only limited observations are available.

A solution to these dilemmas is to shift from deterministic model to probabilistic model (B. Pan et al., 2021). Specifically, we prefer to build a probabilistic model that explicitly represents the inherent statistical structure of the atmosphere as revealed by climate observations or simulations. Thereafter, we hope to effectively and efficiently combine the learned climatological prior with incomplete observations, so as to obtain strong posterior of spatial pattern estimates. This problem setup poses two stringent requirements on the underlying probabilistic model. First, the model must faithfully approximate the high-dimensional climatological distribution as generated by the chaotic evolution of climate dynamics. Second, the model must enable flexible probabilistic inference, allowing us to efficiently obtain posterior atmospheric state estimates given arbitrary observational constraints.

To fulfill these requirements, we resort to generative machine learning, in particular, probabilistic diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song, Sohl-Dickstein, et al., 2020; Kingma et al., 2021). Probabilistic diffusion models learn to approximate complex, high-dimensional probability distributions in an iterative manner, achieving unprecedented fitting capacity and controlling flexibility (B. Pan et al., 2023; Nai et al., 2024). To demonstrate the idea, we consider a case example of inferring the spatial pattern of 2 m temperature based on sparse observations from operational meteorology stations. We learn probabilistic diffusion models to approximate the climatological distribution of 2 m temperature spatial patterns from climate reanalysis or simulation data. After carefully assessing the model’s ability to reproduce climatology, we develop tools to “inpaint” arbitrary observation constraints into the sample generation process, yielding probabilistic 2 m temperature spatial pattern estimates. Finally, we apply this methodology to evaluate each observation's value in reducing state estimation uncertainty, and guide optimal observation network design by pinpointing the most informative sites.

2 Methodology

2.1 Data and problem setup

We consider the task of inferring the spatial pattern of 2 m temperature over East Asia ($15°\text{N} - 45°\text{N}, 95°\text{E} - 125°\text{E}$), using station observations covering $\sim 1\%$ grids of the considered region. To achieve this, we learn climatological distribution of 2 m temperature spatial pattern using climate reanalysis or simulation data. The reanalysis data are hourly, 0.25° 2 m temperature data from the fifth-generation global climate and weather reanalysis (ERA5) developed at European Centre for Medium-Range Weather Forecasts (Hersbach et al., 2020, ECMWF). The simulation data are 3-hourly, 0.25° 2 m temperature historical simulation from the Flexible Global Ocean-Atmosphere-Land System Model version f3-H (Bao et al., 2020, FGOALS-f3-H), which participates in the sixth phase of the Coupled Model Intercomparison Project (Eyring et al., 2016, CMIP6). The station observation data are obtained from the Chinese National Climatic Data Center (X. Pan et al., 2021).

Formally, we denote the spatial pattern of 2 m temperature for the target region as $\mathbf{x}$, which is a 120×120 dimensional random variable here. Our objective is to approximate the distribution of $\mathbf{x}$, based on large number of samples from climate reanalysis or simulation:

$$p_{\theta^*} = \arg\max_{p_\theta} \sum \log p_\theta(\mathbf{x}) \tag{1}$$

Here p_θ is parameterized probability density function approximator, θ^* is the optimal parameter, optimized by maximizing the overall likelihood of p_θ assigned to the training samples.

Given p_{θ^*} and sparse observations, we need to provide probabilistic estimates of 2 m temperature spatial patterns, i.e., $p_{\theta^*}(\mathbf{x}|\mathbf{x}\odot\mathbf{m})$. Here, $\odot$ is dot product, $\mathbf{m}$ is observation mask, with value 1/0 denoting the existence/absence of observations for each geogrid. $p_{\theta^*}(\mathbf{x}|\mathbf{x}\odot\mathbf{m})$ should yield samples that are spatially coherent and faithful to observational constraints. Also, $p_{\theta^*}(\mathbf{x}|\mathbf{x}\odot\mathbf{m})$ should offer accurate uncertainty quantification. For instance, geogrids close to observation stations should typically have low state estimate uncertainties, while distant ones have high uncertainties. Finally, we prefer $p_{\theta^*}(\mathbf{x}|\mathbf{x}\odot\mathbf{m})$ to be adaptive to changes in observation configurations, such as the abortion or inclusion of observation stations, or rearrangement of station network layout. Below we illustrate how to achieve these requirements using the proposed methodology.

2.2 Learning climatology with probabilistic diffusion model

We elucidate how to learn climatological distribution of the target random variable using probabilistic diffusion model, thereafter leverage this learned prior for the inference task (Sec. 2.3). For clarity, we only cover key steps necessary for establishing our methodology. Details can be found in the literature referenced through the description.

To approximate a target distribution using probabilistic diffusion model, we train a series of deep neural networks that can be chained to establish bijective mapping between the target distribution and a prior distribution (Sohl-Dickstein et al., 2015; Ho et al., 2020). Specifically, we define the following Gaussian process:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{(1-\beta_t)}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

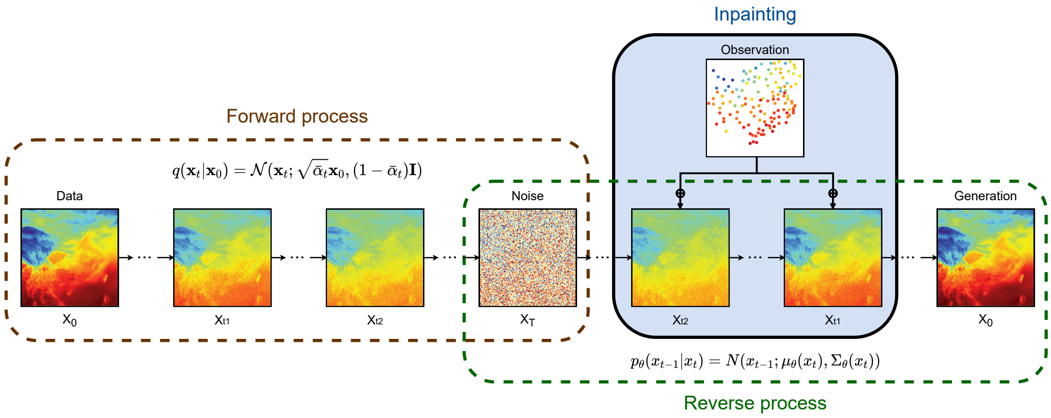


Figure 1. Overview of the Climate Inpainting (CLIN) methodology. A pre-defined forward Gaussian process (left) turns distribution of target climate variable into a prior distribution, i.e., standard Gaussian. A learned reverse Gaussian process (right) turns the prior distribution into the distribution of the target climate variable. We "inpaint" sparse observations throughout the reverse Gaussian process (right top), so as to obtain spatial pattern estimates of the target variable.

Here $p(\mathbf{x}_0) = p(\mathbf{x})$, which is the target distribution; $p(\mathbf{x}_T)$ is the prior distribution; we bridge $\mathbf{x}_0$ and $\mathbf{x}_T$ using $\mathbf{x}_{t\in[1,T]}$, which are latent variables with increasing noise level; $\mathcal{N}$ is Gaussian distribution; $\mathbf{I}$ is identity matrix; β_t is diffusion coefficient, which is pre-defined so that, give large enough T, $p(\mathbf{x}_T|\mathbf{x}_0)$ is drawn close to $p(\mathbf{x}_T)$, which is $\mathbf{x}_0$ agnostic. This setup offers analytical solution for $p(\mathbf{x}_{t+\tau}|\mathbf{x}_t), \forall\tau \in [0, T-t], t \in [0, T]$, facilitating convenient inference as detailed in Sec. 2.3.

To achieve generative modeling, we reverse Eq. 2 using the following variation distributions:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta, \mathbf{\Sigma}_\theta) \tag{3}$$

Here $\mathbf{\Sigma}_\theta$ is represented as an interpolation between its analytical lower and upper bound (Dhariwal & Nichol, 2021); μ_θ can be optimized by maximizing the variational lower bound (ELBO) on the log-likelihood of the training samples (Sohl-Dickstein et al., 2015; Kingma et al., 2021). In practice, we represent μ_θ as function of neural network parameterization for $\nabla p(\mathbf{x}_t|\mathbf{x}_0)$, which is known as the *score function* (Song, Garg, et al., 2020; Song, Sohl-Dickstein, et al., 2020). This simplifies the ELBO objective function to the following form:

$$L = \mathbb{E}_{t\in[1,T],\mathbf{x}_0\sim p(\mathbf{x}_0)}||\nabla p(\mathbf{x}_t|\mathbf{x}_0) - \epsilon_\theta||^2 \tag{4}$$

Here ϵ_θ is a neural network parameterization for $\nabla p(\mathbf{x}_t|\mathbf{x}_0)$. Given the trained score estimates, we can derive $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta, \Sigma_\theta)$ and sample it, starting with $p(\mathbf{x}_T)$, ending with $p(\mathbf{x}_0)$.

2.3 CLIN: inferring weather states using partial observations

We combine the learned climatology prior with station observations to infer the posterior probability distribution of the target variable, using a *repainting* methodology (Lugmayr et al., 2022; Zhang et al., 2023). Specifically, given a pre-trained diffusion model that sequentially applies $p_{\theta^*}(\mathbf{x}_t|\mathbf{x}_{t+1}) = \mathcal{N}(\mathbf{x}_t; \mu_{\theta^*}, \Sigma_{\theta^*})$ to transform $p(\mathbf{x}_T)$ to $p(\mathbf{x}_0)$, within a pre-selected time window of Ω, for grid points where we have observations, we replace values of $\mathbf{x}_t$ with observations noisified to time step t, by sampling $p(\mathbf{x}_t \odot \mathbf{m} | \mathbf{x}_0 \odot \mathbf{m})$. This replacement does not consider the generated parts of $\mathbf{x}_t$, therefore, the observations could not explicitly constrain the variability of unobserved parts.

To address this issue, for any $t \in \Omega$, after the replacement, instead of progressing to $t-1$ directly, we rewind to time step $t-\tau$ by sampling $p(\mathbf{x}_{t-\tau} | \mathbf{x}_t)$. We thereafter repeat the denoising steps from $t-\tau$ to t for k rounds, and carry out observation replacement for $\mathbf{x}_t$ at each round. This allows us to jointly modify both observed and unobserved regions throughout the denoising steps, yielding generated samples that are spatially coherent, faithful and adaptive to observation constraints, and uncertainty-aware. This methodology is referred to as *inpainting*, we hence name our methodology as CLIN, short for Climate Inpainting. A formal algorithm description is given below. Details for data processing, neural network architecture, hyperparameters for training and inference, are given in Supporting Information.

Algorithm 1 CLIN

Require: trained diffusion model p_{θ^*}, observations $\mathbf{x}_0 \odot \mathbf{m}$, repainting time step set Ω, rewinding step τ, rewinding round K

Ensure: observation constrained, spatially coherent sample $\mathbf{x}_0$

Initialize $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T-1, \ldots, 1$ **do**
  $\mathbf{x}_t \sim p_{\theta^*}(\mathbf{x}_t | \mathbf{x}_{t+1})$ ▷ Reverse sampling
  if $t \in \Omega$ **then**:
    for $k = 1, \ldots, K$ **do**
      $\mathbf{x}_t^{\text{obser}} \sim p(\mathbf{x}_t \odot \mathbf{m} | \mathbf{x}_0 \odot \mathbf{m})$
      $\mathbf{x}_t \leftarrow \mathbf{x}_t \odot (\mathbf{I} - \mathbf{m}) + \mathbf{x}_t^{\text{obser}}$ ▷ Condition on observations
      $\mathbf{x}_{t+\tau} \sim p(\mathbf{x}_{t+\tau} | \mathbf{x}_t)$ ▷ Rewind in time by τ steps
      for $i = t+\tau-1, \ldots, t$ **do**
        $\mathbf{x}_i \sim p_{\theta^*}(\mathbf{x}_i | \mathbf{x}_{i+1})$ ▷ Reverse sampling within a rewinding round
      end for
    end for
  end if
end for
return $\mathbf{x}_0$

3 Results

The accuracy for state estimation depends on 1) how well we can approximate the climatological distribution, and 2) based on a learned climatological prior, how well we can combine it with limited observations to obtain probabilistic state estimates. Below we assess model's performance for these two aspects (Sec. 3.1 and 3.2). We further employ the model to quantify the extent to which observations reduce uncertainty in state estimation, offering insights for optimal observation design (Sec. 3.3).

3.1 Climatology

We compare grid-scale and field-scale statistics of 10,000 reference/generated samples to evaluate how well the probabilistic diffusion models reproduce their training data's climatology. Two models trained with climate reanalysis (ERA5) and historical climate simulation (FGOALS) data, hereafter referred to as $\text{CLIN}_{\text{ERA5}}$ and $\text{CLIN}_{\text{FGOALS}}$, are deployed and evaluated.

The grid-scale assessment considers the mean, variance, skewness, minimum, and maximum of climatological distribution at each grid (Fig. 2). These statistics from ERA5 (Fig. 2 Row 1) and FGOALS (Fig. 2 Row 3) generally agree well, due to shared constraints from geophysical laws and geographic boundaries. The key spatial patterns are the latitudinal gradient, the influence of topography (e.g., the Tibetan Plateau), and the land-sea contrast, which are most evident in the mean, minimum and maximum maps. The variance and skewness maps reveal more regional variations. A notable discrepancy is that, compared to ERA5, FGOALS tends to hold larger skewness for most of the land regions in Southern China and Philippine Island, implying a more frequent present of high 2 m temperature for these regions.

$\mathrm{CLIN}_{\mathrm{ERA5}}$ (Fig. 2 Row 2) and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ (Fig. 2 Row 4) can well reproduce the considered statistics of their training data, achieving high spatial correlation coefficient (~ 0.99) and low root mean squared error ($\sim 0.1°\mathrm{C}$) in matching these statistics. Besides reproducing the large scale patterns, both models accurately capture high frequency local variations influenced by complex topography, such as for mountainous regions and coastal areas. Also, the climatology difference between ERA5 and FGOALS are well reproduced by the corresponding CLIN models.

We further carry out grid-wise Kolmogorov-Smirnov tests to assess whether the generated and referential samples are likely to have come from the same underlying distribution: 96/76% grid points (stippled grids in Fig. 2) within the considered region pass a 95% confidence interval test for the $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ model. These results suggest that the CLIN model can well reproduce climatological distribution of its training data at grid scale.

We hereafter compare the referential and generated distributions using field-scale statistics. We first examine the linear spatial structure of the 2 m temperature spatial patterns using a principal component analysis (Supporting Information Fig. S2): we decompose the spatial pattern of the target random variable into a set of orthogonal modes that capture the maximum amount of variance, and compare the spatial modes (Empirical Orthogonal Functions, EOFs), as well as the variance explained by these modes. For EAR5, the first to third leading principal components explained 90/2.7/2.0% of the total variance. While for $\mathrm{CLIN}_{\mathrm{ERA5}}$, the first to third leading principal components explained 91/2.6/1.5% of the total variance, which closely matches results for the ERA5 referential data. More importantly, we obtain spatial correlation coefficient of 0.994/0.990/0.986 between the first to third EOF of EAR5 and $\mathrm{CLIN}_{\mathrm{ERA5}}$. While the spatial modes of FGOALS differs considerably with ERA5, $\mathrm{CLIN}_{\mathrm{FGOALS}}$ closely matches FGOALS: the first to third leading principal components explained 83.6/5.1/2.2% or 83.9/4.9/2.1% of the total variance for FGOALS or $\mathrm{CLIN}_{\mathrm{FGOALS}}$. The spatial correlation coefficient between the first to third EOF of FGOALS and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ are 0.999/0.997/0.994. These results suggest that the CLIN model can well reproduce the linear spatial mode of the considered climatological distribution.

Lastly, we examine the distribution of spatial variability across different spatial scales in the referential/generated dataset: we carry out 2D Fourier transform on the referential/generated samples, and draw the radial averaged squared magnitude of the complex Fourier coefficients as function of wave numbers (Fig. 3). The radially averaged power spectrum density of the considered referential and generated data samples follow a similar power-law scaling, suggesting that the CLIN model can well reproduce the spatial variability across scales.

To sum up, the analysis of both grid-scale and field-scale statistics demonstrates that the CLIN methodology accurately reproduces the essential characteristics and patterns of the climatological distribution present in the training data. We can thereafter leverage this learned climatological prior for the state inference task.

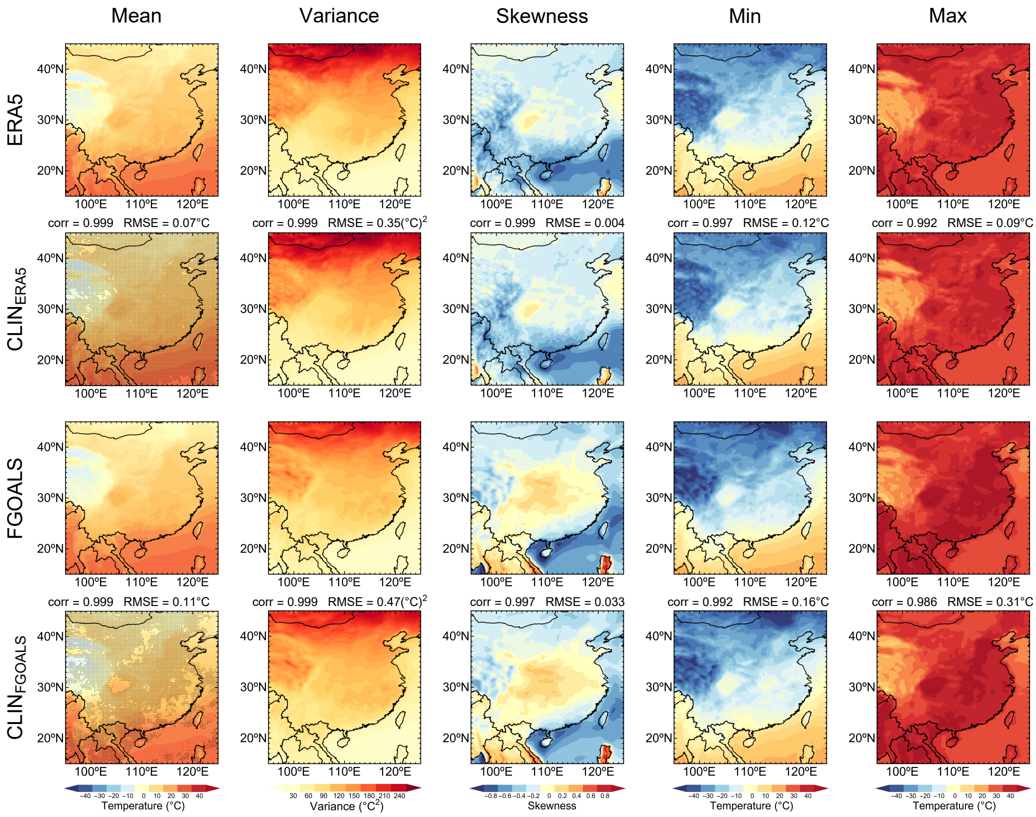


Figure 2. Grid-scale comparison of climatological statistics for climate reanalysis (ERA5, Row 1), climate simulation (FGOALS, Row 3), and probabilistic diffusion models trained using these datasets (CLIN$_{\text{ERA5}}$, Row 2; and CLIN$_{\text{FGOALS}}$, Row 4). The considered statistics are mean, variance, skewness, minimum, and maximum. The spatial correlation coefficient (corr) and root mean squared error (RMSE) between the referential dataset statistics and generated dataset statistics are labeled. Stipples denote grids that pass the Kolmogorov-Smirnov test at 95% confidence interval.

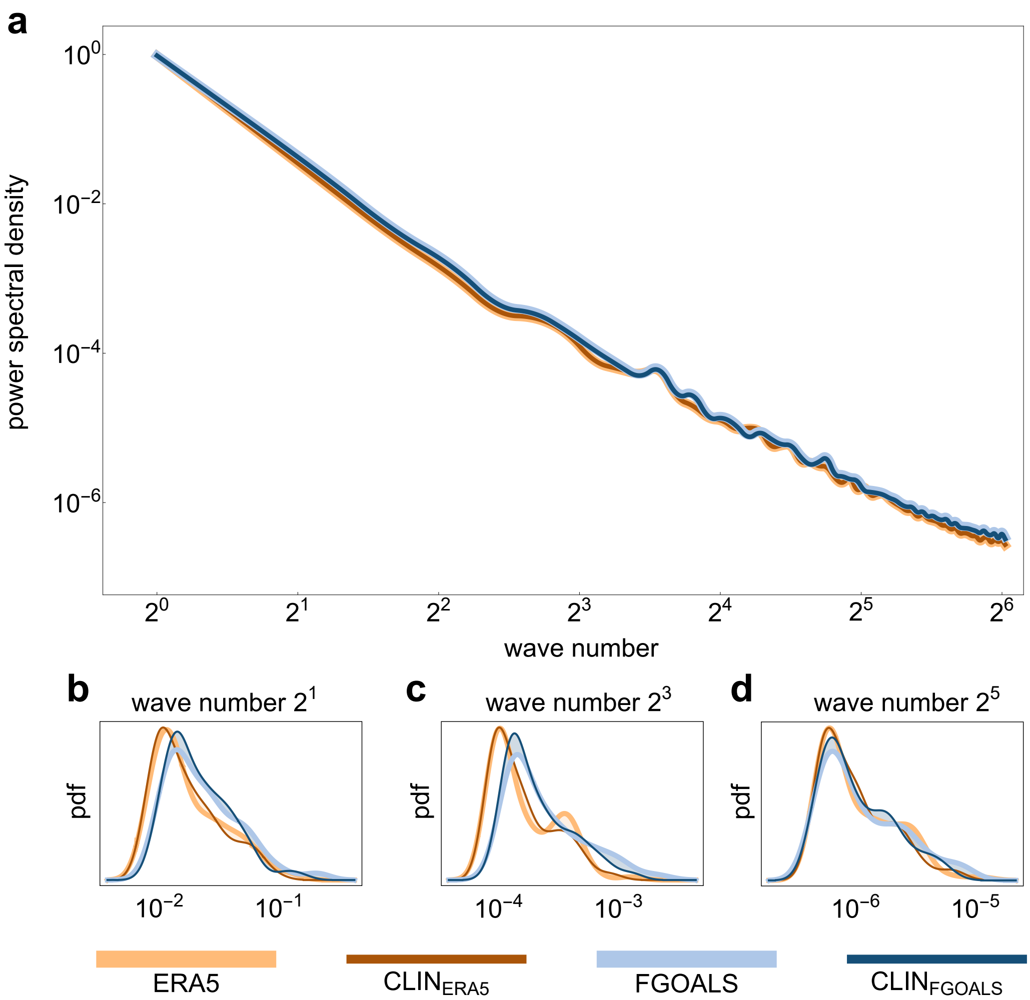


Figure 3. Radial averaged power spectrum density as function of wave number for 2 m temperature spatial pattern. **a**: results for ERA5, FGOALS, $\mathrm{CLIN}_{\mathrm{ERA5}}$, and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ averaged over 100 ensemble members. **b-d**: probability distribution of power spectrum density at wave number 2^1, 2^3, 2^5 for ERA5, FGOALS, $\mathrm{CLIN}_{\mathrm{ERA5}}$, and $\mathrm{CLIN}_{\mathrm{FGOALS}}$.

3.2 Inferring weather states using partial observations

Given a learned climatological prior, we assess how well we can combine it with partial observations to obtain probabilistic estimate of the 2 m temperature spatial patterns. The climatological priors are probabilistic diffusion models trained using climate reanalysis (ERA5) and climate simulation (FGOALS) data. The observations are from 131 operational meteorological stations across China. We randomly select 120 of these stations to inpaint into the generation process, and leave the rest 11 stations for test. For regions without station observations, we consider ERA5 data as benchmark. Below we report case example results (Sec. 3.2.1) and a 1-year round skill assessment (Sec. 3.2.2) .

3.2.1 Case study

We consider four case examples covering different hours of a day and different seasons (Fig. 4). To make probabilistic inference of spatial patterns using partial observations, we gradually inpaint station observations into the generation process of $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$, creating 100 ensemble members for each model and each case. We report the ERA5 spatial pattern (Fig. 4 Row 1), the ensemble mean (Fig. 4 Row 2 and 5), the standard deviation of the ensemble (Fig. 4 Row 3 and 6), the mean squared error between ERA5 and the ensemble members (Fig. 4 Row 4 and 7) for $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$.

Both the repainted $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ ensemble mean results closely match the ERA5 spatial pattern, regarding latitudinal gradient, influence of topography, and the land-sea contrast, yielding spatial correlation coefficient of $0.980\pm0.02/0.977\pm0.02$ for the four considered case examples. These results suggest that the proposed methodology allows effectively propagation of information from limited ($\sim 1\%$) observed locations to a broad range of unobserved parts.

Next, we test if the CLIN methodology offers reliable uncertainty quantification (Fig, 4 Row 3 and 6). A larger ensemble variance indicates greater uncertainty in the estimate, while a smaller variance suggests more confidence in the estimate. As is expected, geogrids close to observation stations tend to have low ensemble variance, while distant ones may have relatively higher ensemble variance. The information constraint from observations may be blocked by topography, such as for Tibetan Plateau and Tian Shan Mountains. While for plain regions, we can expect a larger extension of observation constraints. We further examine the relationship between the spread of the ensemble members and their estimation skill, by computing the correlation between ensemble variance and ensembles' mean squared error score. The high spread skill correlation for $\mathrm{CLIN}_{\mathrm{ERA5}}$ (0.90 ± 0.08) and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ (0.94 ± 0.04) suggest that ensemble spread is a good predictor of model's estimation skill. This means that the CLIN model can capture the underlying uncertainties and provide reliable estimates of spatial estimation confidence.

To sum up, the case studies confirm that the CLIN methodology can make successful probabilistic inference of 2 m temperature spatial patterns using limited observations. The results are spatially coherent, well-constrained by observations, and offer reliable uncertainty quantification. It is worth noting that there are unneglectable mismatches between station observations and ERA5/FGOALS, regarding either climatological statistics or values. These mismatches introduce domain shift error, which is frequently encountered as we deploy a machine learning model in real-world scenarios where the data distribution differs from the training data. Below we dissect this error source by inpainting with different data sources in a 1-year round evaluation.

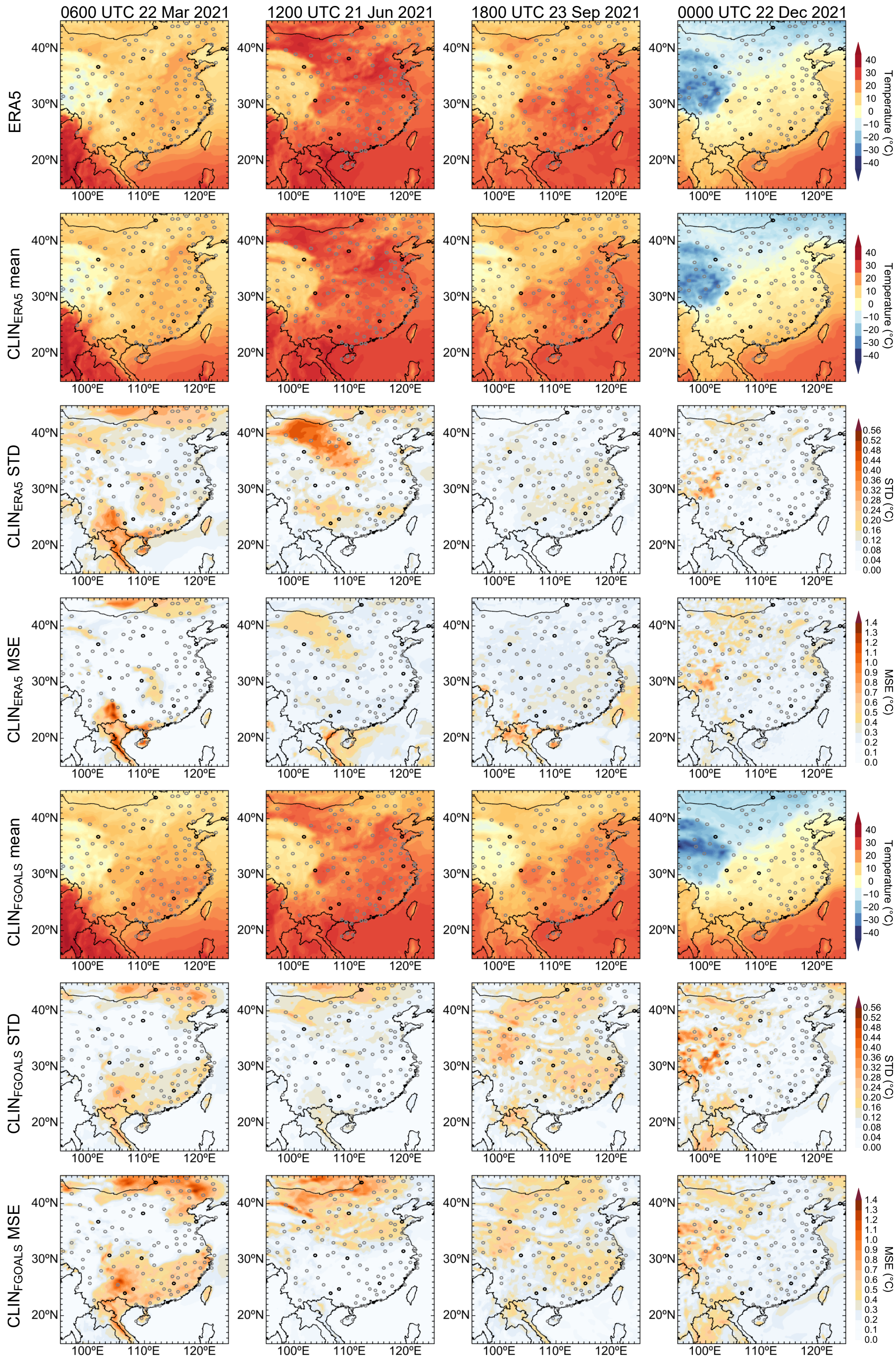


Figure 4. Case examples for probabilistic inference for 2 m temperature spatial pattern using partial observations. For CLIN$_{\mathrm{ERA5}}$ and CLIN$_{\mathrm{FGOALS}}$, 100 ensemble members are created by repainting observations. The ERA5 spatial pattern (Row 1), the ensemble mean (Row 2 and 5), the standard deviation of the ensemble (Row 3 and 6), the mean squared error between ERA5 and the ensemble members (Row 4 and 7) for CLIN$_{\mathrm{ERA5}}$ and CLIN$_{\mathrm{FGOALS}}$ are plotted.

3.2.2 Skill evaluation

We conduct a year-long evaluation of the models' performance in inferring spatial patterns, using data from Year 2021, which are not included in the models' training process. We compare ERA5 with $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$, both inpainted using station observations, and present the spatial distribution of their RMSE in Fig. 5a and Fig. 5b. To further investigate different uncertainty sources in the state inference task, we also consider inpainting $\mathrm{CLIN}_{\mathrm{ERA5}}$ using ERA5 data at the observation stations. The RMSE between this inpainted $\mathrm{CLIN}_{\mathrm{ERA5}}$ and the ERA5 whole-field data is shown in Fig. 5c.

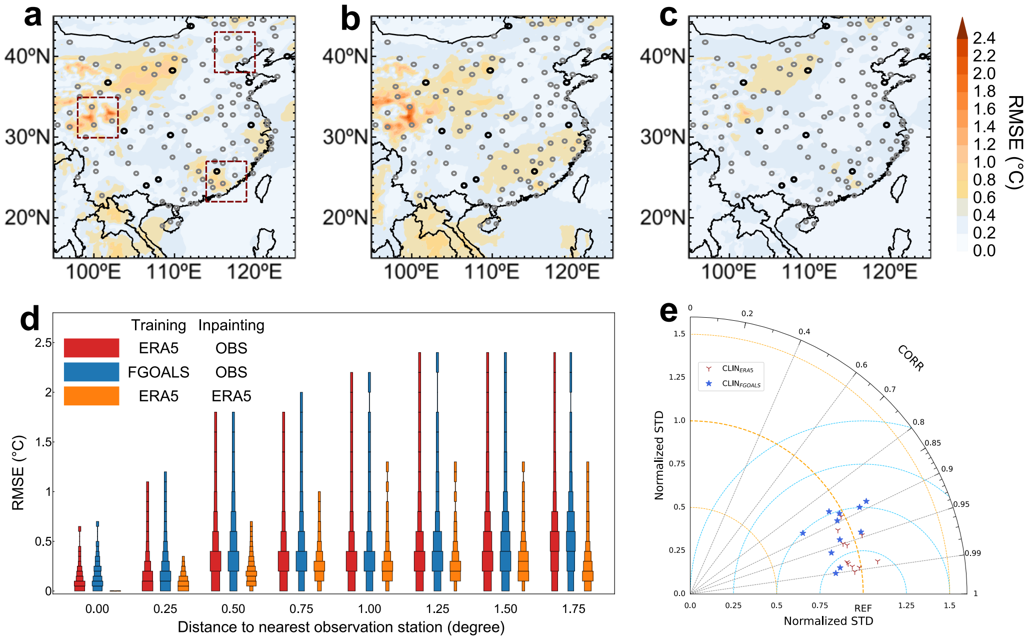


Figure 5. Skill evaluation for CLIN models to estimate spatial pattern of 2 m temperature using data for Year 2021. **a**: root mean squared error (RMSE) between ERA5 reanalysis and $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using station observations; **b**: RMSE between ERA5 reanalysis and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ inpainted using station observations; **c**: RMSE between ERA5 reanalysis and $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using ERA5 data at station observations; **d**: distribution of RMSE as function of grid's distance to nearest observation station for the three considered methods; **e**: Taylor diagram comparing the left-out station observations with $\mathrm{CLIN}_{\mathrm{ERA5}}$ (orange) and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ (blue) results. Both $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ are constrained by 120 station observations here. We delineate three representative regions to evaluate the value of observations in Sec. 3.3

The RMSE between ERA5 and observation inpainted $\mathrm{CLIN}_{\mathrm{ERA5}}/\mathrm{CLIN}_{\mathrm{FGOALS}}$ is $0.25 \pm 0.21°\mathrm{C}/0.31 \pm 0.20°\mathrm{C}$, suggesting that the CLIN methodology enables accurate spatial pattern estimates. Both models exhibit low uncertainty in plain terrain regions or over the ocean, despite that no ocean observations were applied. This suggests that the learned climatological prior effectively captures the spatial patterns and variability in these regions, allowing the models to make confident estimates using limited and far-away observational constraints. On the other hand, both models exhibit higher uncertainty in regions with complex terrain, such as the Tibetan Plateau and the mountainous areas of Southeast China. Additionally, land areas with complicated terrain but lacking observational constraints, such as Southeast Asia, also show large uncertainty in the model estimates.

The uncertainty in state inference comes from the following three sources (Tab. 1). The first is domain shift error, which is due to distribution mismatch among data applied for model training, data applied for inpainting, and data applied for skill evaluation. The second is model error, which is due to the approximation/optimization/statistical error in applying probabilistic diffusion model to fit climatological prior, or due to errors in inpainting. These two types of uncertainties are *epistemic*, as they could be reduced by gathering more data, improving the model, or incorporating knowledge about data distribution differences. The third source of uncertainty is intrinsic/aleatoric, which is due to existence of multiple plausible spatial patterns given partial observational constraints, reflecting the inherent randomness in the system being modeled.

To disentangle these uncertainty sources, we consider the following comparisons.

1. We compare the RMSE of $\mathrm{CLIN}_{\mathrm{ERA5}}$ (Fig. 5a) and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ (Fig. 5b). $\mathrm{CLIN}_{\mathrm{ERA5}}$ achieves an overall lower RMSE, which can be attributed to a relieved domain shift error from the following two aspects: a. compared to FGOALS, ERA5 better matches the "true" climatology as partially revealed by the scattered observations; b. we consider ERA5 data as "ground truth" for evaluating model performance, which gives advantage to CLIN model trained using ERA5 data.
2. We compare $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using observation data (Fig. 5a) and $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using scattered ERA5 data (Fig. 5c). The latter achieve significantly lower RMSE (0.19 ± 0.12°C), suggesting a relatively low model error and a relatively low intrinsic uncertainty of the considered task. The difference between these cases highlights the domain shift error as the observation distribution differs from ERA5.
3. We compare the performance of $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ in predicting the observations at test stations that are excluded during repainting (Fig. 5e). For these test stations, both $\mathrm{CLIN}_{\mathrm{ERA5}}$ and $\mathrm{CLIN}_{\mathrm{FGOALS}}$ results show high correlation coefficient (0.87-0.99) and low root mean squared error (0.2-0.4°C) with the observations , with $\mathrm{CLIN}_{\mathrm{ERA5}}$ performing slightly better than $\mathrm{CLIN}_{\mathrm{FGOALS}}$; $\mathrm{CLIN}_{\mathrm{ERA5}}$ holds a normalized standard deviation close to 1, which closely matches the observations, while $\mathrm{CLIN}_{\mathrm{FGOALS}}$ holds a normalized standard deviation slightly less than 1, suggesting a smaller temporal variability.

Table 1. Uncertainty sources for state inference using partial observations

Uncertainty source	Type	Illustration
Domain shift	Epistemic	Distribution mismatch among data applied for model training, data applied for inpainting, and data applied for skill evaluation.
Model error	Epistemic	1. Approximation/optimization/statistical error in fitting climatological prior. 2. Error in constraining the prior with observations.
Intrinsic uncertainty	Aleatoric	Existence of multiple plausible spatial patterns given observational constraints.

Finally, we quantify the spatial extension of observational constrains by showing models' RMSE skill as function of grid's distance to nearest observation station (Fig. 5d). We consider $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using observation data and ERA5 data, as well as $\mathrm{CLIN}_{\mathrm{FGOALS}}$ inpainted using observation data. For all these cases, models' performances at an arbitrary grid depends closely on the grid's proximity to observations. Meanwhile, there is large variation of models' RMSE skills for grids that are at least 1° away from any observation stations. Below we further investigate the value of individual observations in constraining the variability of its nearby spatial patterns, and offer guidelines for better observation planning.

3.3 On the value of observations

We apply the CLIN methodology to quantify the value of observations in constraining state estimation uncertainty, using three representative regions delineated in Fig. 5a. To achieve this, we add or remove observational stations and evaluate the impact on the estimation error (Fig. 6). Here, the first column shows the RMSE spatial pattern for the original $\mathrm{CLIN}_{\mathrm{ERA5}}$ model estimates in each target region; the second column (Adding) demonstrates the impact of adding an observation station in a high-error area; the third column (Removing) illustrates the effect of removing an existing observation station; the fourth column (Terrain) provides a topographical context for each target region.

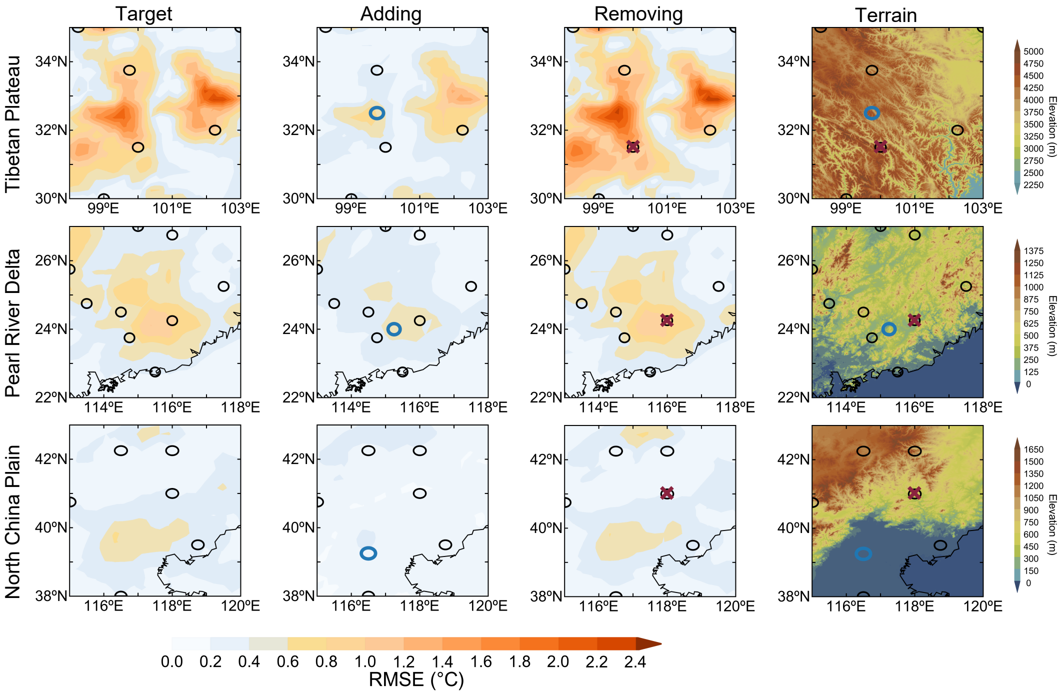


Figure 6. Evaluation of CLIN in reconstructing 2m temperature spatial pattern using different observation setups. Column 1: RMSE between $\mathrm{CLIN}_{\mathrm{ERA5}}$ inpainted using observation data and ERA5 for three selected regions delineated in Fig. 5. Column 2: RMSE after including a pseudo new observation. This new observation data is from ERA5. Column 3: RMSE after including a pseudo new observation. Column 4: elevation map of the considered regions. The results are based on a year-long (Year 2021) evaluation.

For the case of Tibetan Plateau (first row), where the terrain is highly complex, with average elevations exceeding 4500 meters, we obtain a relatively high RMSE given existing observation constrains, particularly in the central and eastern parts of the region. Adding a station in the high-error area significantly reduces the RMSE for a broad range of the considered region, this impact is more pronounced here as compared to the other two cases, highlighting the importance of observational constraints in areas with complex terrain. Removing a station results in a noticeable increase in RMSE in the surrounding areas. Similarly, the effect of station removal is more evident compared to the other two cases, suggesting that the model heavily relies on the limited observational data to constrain its estimates in this complex terrain. The loss of a station in a critical location can greatly impact the model's ability to capture the local temperature patterns.

For the case of Peal River Delta (second row), the terrain is characterized by a mix of lowlands and hilly regions, with elevations ranging from 0 to 1000 meters. The original RMSE is low overall, with some higher values in the central and northwest mountain regions. Adding a station in the high-error area effectively reduces the RMSE. Meanwhile, removing a station leads to a hardly noticeable increase in RMSE in the surrounding areas.

For the case of North China Plain (third row), the northern part is featured by mountainous terrains exceeding 1000 meters, and the southern part has flat topography and homogeneous terrain. Adding a station in the central of southern plain area reduces the RMSE significantly, as existing observations are either from the northern mountain areas, or is too far away. Same as previous case, removing a station has minimal impact on the RMSE distribution.

To sum up, we discusses the application of the CLIN methodology to evaluate the impact of observational data on state estimation uncertainty across three diverse regions. It emphasizes the importance of strategic addition and removal of observational stations in improving estimation accuracy, particularly in areas with complex terrain. The findings highlight how existing observation constraints influence RMSE distribution, with significant reductions observed when stations are added in high-error areas. Conversely, removal of stations leads to increased RMSE, underscoring the model's reliance on limited observational data. Overall, we provide valuable insights for optimizing the design of observation networks, leading to a reduction in uncertainties and biases in weather and climate analysis.

4 Conclusion

Accurate state estimation of Earth atmosphere marks a daunting task due to its high-dimensionality and chaotic nature. We demonstrated the potential of deep generative models, specifically probabilistic diffusion models, in learning the inherent low-dimensional statistical structure of atmospheric circulation from climate reanalysis and simulation data. By leveraging this learned climatological prior, we developed a methodology named CLIN (Climate Inpainting) to effectively infer weather states from partial observations.

For the case study of estimating 2 m temperature spatial patterns, the learned climatological prior accurately reproduced the essential characteristics and patterns of the training data at both grid-scale and field-scale. This learned prior effectively captured multi-scale climate patterns, providing regularization and stability to the state estimation task.

Combining the learned climatological prior with station observations, CLIN yielded strong posterior estimates of 2 m temperature spatial patterns. The estimates were spatially coherent, well-constrained by observations, and provided reliable uncertainty quantification. Regions near observation stations exhibited low ensemble variance, indicating high confidence in the estimates, while distant regions showed relatively higher ensemble variance. The high spread-skill correlation confirmed that the ensemble spread was a good predictor of the model's estimation skill.

Moreover, CLIN allowed us to quantify the value of each observation station in reducing state estimation uncertainty. By adding or removing stations and evaluating the impact on the estimation error, we demonstrated the potential of this approach in guiding the design of optimal observation networks.

Our study showcases the power of deep generative models in extracting and utilizing the information produced by the chaotic evolution of the climate system. The proposed CLIN methodology opens up new opportunities for data-driven weather state estimation, potentially complementing traditional data assimilation approaches.

Future work could focus on extending CLIN to handle indirect observations (i.e., remote sensing) and multiple interdependent variables, incorporating temporal dynamics, and adapting to long-term climate trends. Addressing the computational demands and data requirements of diffusion models is another important direction for making this approach more practical and accessible.

In conclusion, this study demonstrates the immense potential of deep generative models in advancing climate data exploration and tackling complex inference tasks in atmospheric sciences. By learning the intrinsic statistical structure of the climate system, these models can effectively bridge the gap between sparse observations and complete weather state estimates, paving the way for more accurate and efficient climate monitoring and prediction.

5 Data Availability

The ERA5 reanalysis data are obtained from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessible at https://cds.climate.copernicus.eu/.

The FGOALS model data are obtained from the Coupled Model Intercomparison Project Phase 6 (CMIP6), hosted by the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory (LLNL), accessible at https://pcmdi.llnl.gov/CMIP6/.

The observational data are freely available for download from the following website: http://www.ncdc.noaa.gov/oa/ncdc.html. The site information used in this study was obtained from the China Meteorological Data Network, hosted by the China National Meteorological Science Data Center (NMDC), accessible at http://data.cma.cn/.

6 Open Research

Model configuration, analysis scripts, data files used for this study will be publicly available upon accept of the work.

Acknowledgments
This research is supported by National Key R&D Program of China (Grant NoS. 2023YFC3007700 and 2023YFC3007705). We appreciate the insightful discussions with Dr. Niklas Boers from Technical University of Munich, and with Dr. Bin Wang and Dr. Juanjuan Liu from Chinese Academy of Science.

References

Bao, Q., Liu, Y., Wu, G., He, B., Li, J., Wang, L., ... others (2020). Cas fgoals-f3-h and cas fgoals-f3-l outputs for the high-resolution model intercomparison project simulation of cmip6. *Atmospheric and Oceanic Science Letters*, *13*(6), 576–581.

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(5), e535.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, *34*, 8780–8794.

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

Ghil, M. (2020). Hilbert problems for the climate sciences in the 21st century–20 years later. *Nonlinear Processes in Geophysics*, *27*(3), 429–451.

Gilpin, W. (2024). Generative learning for nonlinear dynamics. *Nature Reviews Physics*, 1–13.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., … others (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Holton, J. R., & Hakim, G. J. (2012). *An introduction to dynamic meteorology.* Academic press.

Kadow, C., Hall, D. M., & Ulbrich, U. (2020). Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, *13*(6), 408–413.

Kanngießer, F., & Fiedler, S. (2024). "seeing" beneath the clouds—machine-learning-based reconstruction of north african dust plumes. *AGU Advances*, *5*(1), e2023AV001042.

Kingma, D., Salimans, T., Poole, B., & Ho, J. (2021). Variational diffusion models. *Advances in neural information processing systems*, *34*, 21696–21707.

Law, K., Stuart, A., & Zygalakis, K. (2015). Data assimilation. *Cham, Switzerland: Springer*, *214*, 52.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11461–11471).

Nai, C., Pan, B., Chen, X., Tang, Q., Ni, G., Duan, Q., … Liu, X. (2024). Reliable precipitation nowcasting using probabilistic diffusion models. *Environmental Research Letters*.

Palmer, T. (2017). The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *Journal of Advances in Modeling Earth Systems*, *9*(2), 730–734.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., … Ma, H.-Y. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509.

Pan, B., Wang, L.-Y., Zhang, F., Duan, Q., Li, X., Pan, X., … others (2023). Probabilistic diffusion model for stochastic parameterization–a case example of numerical precipitation estimation. *Authorea Preprints*.

Pan, X., Guo, X., Li, X., Niu, X., Yang, X., Feng, M., … others (2021). National tibetan plateau data center: promoting earth system science on the third pole. *Bulletin of the American Meteorological Society*, *102*(11), E2062–E2078.

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, *44*(24), 12–396.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256–2265).

Song, Y., Garg, S., Shi, J., & Ermon, S. (2020). Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in artificial intelligence* (pp. 574–584).

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Toth, Z., Talagrand, O., Candille, G., & Zhu, Y. (2003). Probability and ensemble forecasts. *Forecast verification: A practitioner's guide in atmospheric science*, *137*, 163.
Wang, B., Zou, X., & Zhu, J. (2000). Data assimilation and its applications. *Proceedings of the National Academy of Sciences*, *97*(21), 11143–11144.
Zhang, G., Ji, J., Zhang, Y., Yu, M., Jaakkola, T. S., & Chang, S. (2023). Towards coherent image inpainting using denoising diffusion implicit models.

Supporting Information for "Learning to infer weather states using partial observations"

Jie Chao[1,2], Baoxiang Pan[2], Quanliang Chen[1], Shangshang Yang[2,3], Jingnan Wang[2,4], Congyi Nai[2,5], Yue Zheng[6], Xichen Li[2], Huiling Yuan[3], Xi Chen[2], Bo Lu[7], Ziniu Xiao[2]

[1]School of Atmospheric Sciences, Chengdu University of Information Technology, Sichuan, China

[2]Institute of Atmospheric Physics, Chinese Academy of Science, Beijing, China

[3]Key Laboratory of Mesoscale Severe Weather, Ministry of Education, and School of Atmospheric Sciences, Nanjing University, Jiangsu, China

[4]College of Computer, National University of Defense Technology, Hunan, China

[5]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

[6]Clustertech LTD, Hong Kong, China

[7]National Climate Center, China Meteorological Administration, Beijing, China

Contents of this file

---

6. Figures S1 Network architecture of diffusion model
7. Figures S2 The first three EOF modes.

S1. Details of probabilistic diffusion model

Here, we provide detailed mathematical formulations and implementation specifics of the deployed probabilistic diffusion model. For more information and useful learning materials, refer to the works of Sohl-Dickstein et al.(2015), Ho et al. (2020), Song et al. (2020) , Kingma et al. (2021), Ho & Salimans (2022) and Luo (2022).

Diffusion models are probabilistic models that describe the evolution of a stochastic process over time. In the context of deep learning diffusion models, the diffusion process and its reverse process are fundamental concepts.

The diffusion process is the forward process through which a model generates data, typically images, from a simple noise distribution (often Gaussian noise) to the target distribution. A step-by-step derivation is provided below.

First, we define the following Gaussian process to transform the target distribution $p(\mathbf{x}_0)$ to a prior distribution $p(\mathbf{x}_T)$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

Here $\mathbf{x}_{t\in[1,T]}$ are latent variables with increasing noise level; $\mathcal{N}$ is Gaussian distribution; $\mathbf{I}$ is identity matrix; β_t is diffusion coefficient, which is pre-defined so that, give large enough T, $p(\mathbf{x}_T|\mathbf{x}_0)$ is drawn close to $p(\mathbf{x}_T)$, which is $\mathbf{x}_0$ agnostic.

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2}$$

We parameterize the Gaussian encoder with mean $\mu_t(x_t) = \sqrt{\alpha_t} x_{t-1}$, and variance $\mathbf{\Sigma}_t(\boldsymbol{x}_t) = (1-\alpha_t)\mathbf{I}$, Here $\alpha_t = 1 - \beta_t$. Mathematically, encoder transitions are denoted as:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \quad (3)$$

$$= \sqrt{\alpha_t \alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} \quad (4)$$

$$= \ldots \quad (5)$$

$$= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \quad (6)$$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad (7)$$

These assumptions depict a systematic process of adding Gaussian noise to the data input over time. As we continue to corrupt the data, it gradually transitions until it is entirely characterized by pure Gaussian noise.

In essence, the reverse process aims to infer the noise distribution that could have generated the observed data. Similar to the diffusion process, the reverse process is represented as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)) \quad p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad (8)$$

Here, $\mathbf{\Sigma}_\theta$ is parameterized as an interpolation between its analytical lower and upper bounds (Dhariwal & Nichol, 2021). The optimization of μ_θ involves maximizing the variational lower bound (ELBO) on the log-likelihood of the training samples (Sohl-Dickstein et al., 2015; Kingma et al., 2021).

Then, diffusion model can be optimized by maximizing the ELBO, which can be derived as follows:

$$\log p(\boldsymbol{x}) = \log \int p(x_{0:T}) dx_{1:T} \tag{9}$$

$$= \log \int \frac{p(x_{0:T}) q(x_{1:T}|x_0)}{q(x_{1:T}|x_0)} dx_{1:T} \tag{10}$$

$$= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \tag{11}$$

$$\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \tag{12}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0) \prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)} \right] \tag{13}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}]} \right] \tag{14}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{0}}(x_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right]$$
$$+ \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right] \tag{15}$$

$$= \mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1) \right] - D_{\mathrm{KL}}(q(x_T|x_0) \parallel p(x_T))$$
$$- \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(x_{t-1}|x_t, x_0) \parallel p_{\boldsymbol{\theta}}(x_{t-1}|x_t)) \right] \tag{16}$$

We now explain the three terms on the right-hand side of the Eq. 16:

- $\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]$ represents the expected log-likelihood of the initial data $\boldsymbol{x}_0$ given the sampled intermediate data $\boldsymbol{x}_1$. For the first step, we have $\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] = 0$.

- $D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))$ denotes the KL divergence between the approximate posterior distribution $q(\boldsymbol{x}_T|\boldsymbol{x}_0)$ and the prior distribution $p(\boldsymbol{x}_T)$ at the final time step $\mathbf{T}$. Where $p(\boldsymbol{x}_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, it implies that $\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T)) \right] = 0$

- $\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \right]$ represents the sum of the expected KL divergences between the approximate posterior distributions $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and the conditional distributions $p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})$ for each intermediate time step $\boldsymbol{t}$ in the reverse diffusion process.

Given the analysis above, maximizing $\log p(\boldsymbol{x})$ can be approximately achieved by minimizing the third term. While minimizing each KL Divergence term individually can be challenging for arbitrary posteriors, we can leverage Bayes' rule to simplify the process:

$$\begin{aligned}
q(x_{t-1}|x_t, x_0) &= \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} && (17)\\
&= \frac{\mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})\mathcal{N}(x_{t-1}; \sqrt{\alpha_{t-1}}x_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})} && (18)\\
&\propto \exp\left\{-\left[\frac{(x_t-\sqrt{\alpha_t}x_{t-1})^2}{2(1-\alpha_t)} + \frac{(x_{t-1}-\sqrt{\alpha_{t-1}}x_0)^2}{2(1-\bar{\alpha}_{t-1})} - \frac{(x_t-\sqrt{\alpha_t}x_0)^2}{2(1-\bar{\alpha}_t)}\right]\right\} && (19)\\
&= \exp\left\{-\frac{1}{2}\left[\frac{(x_t-\sqrt{\alpha_t}x_{t-1})^2}{1-\alpha_t} + \frac{(x_{t-1}-\sqrt{\alpha_{t-1}}x_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(x_t-\sqrt{\alpha_t}x_0)^2}{1-\bar{\alpha}_t}\right]\right\} && (20)\\
&= \exp\left\{-\frac{1}{2}\left(\frac{1}{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\right)\left[x_{t-1}^2 - 2\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}x_{t-1}\right]\right\} && (21)\\
&\propto \mathcal{N}(x_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\alpha_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)}) && (22)
\end{aligned}$$

Hence, it is demonstrated that at each step $\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ follows a normal distribution. We use the KL Divergence between two Gaussian distributions for calculation.

$$
\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}\left(q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right) \| p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)\right) && (23)\\
=&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}\left(\mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)\right) \| \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_q(t)\right)\right) && (24)\\
=&\arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left[\log \frac{\left|\boldsymbol{\Sigma}_q(t)\right|}{\left|\boldsymbol{\Sigma}_q(t)\right|} - d + \operatorname{tr}\left(\boldsymbol{\Sigma}_q(t)^{-1} \boldsymbol{\Sigma}_q(t)\right) + \left(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\right)^T \boldsymbol{\Sigma}_q(t)^{-1}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\right)\right] && (25)\\
=&\arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left[\left(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\right)^T\left(\sigma_q^2(t) \mathbf{I}\right)^{-1}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\right)\right] && (26)\\
=&\arg\min_{\boldsymbol{\theta}} \frac{1}{2 \sigma_q^2(t)}\left[\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\right\|_2^2\right] && (27)
\end{aligned}
$$

After optimizing the Diffusion Model, the sampling procedure simplifies to sampling Gaussian noise from $p(\boldsymbol{x}_T)$ and iteratively running the denoising transitions $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ for T steps to generate a novel $\boldsymbol{x}_0$. In practice, we denote $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ as function of neural network parameterization for $\nabla p(\mathbf{x}_t|\mathbf{x}_0)$, which is commonly known as the *score function* (Y. Song et al., 2020).

S2. CLIN

In our approach, we merge the acquired climatology prior with station observations to deduce the posterior probability distribution of the target variable. This allows us to jointly modify both observed and unobserved regions throughout the denoising steps, yielding generated samples that are spatially coherent, faithful and adaptive to observation constraints, and uncertainty-aware. The specific parameters of CLIN are presented in the following table. S1.

S3. Model parameters schedule

We trained the neural network on the NVIDIA Tesla V100 32GB GPU using CUDA version 12.3. The neural network architecture details of the diffusion model are illustrated in Fig. S1. Typical hyperparameter configurations for diffusion models are often derived from the(Ho & Salimans, 2022).

The specific hyperparameters of the model are presented in the following table. S1.

we embed the time information, and stack the time embedding as an additional channel to all UNet blocks. Each contracting block consists of a long sequence of $\{C_{3*3} + N + ReLU\}_3$ operations and a short sequence of $\{C_{1*1}\}_1$ operations, concatenated as a residual block. Here, C_{n*n} is convolution layer with kernel receptive field of size $n * n$. N is group normalization, ReLU is rectified linear unit function. Each expand block consists of a long sequence of $\{R_2 + C_{3*3} + N + ReLU\}_3$ operations and a short sequence of $\{R_2, C_{1*1}\}_1$ operations, concatenated as a residual block. Here, R_n resizes the data by n times using linear interpolation. We begin with a channel size of 64 and double/shrink the channel size by 2 along each contracting/expanding block.

S4. Evaluation metrics

S4.1 Pearson correlation coefficient (corr)

The Pearson correlation coefficient ($corr$) between prediction $\hat{x}$ and observation x is calculated as follows:

$$corr = \frac{\sum_{i=1}^{n}(\hat{x}_i - \bar{\hat{x}})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(\hat{x}_i - \bar{\hat{x}})^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{28}$$

S4.2 Root mean square error(RMSE)

The root mean square error (RMSE) between prediction $\hat{x}$ and observation x is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{29}$$

S4.3 Empirical Orthogonal Function

Empirical Orthogonal Function (EOF) analysis, also known as Principal Component Analysis (PCA) in some contexts, is a widely used statistical method in various fields, including meteorology, oceanography, climatology, and geophysics.

It is employed to analyze and extract the dominant patterns of variability present in a multivariate dataset, such as spatial patterns in climate data or in oceanographic data. The detailed calculation method for EOF is based on the PrincipalComponents function in Mathematica.

S4.4 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (KS test) is a statistical method used to compare the empirical cumulative distribution function (CDF) of a sample dataset with a reference probability distribution or another sample dataset. It is particularly useful for assessing whether the two datasets are drawn from the same underlying distribution or if they differ significantly.

The KS test operates by computing the maximum difference (or maximum deviation) between the two cumulative distribution functions. This maximum difference, often denoted as the KS statistic (D), represents the largest vertical distance between the empirical CDF and the theoretical (or reference) CDF. The KS statistic is then compared against critical values from the Kolmogorov-Smirnov distribution, which depends on the sample size and the significance level chosen for the test.

The specific computation method for the Kolmogorov-Smirnov test is derived from Mathematica's KolmogorovSmirnovTest function.

S4.5 Power spectrum density

The radial averaged power spectrum density (PSD) is a quantitative measure used in various fields of science and engineering, including signal processing, optics, and geophysics. It provides valuable insights into the distribution of power across different spatial frequencies in a given signal or image. In this paper, the PSD is calculated by first computing the Fourier transform of the signal or image to obtain its frequency domain representation. The power spectrum density is then computed as the squared magnitude of the Fourier transform. The PSD further averages the power spectrum density over concentric circles or spherical shells centered at the origin, hence the term "radial averaged." This averaging process is performed to capture the isotropic characteristics of the signal or image, ensuring that contributions from all directions are considered equally.

The PSD is particularly useful for analyzing signals or images with rotational symmetry or spatial periodicity. By averaging the power spectrum density radially, it becomes possible to discern patterns or structures that are not readily apparent in the original signal or image. Additionally, the PSD can be used to quantify the dominant spatial frequencies present in the signal or image, providing valuable information for further analysis or interpretation.

In short, the radial averaged power spectrum density offers a comprehensive view of the spatial frequency content of a signal or image, facilitating insights into its underlying structure and characteristics.

The specific calculation method for the PSD is derived from the pySTEPS library in Python.

References

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis.

Advances in neural information processing systems, *34*, 8780–8794.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Kingma, D., Salimans, T., Poole, B., & Ho, J. (2021). Variational diffusion models. *Advances in neural information processing systems*, *34*, 21696–21707.

Luo, C. (2022). Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256–2265).

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Table S1. Hyperparameters of Diffusion model

Hyperparameter	**Setting**	**Parameter**
Learning Rate	10^{-4}	$\alpha_t^2 = 1 - \sigma_t^2 = \frac{1}{1+e^{-\lambda_t}}$
Batch Size	64	$\lambda_t = -2 \log \tan(at + b)$
Channel	64	$b = \arctan(e^{-\frac{\lambda \max}{2}})$
Optimizer	Adam	$a = \arctan(e^{-\frac{\lambda \min}{2}}) - b$
Number of Iterations	1000	$t = \frac{i}{1000}$, Where $i = 0, 1, 2, \ldots, 1000$
$\lambda \min$	-20	$\text{embedding}(t) = [\sin(2\pi\omega t); \cos(2\pi\omega t)]$
$\lambda \max$	20	$\omega \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

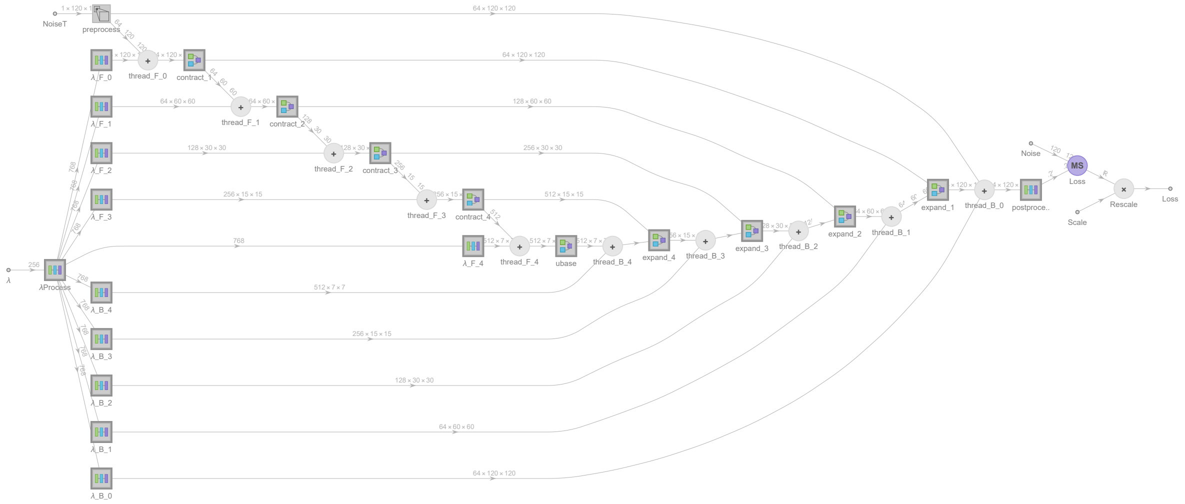


Figure S1. Network Architecture of diffusion model. Each contracting block consists of a long sequence of $\{C_{3*3} + N + ReLU\}_3$ operations and a short sequence of $\{C_{1*1}\}_1$ operations, concatenated as a residual block. Here, C_{n*n} is convolution layer with kernel receptive field of size $n*n$. N is group normalization, ReLU is rectified linear unit function. Each expand block consists of a long sequence of $\{R_2 + C_{3*3} + N + ReLU\}_3$ operations and a short sequence of $\{R_2, C_{1*1}\}_1$ operations, concatenated as a residual block.

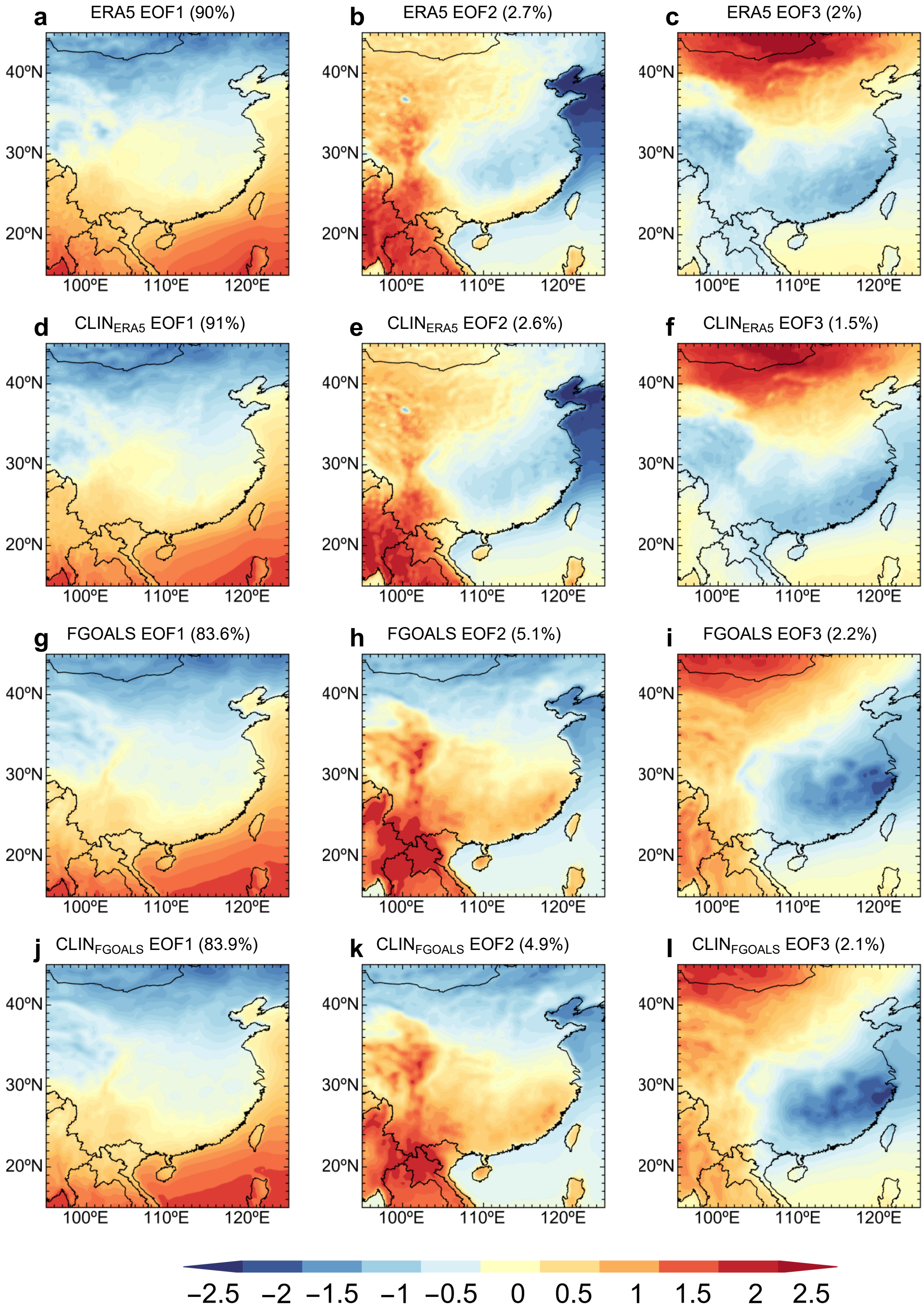


Figure S2. The first three EOF modes. ERA5 (a-c), *Clin*_ERA5 (d-f), FGOALS (g-i) and CLIN_FGOALS (j-l).