VegeDiff: Latent Diffusion Model for Geospatial Vegetation Forecasting

Sijie Zhao, Hao Chen*, Xueliang Zhang*, Pengfeng Xiao, and Lei Bai

***Abstract*—In the context of global climate change and frequent extreme weather events, forecasting future geospatial vegetation states under these conditions is of significant importance. The vegetation change process is influenced by the complex interplay between dynamic meteorological variables and static environmental variables, leading to high levels of uncertainty. Existing deterministic methods are inadequate in addressing this uncertainty and fail to accurately model the impact of these variables on vegetation, resulting in blurry and inaccurate forecasting results. To address these issues, VegeDiff is proposed for the geospatial vegetation forecasting task. To our best knowledge, VegeDiff is the first to employ a diffusion model to probabilistically capture the uncertainties in vegetation change processes, enabling the generation of clear and accurate future vegetation states. VegeDiff also separately models the global impact of dynamic meteorological variables and the local effects of static environmental variables, thus accurately modeling the impact of these variables. Extensive experiments on geospatial vegetation forecasting tasks demonstrate the effectiveness of VegeDiff. By capturing the uncertainties in vegetation changes and modeling the complex influence of relevant variables, VegeDiff outperforms existing deterministic methods, providing clear and accurate forecasting results of future vegetation states. Interestingly, this study demonstrate the potential of VegeDiff in applications of forecasting future vegetation states from multiple aspects and exploring the impact of meteorological variables on vegetation dynamics. The code of this work will be available at *https://github.com/walking-shadow/Official_VegeDiff*.**

***Index Terms*—Latent diffusion model, Geospatial vegetation forecasting, Variational Autoencoder, High resolution, Remote sensing**

I. Introduction

GEOSPATIAL forecasting on Earth involves analyzing historical data and related influential factors of the Earth's surface to identify changing patterns and forecast future states. In the context of global climate change and frequent extreme weather events [1, 2, 3, 4], understanding potential changes and future states, such as land use/land cover transformations [5, 6], future crop yields [7, 8], and vegetation growth [9, 10], is crucial for policy-making and decision-making. The Earth's surface exhibits significant complexity and variability, with numerous factors such as meteorological and topographic variables playing pivotal roles [11, 12]. Minor historical differences can lead to vastly different future states, indicating a high degree of uncertainty in geospatial changes.

The rapid growth of Earth observation data has made it feasible to apply deep learning for geospatial forecasting [13, 14, 15]. However, research in this area remains scarce, with most studies relying solely on historical geospatial data for future forecasting [16]. This approach is insufficient because geospatial changes are heavily influenced by various factors, making it challenging for models to learn patterns solely from geospatial state variables. Therefore, it is essential to incorporate related variables when forecasting geospatial changes.

Given the urgent challenges posed by global climate change, understanding the evolution of the Earth's surface is paramount. Meteorological variables, which inherently exhibit diurnal and other temporal variations (i.e., dynamic temperature and precipitation patterns), play a crucial role in influencing geospatial vegetation changes [17, 18]. When these dynamic meteorological factors interact with static environmental conditions (e.g., topography), they can immensely impact vegetation distribution and dynamics. The variability and complexity of geospatial vegetation changes are highly pronounced, necessitating models to capture uncertainties in the change process. On the one hand, the future geospatial vegetation states are influenced by historical vegetation states and a multitude of related variables. Minor differences in historical vegetation and related variables can be amplified during the change process, leading to significant disparities in future vegetation states. On the other hand, meteorological variables, topographic variables, and other related factors have a significant impact on the vegetation change process. These variables interact differently under various historical vegetation states, highlighting the complexity of the vegetation change process.

Using the geospatial vegetation forecasting task introduced in EarthNet2021X [19], the vegetation state is forecasted under the constraints of dynamic meteorological variables (wind speed, relative humidity, shortwave downwelling radiation, rainfall, sea-level pressure, and temperature (daily mean, min & max)) and static environmental variables (digital elevation model and land cover), as shown in Figure 1 (a). Specifically,

This work was supported in part by the Shanghai Artificial Intelligence Laboratory, in part by the National Natural Science Foundation of China under Grant 42071297, in part by the AI and AI for Science Project of Nanjing University (Grant No. 020914380141), in part by the Fundamental Research Funds for the Central Universities under Grant 020914380119, and in part by the Youth Innovation Team of China Meteorological Administration (CMA2024QN02).

Sijie Zhao, Xueliang Zhang, and Pengfeng Xiao are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: zsj@smail.nju.edu.cn; zxl@nju.edu.cn; xiaopf@nju.edu.cn).

Hao Chen, and Lei Bai are with the Shanghai Artificial Intelligence Laboratory, Shanghai 200000, China (e-mail: chenhao1@pjlab.org.cn, bailei@pjlab.org.cn).

Corresponding Author: Hao Chen and Xueliang Zhang.

two dynamic variables are utilized: high-resolution geospatial vegetation states and low-resolution dynamic meteorological variables from time 1 to $T + K$, accompanied by static environmental variables corresponding to the geospatial vegetation states. The task paradigm involves forecasting the geospatial vegetation states from time $T + 1$ to $T + K$, given the geospatial vegetation states from time 1 to T, dynamic meteorological variables from time 1 to $T + K$, and the static environmental variables. Geospatial vegetation states refer to the spatial distribution of vegetation on the Earth's surface. Dynamic meteorological variables are characterized by their continuous changes, primarily reflecting daily fluctuations in this context. In contrast, static environmental variables change very slowly and can be approximated as static within this paradigm. Therefore, geospatial vegetation forecasting aims to predict how the overall spatial distribution of geospatial vegetation will change in the future, influenced by the interplay between relatively constant static environmental variables and the daily dynamic variations of meteorological variables.

Based on this paradigm, many studies have utilized deep learning approaches for forecasting geospatial vegetation and have achieved commendable results [20, 19, 21, 22]. However, three main issues prevent these models from effectively forecasting future vegetation states, as shown in Figure 1 (b) : 1) Difficulty in handling the high uncertainty of vegetation changes. The complexity and variability of geospatial vegetation changes mean that minor differences in historical vegetation and related variables can lead to substantial differences in future states, requiring the ability of models to capture the high uncertainty of vegetation changes. However, these models are deterministic models which would generate blurry and inaccurate vegetation forecasting results that lack crucial details. 2) Failure to utilize vast amounts of remote sensing data. These deterministic models are trained on the limited data relevant to the vegetation forecasting task, overlooking the plethora of available remote sensing data that could enhance the model's understanding of vegetation states. 3) Inadequate approaches to address the causal relationships of related variables on vegetation changes. These deterministic models simply concatenate remote sensing images, dynamic meteorological variables, and static environmental factors, feeding them into the model without properly modeling the complex interactions among these variables on vegetation changes.

To address the aforementioned issues, a probabilistic model based on latent stable diffusion is introduced for geospatial vegetation forecasting, aiming to model the high uncertainties in vegetation changes and generate accurate and clear vegetation forecasting results. VegeDiff is proposed to forecast geospatial vegetation changes, leveraging the latent space of a well-trained vegetation autoencoder to represent vegetation states and appropriate approaches to model the influence of related variables on vegetation changes, as depicted in Figure 1 (c).

Specifically, 1)Diffusion model is firstly introduced to model the high uncertainty in geospatial vegetation forecasting probabilistically. Currently, diffusion models are primarily employed in meteorological applications in Earth forecasting and have demonstrated superior performance in precipitation and weather forecasting, but their application in geospatial vegetation forecasting tasks is lacking. By employing the diffusion model in vegetation forecasting, the denoising processes of this probabilistic model is leveraged to model the high uncertainties in vegetation changes, thereby enabling the capture of multiple potential futures of geospatial vegetation states and producing accurate and clear forecasting results. 2) A vegetation autoencoder is designed to obtain a well-represented latent space for geospatial vegetation states. Since combinations of blue, green, red, and near-infrared channels (RGBN) can calculate many vegetation indices that effectively indicate geospatial vegetation states, RGBN remote sensing images are utilized to represent geospatial vegetation states. Current variational autoencoder [23] models are trained on vast amounts of natural RGB images, which are not suitable for RGBN remote sensing images in terms of channel and scene adaptation. Therefore, a variational autoencoder is pre-trained on a large amount (10M) of RGBN remote sensing images and fine-tuned it on a relatively small amount (20K) of RGBN vegetation remote sensing data, which enhances the ability of latent space to represent geospatial vegetation states, facilitating better forecasting in the latent space. 3) VegeNet is proposed to model the effects of dynamic meteorological variables and static environmental variables on vegetation states. Historical vegetation states undergo complex transformations to form future vegetation states under the influence of dynamic meteorological variables and static environmental variables. VegeNet models both the local effects of static environmental variables and the global effects of dynamic environmental variables on vegetation states, structuring the model according to the causal relationships affecting geospatial vegetation changes.

Overall, the principal contributions of our work are as follows:

1) A probabilistic model is firstly introduced into the geospatial vegetation forecasting task. VegeDiff employs the diffusion process to model the uncertainties in the vegetation change process, capturing multiple potential futures of geospatial vegetation states and generating clear and accurate forecasting results.
2) A vegetation autoencoder is designed to achieve a robust representation of geospatial vegetation states. This vegetation autoencoder was pre-trained on 10M RGBN remote sensing data and fine-tuned on 20K remote sensing vegetation data, enabling its latent space to effectively represent the geospatial vegetation states.
3) VegeNet is designed to model the impact of static environmental and dynamic meteorological variables on geospatial vegetation changes. VegeNet decouples the effects of static environmental variables and dynamic meteorological variables on the geospatial vegetation change process, effectively modeling the transformation process of vegetation under the influence of these variables.

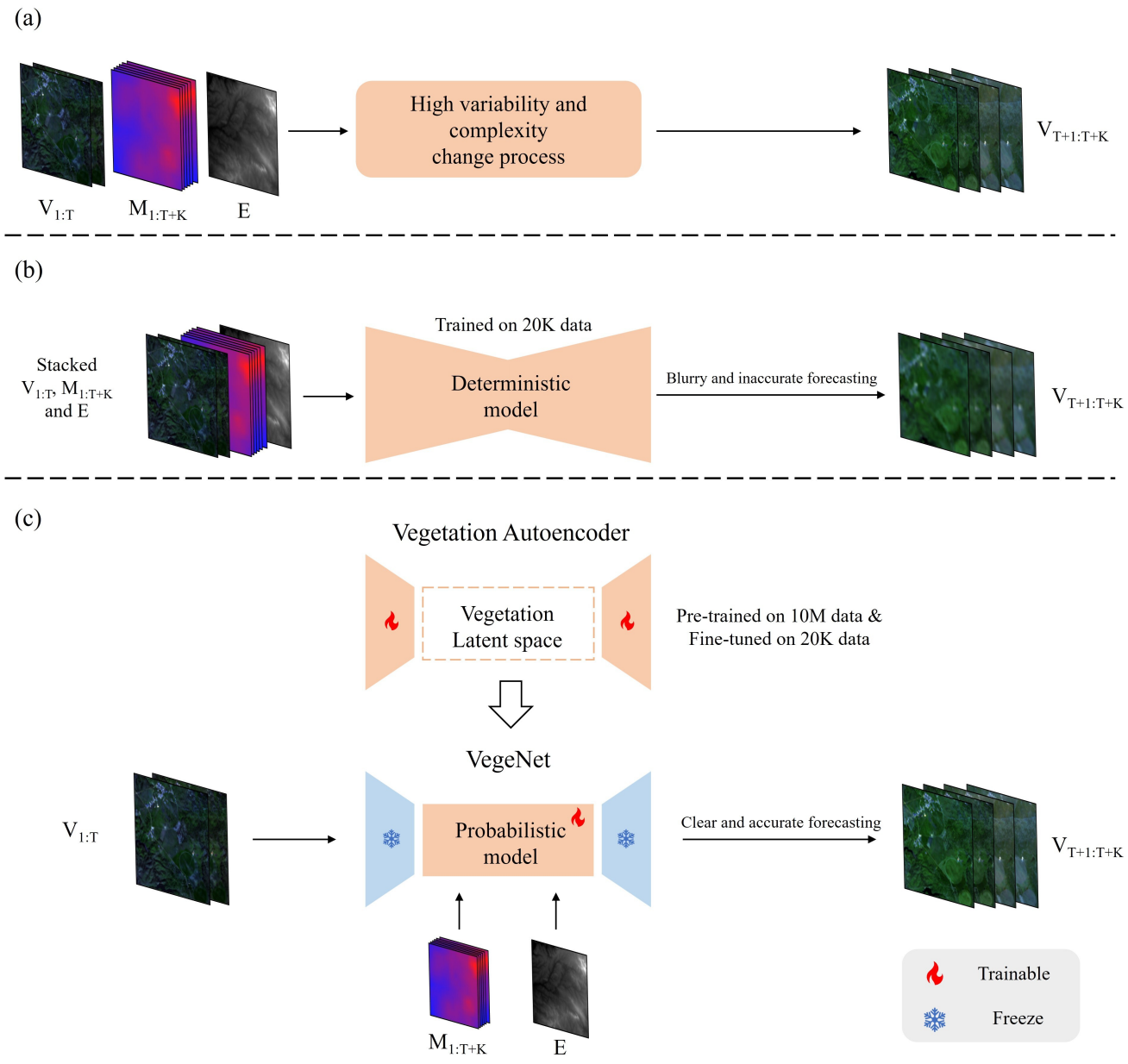

Fig. 1. Illustration of the geospatial vegetation forecasting task and models, where V indicates geospatial vegetation states, M indicates dynamic meteorological variables, and E indicates static environmental variables. (a) The overview of the geospatial vegetation forecasting task. (b) The overview of deterministic models performing the geospatial vegetation forecasting task. (c) The overview of VegeDiff performing the geospatial vegetation forecasting task.

II. Related Works

A. Geospatial Vegetation Forecasting

Geospatial vegetation forecasting seeks to forecast the future changes in both the spatial distribution of geospatial vegetation, which is shaped by the interaction between stable static environmental factors and the daily fluctuations of meteorological variables [20]. Geospatial vegetation state emphasizes the spatial distribution and detailed patterns of vegetation as captured by the remote sensing images. In recent years, with the surge in time series data of meteorological variables and remote sensing images [24, 25], many deep learning-based methods have been applied to this task, achieving superior performance [20, 19, 21, 22]. Requena-Mesa et al. [20] advocated viewing this task as a guided video forecasting task and have constructed a dataset specifically for geospatial vegetation forecasting called EarthNet2021. Diaconu et al. [21] used a ConvLSTM-based model on this dataset and conducted ablation experiments to validate the effectiveness of dynamic meteorological and static environmental variables in forecasting future geospatial vegetation states. Robin et al. [26] developed a dataset for climatically volatile regions of Africa to investigate the impact of extreme weather on geospatial vegetation states. Benson et al. [19] improved the EarthNet2021 dataset to the EarthNet2021X dataset, focusing more on areas with drastic vegetation changes.

However, the deep learning models used in these stud-

ies are deterministic and face three key limitations. This study propose to use diffusion models to forecast geospatial vegetation changes probabilistically. As probabilistic models, diffusion models can use diffusion processes to effectively model uncertainty in the geospatial vegetation change process. First, they struggle to capture the high uncertainty inherent in vegetation changes. Second, while vast quantities of high-resolution RGBN remote sensing data are available for detailed vegetation patterns, existing methods do not fully exploit this rich data source. Third, the simple concatenation of remote sensing imagery with dynamic and static variables overlooks the relationships between geospatial vegetation states and these variables. In contrast, our approach advances the state of the art via three approcaches: (i) firstly intorducing a diffusion model that probabilistically models multiple potential future states to better represent uncertainty; (ii) developing a novel vegetation autoencoder pre-trained on a massive set of RGBN images—which is then fine-tuned on vegetation-specific data—to construct an effective latent representation of vegetation states; and (iii) designing VegeNet to explicitly disentangle and model the local effects of static environmental factors and the global impacts of dynamic meteorological variables. This combined strategy not only provides a robust probabilistic forecasting framework but also leverages the full potential of available remote sensing data while addressing the relationships between geospatial vegetation change and related variables.

B. Diffusion Models for Earth Forecasting

Diffusion models (DMs) [27] have emerged as a potent framework for generating high-quality images through a process known as stochastic denoising [28, 29, 30, 31]. These models operate by gradually transforming a distribution of random noise into a distribution of images, closely resembling the target data distribution. The process involves an iterative procedure in which an initial noise image is progressively denoised through a series of steps, guided by a neural network that has been trained to perform this transformation effectively.

Building upon the foundational principles of diffusion models, latent diffusion models (LDMs) [32] introduce significant advancements by operating in a compressed, latent space rather than directly in the pixel space [33, 34, 35]. This modification brings forth several key improvements. First, by operating in a lower-dimensional latent space, these models can achieve faster convergence and require less computational resources, making the generation process more efficient. Second, latent space diffusion models have demonstrated an enhanced ability to capture and reproduce the complex, high-level semantics of the target distribution, leading to the generation of images with superior quality and greater detail. This is primarily because the latent space provides a more abstract representation of the data, enabling the model to focus on the underlying structure and semantics rather than pixel-level details.

As the effectiveness of LDMs in generating images has been proven, their application has extended to video generation [36, 37, 38, 39, 40]. PVDM [38] introduces a method of projecting videos into a low-dimensional latent space represented as 2D vectors, facilitating simultaneous training for both unconditional and frame-conditional video generation. LFDM [39] utilizes a flow forecastor for estimating latent flows between video frames, thereby training an LDM to generate temporal latent flows. VideoFusion [40] separates the transition noise in LDMs into individual frame noise and temporal noise and synchronously trains two networks to accurately represent this noise decomposition.

Due to the effectiveness of LDMs in video generation, some studies have employed LDMs for Earth forecasting tasks [41, 42, 43, 44]. LDCast [43] introduces a latent diffusion model for precipitation nowcasting, highlighting its capability for effective uncertainty quantification. Prediff [44] developed a conditional latent diffusion model for the same purpose, incorporating an explicit knowledge control mechanism to ensure that forecasts align with domain-specific physical constraints.

However, the application of LDMs in Earth forecasting has primarily focused on meteorological contexts, with their potential in other areas of Earth forecasting remaining relatively unexplored. This presents an opportunity for significant advancements. Therefore, this study proposed the use of LDMs for geospatial vegetation forecasting, leveraging their robust modeling capabilities to forecast the dynamic processes of vegetation change on Earth's surface.

III. Methodology

A. Preliminary: Latent Diffusion Models

Latent diffusion models (LDMs) represent a groundbreaking development in the field of generative modeling, particularly within the domain of computer vision. LDMs leverage the concept of diffusion processes, traditionally utilized in thermodynamics and statistical mechanics, to model the generation of complex data distributions through a series of gradual, probabilistic transformations in a latent space. This approach is distinguished by its capacity to model and manipulate high-dimensional data distributions with unprecedented precision and versatility.

The core operation of LDMs revolves around the iterative application of a forward diffusion process, which gradually adds noise to the data in the latent space over a series of time steps, transforming the data distribution from its original, complex form to a simpler, noise-dominated distribution. This is mathematically represented as:

$$x_t = \alpha_t x_{t-1} + (1 - \alpha_t)\epsilon \tag{1}$$

where x_t represents the data at step t, α_t is a coefficient determining the amount of noise to add, and ϵ is the noise vector sampled from a standard Gaussian distribution. The reverse process, or the denoising phase, aims to reconstruct the original data from the noise by iteratively estimating the noise component and subtracting it from the noisy data, effectively inverting the diffusion process. The denoising process is often modeled with a neural network that learns to forecast the noise, $\hat{\epsilon}$, added at each step, thus enabling the recovery of the clean data:

$$x_{t-1} = \frac{1}{\alpha_t}(x_t - \frac{1 - \alpha_t}{\alpha_t}\hat{\epsilon}(x_t, t)) \tag{2}$$

The latent space in LDMs plays a pivotal role, serving as a compact and computationally efficient representation of the data, which significantly enhances the model's ability to handle high-dimensional inputs without the exponential increase in computational demand typically associated with such tasks. This efficiency is partly due to the reduced dimensionality of the latent space, which also tends to capture the most salient features of the data, thus facilitating a more focused and effective diffusion and denoising process.

Traditional methods such as CNNs and transformer-based models have achieved impressive performance in various forecasting tasks through effective feature extraction and modeling complex spatial-temporal dependencies. However, these deterministic approaches often fall short when dealing with tasks characterized by high inherent uncertainty, such as geospatial vegetation forecasting under the effects of dynamic meteorological variables and static environmental variables.

In contrast, latent diffusion models (LDMs) provide a natural framework for probabilistic modeling by explicitly simulating the evolution of data through a stochastic diffusion process. Unlike CNNs or transformers that generate deterministic outputs, LDMs iteratively add and remove noise, thereby capturing the underlying aleatoric uncertainty in the vegetation change process. This property is particularly beneficial in our context, where the vegetation state is influenced by both dynamic meteorological variables and static environmental factors.

Moreover, the iterative denoising process of LDMs can be interpreted as a gradual refinement of data representations, which aligns with the physical process of vegetation change under continuously varying external conditions. In contrast, deterministic models such as CNNs may produce blurry or averaged predictions when faced with multimodal future states, and attention-based mechanisms, while effective in modeling spatial dependencies, do not inherently model uncertainty in the same principled probabilistic fashion.

Therefore, the adoption of a latent diffusion framework in the study is well motivated by its ability to capture diverse plausible future vegetation states through an explicit probabilistic formulation, leading to clearer and more accurate forecasting results.

B. Overall Structure of VegeDiff

VegeDiff is composed of a vegetation autoencoder, a diffusion process, and a denoising process, as depicted in Figure 2. The training process of VegeDiff involves two main parts: training the vegetation autoencoder and training the VegeNet.

To efficiently forecast future geospatial vegetation, a vegetation autoencoder is initially trained on 10M RGBN remote sensing data. This variational autoencoder provides a latent space with a robust representation of geospatial vegetation features and allows for more efficient forecasting at a lower image resolution in the latent space.

In the context of diffusion models, the denoising process applied to noisy images is inherently probabilistic. This characteristic allows the denoising model to effectively simulate the inherent uncertainties in geospatial vegetation changes. By iteratively refining the noisy input through a series of probabilistic steps, the denoising model captures the complex and stochastic nature of vegetation dynamics. The denoising model, VegeNet, is trained with all parameters of the vegetation autoencoder frozen due to its completed training. Given the high uncertainty in the vegetation change process, our approach diverges from past research that used deterministic models to forecast the future state of geospatial vegetation. Instead, a diffusion model is employed to forecast the future geospatial vegetation state. As probabilistic models, they are capable of effectively modeling the uncertainties inherent in geospatial vegetation changes.

VegeDiff forecasts the future state of geospatial vegetation based on past vegetation state, utilizing dynamic meteorological variables and static environmental variables, as illustrated in Figure 2. Specifically, past remote sensing images from time 1 to T, denoted as $V_{1:T}$, are processed through the vegetation autoencoder to obtain latent space features $Z_{1:T}$. Since the distribution of geographical features in remote sensing images generally remains constant and vegetation changes are closely related to past states, future states are generated based on these past vegetation states rather than from pure noise. Therefore, in the diffusion process, the past latent space features $Z_{1:T}$ are averaged over the temporal dimension and combined with Gaussian noise weighted by weight w to produce the noisy features $Noise_{T+1:T+k}$, which are then concatenated with $Z_{1:T}$ to form the latent space features $Z_{1:T+K}$.

As a probabilistic model, VegeNet can effectively model the high uncertainty in vegetation changes and accurately forecast future vegetation states. Specifically, in the denoising process, VegeNet leverages past remote sensing image features $Z_{1:T}$ and incorporates meteorological features $M_{1:T+k}$ and static environmental features E, progressively denoising the noise features $Noise_{T+1:T+k}$ to forecast future latent space features $Z_{T+1:T+K}$. The future features $Z_{T+1:T+K}$ are finally decoded by the vegetation autoencoder, reconstructing the future remote sensing images $V_{T+1:T+k}$, thus providing forecasting results of future geospatial vegetation states.

C. Vegetation Autoencoder

Due to the high complexity of geospatial vegetation changes, it is necessary to perform vegetation forecasting within a feature space that accurately represents the geospatial vegetation states. Many vegetation indices, which represent the state of geospatial vegetation, can be calculated using the blue, green, red, and near-infrared channels. Therefore, this study choose to use RGBN remote sensing images to construct a feature space that effectively represents vegetation states. To develop such a feature space, extensive data are required to train the model. Unlike models trained solely on geospatial vegetation forecasting datasets, the vegetation autoencoder is pre-trained based on the variational autoencoder [23] of latent diffusion models [32] with 10M RGBN remote sensing data to create a robust feature space representative of RGBN images. Subsequently, the model is fine-tuned with 20K remote sensing data from the geospatial vegetation forecasting dataset, further refining the feature space to accurately represent various

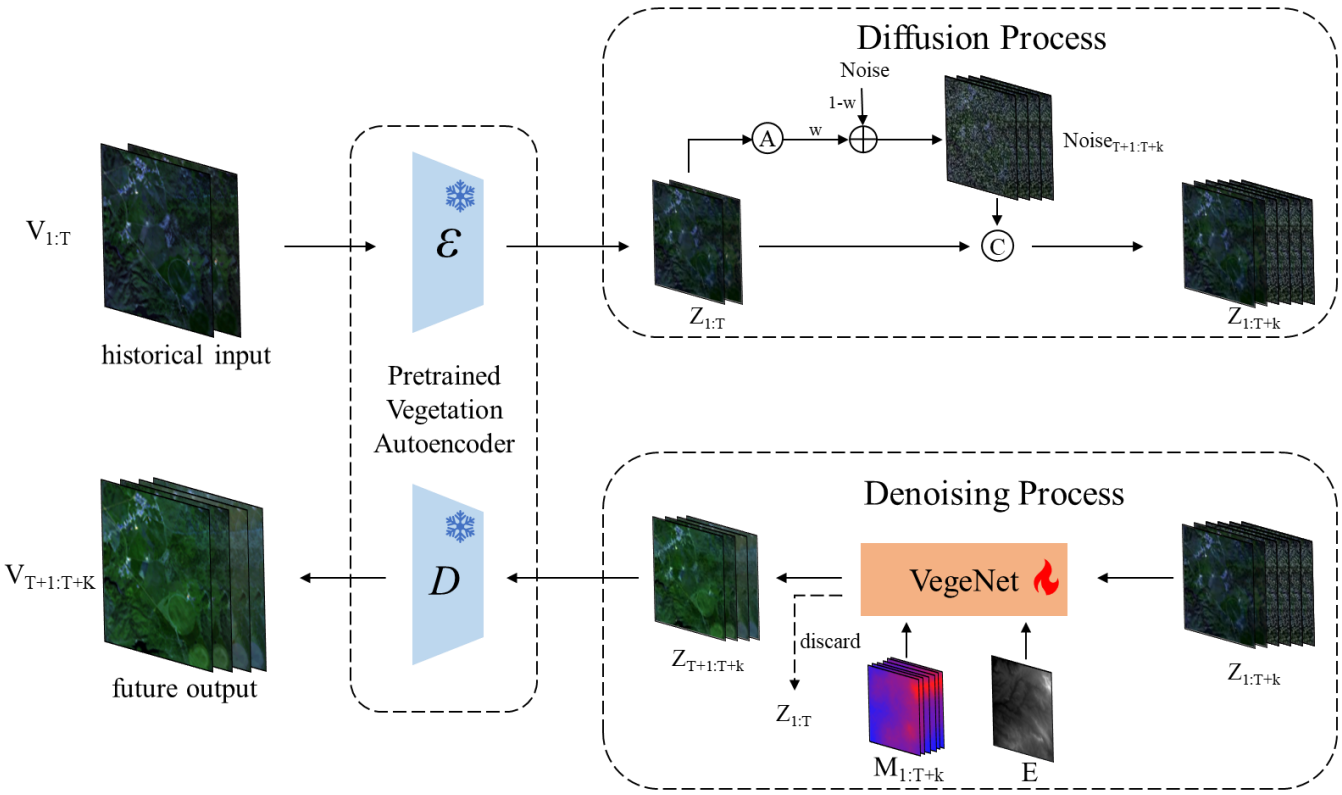

Fig. 2. The overall structure of VegeDiff. VegeDiff models the high uncertainty of geospatial vegetation change process with diffusion model.

vegetation states. Through the two-stage training process, the vegetation autoencoder can learn the characteristics of RGBN remote sensing images from a massive dataset, forming a vegetation latent space that effectively represents various geospatial vegetation states.

D. VegeNet

The process of geospatial vegetation change is complexly influenced by dynamic meteorological variables and static environmental variables. Therefore, modeling the effects of these variables is crucial for forecasting geospatial vegetation. Simply concatenating all variables and inputting them into the model is not advisable, as it would confuse the impact of dynamic meteorological and static environmental variables on geospatial vegetation states. Thus, VegeNet is introduced, which models the complex effects of dynamic meteorological variables and static environmental variables on past vegetation states to accurately forecast future geospatial vegetation states.

VegeNet is based on the DiT [45] architecture and decouples the effects of meteorological and environmental variables, enabling accurate future vegetation forecasting, as shown in Figure 3. The DiT architecture integrates transformer blocks into the diffusion model framework. Instead of relying on conventional convolutional layers, DiT employs self-attention mechanisms to capture long-range dependencies, allowing for efficient scaling and improved performance in forecasting tasks. The time series of remote sensing image features $Z_{1:T+K}$ are divided into patches of size $P \times P$ using the patchify operation, then downsampled by a factor of P through convolution operations and flattened in the spatial dimension to produce one-dimensional sequences, referred to as $Z_{1:T+K}$ tokens. Static environmental variables E are embedded via a convolutional layer and downsampled by P to produce E tokens. Dynamic meteorological variables $M_{1:T+k}$ are embedded through an MLP layer to produce $M_{1:T+k}$ tokens. All these embedded vectors are then fed into DiT blocks to model the geospatial vegetation changes. After processing through N DiT blocks, the remote sensing image feature time series undergoes unpatchify operations to upsample by a factor of P, generating the time series of geospatial vegetation states $Z_{1:T+K}$. As only future vegetation states are needed, past vegetation states $Z_{1:T}$, are discarded to retain future vegetation states $Z_{T+1:T+K}$.

The DiT block models the complex interactions of dynamic meteorological and static environmental variables with geospatial vegetation states, as shown in Figure 3. Given the significantly lower spatial resolution of meteorological variables compared to remote sensing images, the effect of these variables on vegetation is global. Thus, dynamic meteorological variable tokens $M_{1:T+k}$ globally adjust the remote sensing image tokens $Z_{1:T+K}$ through the adaLN-zero method [46, 47, 45]. This method utilizes global information from meteorological variables at each time to derive adjustment parameters as normalization parameters, adjusting remote sensing image features at the corresponding time to model the global impact. The vegetation changes at specific locations are influenced by nearby vegetation, and neighborhood vegetation changes tend to be similar. Therefore, $Z_{1:T+K}$

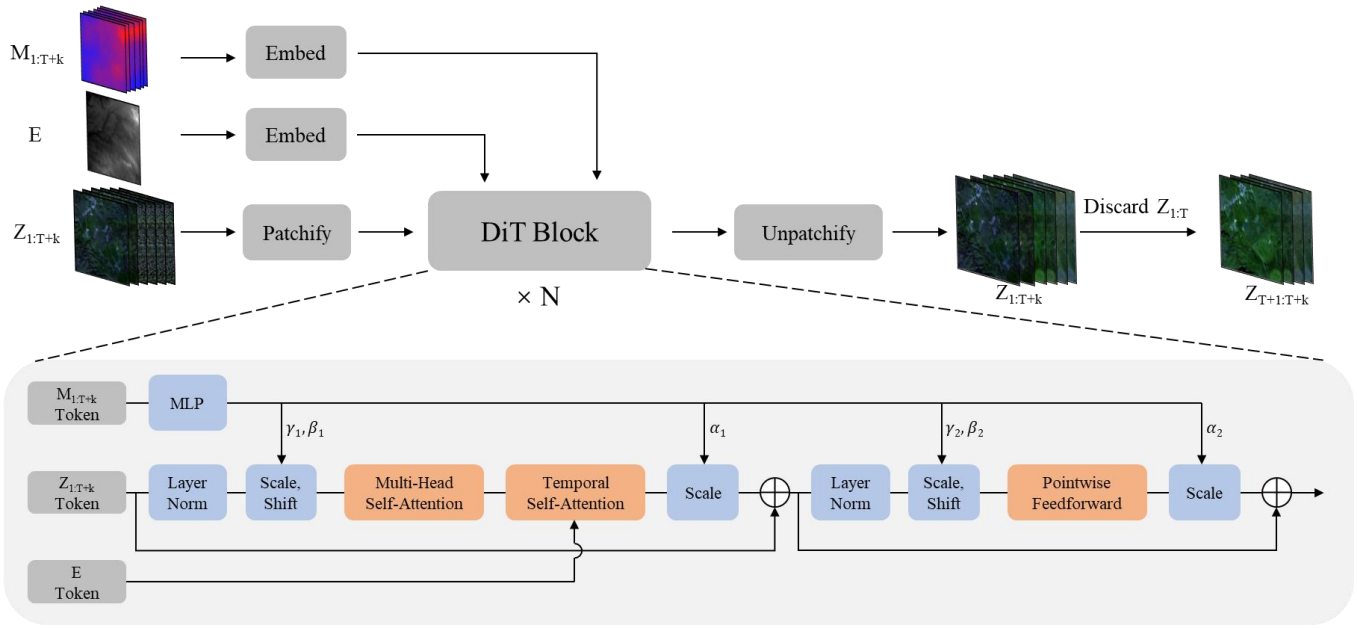

Fig. 3. The overall structure of VegeNet.

tokens utilize multi-head self-attention operations in the spatial dimension, enabling the model to focus on global vegetation states, which aids in forecasting future vegetation states. Since static environmental variables typically have a similar spatial resolution to remote sensing images, their impact on vegetation is local. Moreover, understanding past vegetation states is crucial for forecasting future states as vegetation change is a continuous process. Hence, in the temporal self-attention module, $Z_{1:T+K}$ tokens and E tokens are concatenated in the temporal dimension and subjected to self-attention operations along the temporal dimension, allowing each future vegetation state at each position to simultaneously consider corresponding past vegetation states and static environmental variables. This approach effectively models the local impact of static environmental variables on geospatial vegetation states and aids in forecasting future vegetation states based on past vegetation states.

IV. Experimental Settings and Results

A. Datasets

Satlas [48] is a large-scale pre-training dataset designed for tasks involving the analysis of satellite images. It integrates over 30 TB of satellite imagery with 137 labeled categories, drawing from public, regularly updated data sources such as Sentinel-2 and NAIP. This dataset supports a range of applications, from combating illegal deforestation to monitoring marine infrastructure. From Satlas, all Sentinel-2 remote sensing images are extracted, retaining only the blue, green, red, and near-infrared channels, resulting in approximately 10M RGBN remote sensing data. These images were divided into training and validation sets at a 9:1 ratio for extensive pre-training of the variational autoencoder, enabling it to gain a deeper understanding of RGBN remote sensing images.

The EarthNet2021 dataset [20] includes more than 32,000 samples, each with high-resolution Sentinel 2 [49] satellite imagery (20 m per pixel) and corresponding mesoscale E-OBS [50] interpolated meteorological data (1.28 km resolution), covering diverse European landscapes, which is designed for the geospatial vegetation forecasting task. Each sample comprises 30 sequential frames with a 5-day interval, capturing four channels and meteorological variables such as precipitation, sea level pressure, and temperature ranges.

EarthNet2021X [19] enhances EarthNet2021 by optimizing cloud masks, adding additional static environmental variables, and dynamic meteorological data, converting the latter from raw to one-dimensional data. It also introduces a vegetation mask that ensures the remote sensing image time series adequately represents the dynamic changes in geospatial vegetation. The minimum NDVI in the vegetation mask segments is above zero, the standard deviation was greater than 0.1, and the observable frames exceeded three in the context period and ten in the target period.

We use the EarthNet2021X dataset for the geospatial vegetation forecasting task. Given 50 days of past vegetation states (10 remote sensing image frames) at 5-day intervals, 150 days of meteorological variables (150 frames), and static environmental factors, the task is to forecast the geospatial vegetation states for the next 100 days (20 remote sensing image frames) at 5-day intervals.

To ensure that the variational autoencoder fully comprehends the geospatial vegetation states, vegetation remote sensing images from the EarthNet2021X dataset are extracted for fine-tuning the variational autoencoder. All images are extracted from the EarthNet2021X remote sensing image time series, which were split into training and validation sets at a 9:1 ratio to fine-tune the variational autoencoder pre-trained on the Satlas dataset.

After fine-tuning the variational autoencoder, VegeNet is trained using the original data split method of the EarthNet2021X dataset. This approach enabled VegeNet to effec-

tively forecasting the future state of geospatial vegetation in the well-trained latent space.

B. Benchmark methods

To assess the effectiveness of the proposed VegeDiff, comparative experiments are conducted with various benchmark methods on the geospatial vegetation forecasting task, ensuring consistency in dataset splitting and data usage across all methods. The benchmark methods can be categorized into non-ML methods, CNN-based models, RNN-based models, and transformer-based models. RNN-based models use autoregressive approaches to forecast the future geospatial vegetation states, while CNN-based and transformer-based models forecast all future vegetation states simultaneously.

The non-machine learning benchmarks include persistence methods [20] (using the last cloud-free NDVI pixel) and historical comparisons [26] (utilizing linearly interpolated data from the previous year). The CNN-based models include SimVP [51] and U-Net [52]. The RNN-based approaches feature ConvLSTM [21] and PredRNN [53], the transformer-based methods include ViT [54] and Swin Transformer [55], and CNN-transformer hybrid approaches are represented by Earthformer [56].

C. Implementation details

1) Data preprocessing and augmentation: To ensure accurate data loading, the same data preprocessing methods used in EarthNet2021X [19] are adopted. However, unlike EarthNet2021X, cloud-covered and non-vegetation pixels are not directly mask out as this would result in incomplete remote sensing images with partial masking. Instead, the cloud mask regions are filled using the adjacent values in the time dimension of the remote sensing image time series. This approach ensures that the vegetation in the cloud mask regions remains consistent with the overall vegetation changes. Additionally, the values in the non-vegetation mask regions are replaced with their mean values in the time dimension, thereby preserving the spatial features of the remote sensing images while ensuring no vegetation change in the time dimension for these regions. By filling the cloud mask and non-vegetation mask regions in the remote sensing image time series, unmasked remote sensing images that fully reflect the vegetation states of the entire area can be generated.

For the cloud-covered portions, the cloud-covered areas in the current frame are replaced with the average of the corresponding areas in the preceding and succeeding frames. For the non-vegetation pixels, the average values of the corresponding areas from the previous 10 frames are used to replace the non-vegetation pixels across the entire time series of remote sensing images.

To demonstrate the effectiveness of the proposed methods, only straightforward data augmentation techniques are employed, avoiding the use of any elaborate tricks. For the geospatial vegetation forecasting task, the data augmentation methods used for the VegeDiff model included flipping (p=0.5) and transposing (p=0.5), which is consistent with the data augmentation methods used by the benchmark method for comparison.

2) Training and Inference: We employed PyTorch [57] to construct and deploy the variational autoencoder on eight RTX A100 GPUs (80G each) and VegeNet on four RTX A100 GPUs (80G each). During the training of the variational autoencoder, the batch size is set to 64 and used Adam [58] with an initial learning rate of 4.5e-6. The variational autoencoder was trained over 50 epochs on the Satlas dataset, with the checkpoint exhibiting the lowest mean squared error (MSE) on the validation set being saved. Subsequently, it was fine-tuned for 10 epochs on the EarthNet2021X dataset, where again the checkpoint with the lowest MSE on the validation set was preserved as the final pre-trained model. For training VegeNet, the batch size is set to 16 and used AdamW [58] with an initial learning rate of 2e-4. VegeNet was trained for 200 epochs on the EarthNet2021X dataset, saving the checkpoint with the lowest root mean squared error (RMSE) on the validation set as the final model.

3) Evaluation metrics: We employed two key metrics for evaluation: Root Mean Square Error (RMSE), and Structural Similarity Index Measure (SSIM). The RMSE provides a sensitive metric that elevates the errors by squaring them before averaging, thus giving weight to larger errors. In our task, both the prediction and the ground truth are stored in arrays of size $T \times C \times H \times W$, where T is the number of time steps, C is the number of channels, H and W are the spatial heights and widths, respectively. The calculation of RMSE is formalized as follows:

$$\text{RMSE} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{C\cdot H\cdot W}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(x_{t,c,h,w}-y_{t,c,h,w}\right)^2} \quad (3)$$

The SSIM, on the other hand, measures the visual impact of differences between the forecasted and actual images. The calculation of SSIM is formalized as follows:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (4)$$

where μ_x and μ_y are the averages of images x and y, σ_x^2 and σ_y^2 are the variances of x and y, σ_{xy} is the covariance between x and y, C_1 and C_2 are constants used to stabilize the division in case the denominators are small. For our multi-temporal prediction, the overall SSIM is the average of the SSIM computed for each time step t and each channel c:

$$\text{SSIM} = \frac{1}{T\cdot C}\sum_{t=1}^{T}\sum_{c=1}^{C}\text{SSIM}(x_{t,c,:,:},\, y_{t,c,:,:})\,, \quad (5)$$

The RMSE emphasizes larger errors by squaring the differences before averaging, which is particularly useful in highlighting significant forecasting failures, while the SSIM quantifies the perceptual similarity between the forecasted and actual images.

To validate the model's performance in forecasting future vegetation states, the RMSE and SSIM are calculated not only on RGBN images but also on Normalized Difference Vegetation Index (NDVI) and Atmospherically Resistant Vegetation Index (ARVI) images. **Normalized Difference Vegetation Index** is defined as:

$$\text{NDVI} = \frac{NIR - RED}{NIR + RED} \quad (6)$$

where NIR and RED represent the reflectance in the near-infrared and red spectral bands, respectively. NDVI is a robust indicator of live green vegetation. This index leverages the high absorption of red channel by chlorophyll and the high reflectance of NIR by plant cell structures, facilitating the monitoring of plant health, biomass, and coverage over time. **Atmospherically Resistant Vegetation Index (ARVI)** is defined as:

$$\text{ARVI} = \frac{NIR - (2 \times RED - BLUE)}{NIR + (2 \times RED - BLUE)} \tag{7}$$

where NIR, RED, and $BLUE$ are the reflectance values in the near-infrared, red, and blue bands, respectively. ARVI enhances the NDVI by introducing a correction for atmospheric effects, particularly aerosol scattering in the red spectral band. This is achieved by incorporating the blue band as a reference to adjust the red reflectance prior to the NDVI computation. Given that remote sensing imagery in the EarthNet2021X dataset is prone to atmospheric disturbances, ARVI offers a more robust and accurate representation of vegetation conditions compared to NDVI.

Therefore, six validation metrics are utilized in total: RMSE and SSIM calculated on RGBN, NDVI, and ARVI images. These metrics comprehensively reflect the model performance in the task of forecasting geospatial vegetation states.

D. Ablation study

To verify the effectiveness of temporal self-attention module and adaLN-zero method in VegeNet, and to explore the role of dynamic meteorological variables and static environmental variables in geospatial vegetation forecasting, ablation experiments are conducted on the EarthNet2021X dataset.

First, to validate the effectiveness of temporal self-attention, this module from VegeNet is removed and the static environmental variables is replicated with T times, concatenating them with the input remote sensing image time series along the channel dimension. Second, to demonstrate the effectiveness of adaLN, the adaLN branch is removed and the dynamic meteorological variables are upsampled to the spatial size of the remote sensing images, then concatenated them with the remote sensing image time series along the channel dimension. Finally, to explore the roles of dynamic meteorological variables and static environmental variables, experiments were conducted using four different settings: no relevant variables, only dynamic meteorological variables, only static environmental variables, and all relevant variables, with the values of unused variables set to zero.

The ablation results for temporal self-attention and adaLN are summarized in Table I, demonstrating that the simultaneous use of both modules yields the best performance. Specifically, the temporal self-attention module enables the model to capture both static environmental variables and geospatial vegetation states from time 0 to $T-1$ when forecasting the state at time T, thereby facilitating accurate future predictions based on historical data. Conversely, adaLN converts meteorological variables at each time step into adjustment parameters for normalization, effectively modeling their global impact on geospatial vegetation states.

The ablation results for dynamic meteorological variables and static environmental variables are shown in Table II, demonstrating that both types of variables help the model forecast geospatial vegetation, and utilizing both variables simultaneously can further enhance the model's performance in forecasting the future state of geospatial vegetation. Since changes in geospatial vegetation are influenced by static variables such as land cover types and DEMs in the geographical environment, as well as dynamic influences from meteorological variables such as precipitation and temperature, both dynamic meteorological variables and static environmental variables effectively assist the model in accurately forecasting future geospatial vegetation states. Using static environmental variables as auxiliary inputs can slightly improve both the overall prediction performance of the model and the predictions for individual vegetation indices. In contrast, incorporating dynamic meteorological variables significantly enhances the model's performance. The improvements provided by these two types of variables are complementary, their combined use further boosts the model's performance.

E. Overall Comparsion

To demonstrate the effectiveness of the proposed VegeDiff, comparative experiments are conducted on the EarthNet2021X dataset. The results, as shown in Table III, indicate that VegeDiff outperforms all comparison methods across various metrics, achieving superior performance in the task of geospatial vegetation forecasting. Compared to non-machine learning benchmarks, VegeDiff achieves significantly lower RMSE and higher SSIM, indicating that it is capable of generating more accurate forecasting results. Similarly, when compared to CNN-based and transformer-based models, VegeDiff exhibits significantly higher SSIM and lower RMSE, which demonstrates its ability to produce clearer forecasting results. VegeDiff leverages millions of RGBN remote sensing images for pretraining the variational autoencoder and fine-tunes it on 20K geospatial vegetation states remote sensing data. This process results in a latent space that effectively represents geospatial vegetation states. Within this latent space, VegeDiff models the effects of dynamic meteorological variables and static environmental variables on past geospatial vegetation states, enabling accurate forecasting of future geospatial vegetation states.

We performed inferences on sampled data in the EarthNet2021X test dataset using the trained ConvLSTM, Earthformer, and VegeDiff models. Based on the inference results, the NDVI and ARVI indices were calculated, and their images were displayed for the 5th, 10th, 15th, 20th, 25th, and 30th days, as illustrated in Figures 4 and Figure 5. Due to the deterministic nature of ConvLSTM and Earthformer, their forecastings tend to be blurry, which is disadvantageous for forecasting future vegetation states. Earthformer, in particular, produces even blurrier results as it divides images into several patches for forecasting. In contrast, VegeDiff employs a diffusion model to model the process of vegetation change. As a probabilistic model, it effectively handles the uncertainties in vegetation change, generating clear and accurate forecasting of vegetation states.

TABLE I
ABLATION STUDY OF TEMPORAL SELF-ATTENTION AND ADALN ON THE EARTHNET2021X DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Temporal Self-Attention	adaLN	RGBN		NDVI		ARVI	
		RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑
		0.11	0.86	0.23	0.79	0.26	0.73
✓		0.05	0.92	0.16	0.87	0.17	0.81
	✓	0.09	0.88	0.22	0.82	0.24	0.76
✓	✓	**0.04**	**0.94**	**0.13**	**0.89**	**0.14**	**0.82**

TABLE II
ABLATION STUDY OF DYNAMIC METEOROLOGICAL VARIABLES (DMVS) AND STATIC ENVIRONMENTAL VARIABLES (SEVS) ON THE EARTHNET2021X DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

DMVs	SEVs	RGBN		NDVI		ARVI	
		RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑
		0.07	0.86	0.17	0.82	0.20	0.76
✓		0.05	0.91	0.14	0.87	0.15	0.79
	✓	0.07	0.88	0.16	0.83	0.18	0.75
✓	✓	**0.04**	**0.94**	**0.13**	**0.89**	**0.14**	**0.82**

TABLE III
OVERALL COMPARSION ON THE EARTHNET2021X DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Model	RGBN		NDVI		ARVI	
	RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑
Persistence [20]	0.09	0.89	0.23	0.84	0.25	0.78
Previous year [26]	0.08	0.91	0.20	0.86	0.21	0.81
ConvLSTM [21]	0.05	0.87	0.16	0.78	0.17	0.67
PredRNN [53]	0.06	0.88	0.17	0.81	0.19	0.71
SimVP [51]	0.05	0.89	0.16	0.80	0.17	0.64
U-Net [52]	0.08	0.85	0.18	0.82	0.20	0.60
ViT [54]	0.06	0.84	0.18	0.79	0.19	0.58
Swin Transformer [55]	0.05	0.86	0.16	0.81	0.17	0.61
Earthformer [56]	0.05	0.84	0.15	0.70	0.16	0.57
VegeDiff	**0.04**	**0.94**	**0.13**	**0.89**	**0.14**	**0.82**

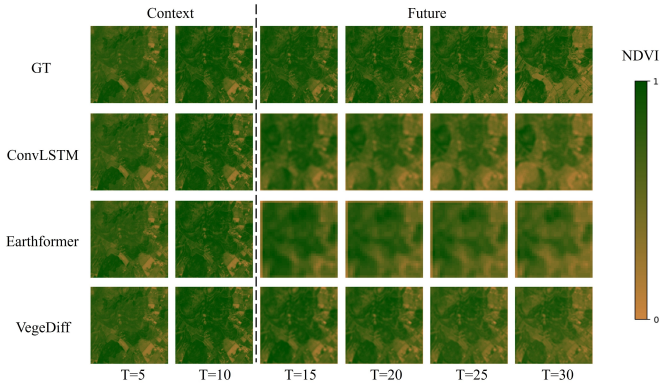

Fig. 4. NDVI images of sample inference results of ConvLSTM, Earthformer and VegeDiff on the EarthNet2021X test dataset. A time step represents a 5-day interval. This indicates that the model utilizes historical information recorded at 5-day intervals over a 50-day period to predict the geospatial vegetation state over a future span of 100 days, also at 5-day intervals.

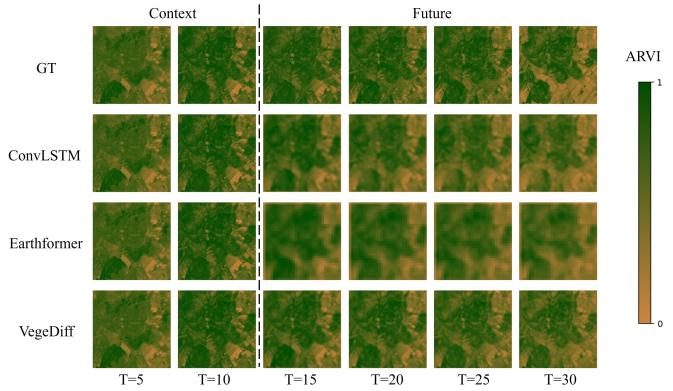

Fig. 5. ARVI images of sample inference results of ConvLSTM, Earthformer and VegeDiff on the EarthNet2021X test dataset. A time step represents a 5-day interval. This indicates that the model utilizes historical information recorded at 5-day intervals over a 50-day period to predict the geospatial vegetation state over a future span of 100 days, also at 5-day intervals.

F. Model Performance over Lead Time

Due to the complex influence of dynamic meteorological variables and static environmental variables, the geospatial vegetation change process exhibits a high degree of complexity and variability. Lead time refers to the interval between the forecasting time and the current time. Therefore, as the lead time of geospatial vegetation forecasting increases, both the degree and uncertainty of geospatial vegetation change increase, posing great challenges for vegetation forecasting. To demonstrate the superiority of VegeDiff in forecasting geospatial vegetation states over long lead time, the performance of VegeDiff and comparative benchmark methods are tested

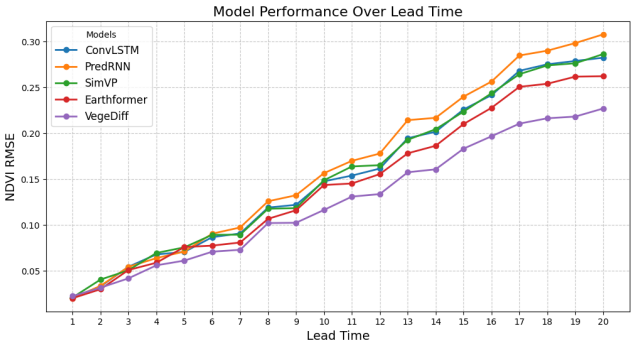

Fig. 6. Illustration of model performance over lead time. Lead time refers to the interval between the forecasting time and the current time. Each lead time corresponds to a 5-day interval. This indicates that the model forecasts the geospatial vegetation state at 5-day intervals over the next 100 days.

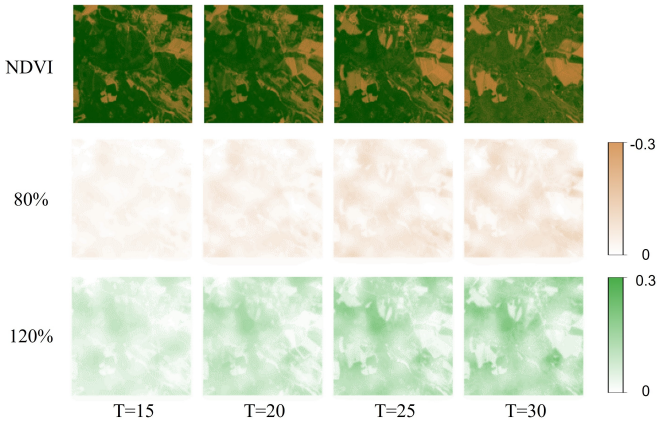

Fig. 7. The influence of precipitation changes on geospatial vegetation changes. The first row represents the actual state of geospatial vegetation NDVI. The second row shows the difference between the NDVI when precipitation is reduced to 80% of its original value and the actual NDVI. The third row displays the difference between the NDVI when precipitation is increased to 120% of its original value and the actual NDVI.

under different lead time. The model performance was evaluated by the RMSE between the predicted NDVI images and the ground truth, with lower RMSE indicating better model performance. The experimental results are shown in Figure 6, where the horizontal axis represents the time interval between the current time and the forecasting time. It can be seen that as the lead time increases, the NDVI RMSE of all models increases. This is because the longer the lead time, the longer the geospatial vegetation states is influenced by other variables, leading to greater degrees of change and uncertainty in geospatial vegetation states, thereby increasing the difficulty of forecasting future vegetation states. Additionally, it is observed that when the lead time exceeds 3, VegeDiff outperforms all other models. In the case of forecasting future geospatial vegetation states over long lead time, VegeDiff significantly outperforms all comparison models, demonstrating its superiority in long lead time geospatial vegetation forecasting. By leveraging a diffusion-based framework, VegeDiff more effectively captures the underlying uncertainties present in the dynamics of surface vegetation changes. This mechanism allows VegeDiff to model the gradual, stochastic evolution of vegetation, resulting in a slower accumulation of error compared to other models.

G. Influence of Meteorological Variables

Since VegeDiff can forecast future vegetation states under the constraints of dynamic meteorological variables and static environmental variables, it can be used to explore the effects of these variables on geospatial vegetation states. Specifically, by modifying one or more dynamic meteorological or static environmental variables, inputting the modified variables into VegeDiff, and comparing the forecasting results with those obtained using the original variables, the impacts of these variables on vegetation changes can be analyzed. This approach helps us understand the influence of different dynamic meteorological variables and static environmental variables on vegetation changes, as well as the response of vegetation in different regions to these variable changes.

Given the significant impact of precipitation on vegetation changes, the study focused on exploring its effect. Precipitation is adjusted to 80%, 90%, 100% (no change), 110%, and 120% of the original values and test the model's performance under these conditions. As shown in Table IV, VegeDiff has the best vegetation states forecasting performance when precipitation remains unchanged (100%). Both increased and decreased precipitation lead to erroneous forecastings, which is similar to the findings of [21]. The results show that VegeDiff is sensitive to the change of precipitation and can accurately model the impact of precipitation on vegetation changes.

We displayed the NDVI deviations produced by VegeDiff under 80% and 120% precipitation conditions, as illustrated in Figure 7. The first row shows the NDVI images on the 15th, 20th, 25th, and 30th days for a sample from the test set of the EarthNet2021X dataset. The second row shows the difference between the forecasting NDVI and the ground truth under 80% precipitation, while the third row shows the difference under 120% precipitation. It shows that VegeDiff underestimates NDVI when precipitation is at 80% and overestimates NDVI at 120%. As the forecasting period increases, the underestimation in the 80% precipitation scenario becomes more pronounced, and the overestimation in the 120% scenario also intensifies. This indicates that reduced precipitation inhibits vegetation growth, while increased precipitation promotes it, which is consistent with the findings of [59]. Additionally, cumulative effects on vegetation NDVI arise from sustained changes in precipitation over time, with decreased or increased precipitation further reducing or enhancing NDVI, respectively. Moreover, different regions exhibit varied responses to changes in precipitation. Generally, areas with lush vegetation are more sensitive to precipitation changes and respond more strongly.

V. Discussion

A. Forecasting Vegetation State from Multiple Perspectives

VegeDiff effectively forecasts future geospatial vegetation states based on past vegetation states, dynamic meteorological variables, and static environmental variables. By forecasting geospatial vegetation states with RGBN remote sensing

TABLE IV
THE INFLUENCE OF PRECIPITATION CHANGES TO THE MODEL PERFORMANCE ON THE EARTHNET2021X TEST DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Precipitation	RGBN		NDVI		ARVI	
	RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑	RMSE ↓	SSIM ↑
80%	0.08	0.87	0.21	0.86	0.20	0.77
90%	0.06	0.92	0.16	0.87	0.17	0.81
100%	**0.04**	**0.94**	**0.13**	**0.89**	**0.14**	**0.82**
110%	0.05	0.93	0.15	0.87	0.15	0.80
120%	0.08	0.88	0.20	0.84	0.18	0.76

images, VegeDiff can compute a wide range of vegetation indices from the forecasting results, providing comprehensive forecasting of future vegetation states. To demonstrate that VegeDiff can forecast the future state of geospatial vegetation from multiple perspectives, three vegetation indices are selected including NDVI, Enhanced Vegetation Index (EVI), and Structure Insensitive Pigment Index (SIPI), and displayed their forecasting results for the 15th, 20th, 25th, and 30th days, as shown in Figure 8.

NDVI is an effective indicator of vegetation health and density, primarily used to assess biomass and monitor vegetation changes over time. EVI offers improved sensitivity in high biomass regions and minimizes atmospheric and canopy background influences. SIPI is less sensitive to chlorophyll content variations and more indicative of the structure and condition of vegetation canopies.

As shown in Figure 8, VegeDiff can forecast geospatial vegetation states from multiple perspectives. The forecasting NDVI results (the first row) provide an overall indication of vegetation growth and effectively reflect vegetation health. The forecasting EVI results (the second row) correct for soil and atmospheric effects, offering a more accurate representation of vegetation health and providing more detailed information on vegetation growth conditions. The forecasting SIPI results (the third row) effectively highlight areas with poor vegetation growth, serving as an early warning for vegetation diseases, geospatial drought, and other issues. Therefore, by forecasting RGBN remote sensing images through VegeDiff, various vegetation indices can be computed to forecast future vegetation states from multiple aspects. This capability provides valuable insights for agricultural management, disaster warning, and other applications.

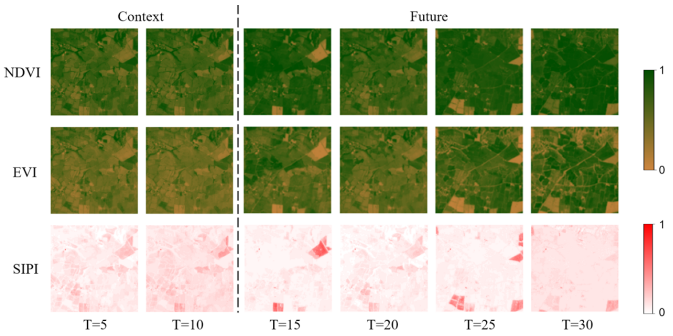

Fig. 8. NDVI, EVI and SIPI forecasting results of VegeDiff on the EarthNet2021X dataset.

B. Expectations and Limitations

The evolution of geospatial vegetation states is governed by complex interactions between dynamic meteorological factors and static environmental variables, leading to significant uncertainty. Traditional deterministic methods often yield vague and inaccurate forecasts because they cannot adequately address these combined effects.

To overcome these challenges, VegeDiff is proposed for geospatial vegetation forecasting. VegeDiff employs a diffusion model to capture the inherent uncertainty in vegetation changes, thereby producing clear and accurate predictions. Its component, VegeNet, models the influence of historical dynamic meteorological and static environmental variables on vegetation, enabling reliable forecasting of future states. Ablation studies and comparative experiments on the EarthNet2021X dataset demonstrate that VegeDiff outperforms existing deterministic methods, particularly in long lead time forecasts.

Furthermore, exploratory experiments with VegeDiff reveal its versatile application potential. By generating multiple vegetation indices from RGBN forecast results and enabling adjustments of meteorological or environmental variables, VegeDiff provides insights into the diverse responses of vegetation to varying conditions. VegeDiff is envisioned as a benchmark that not only advances probabilistic modeling in geospatial vegetation forecasting but also opens up new avenues for practical applications.

Despite its strong performance, VegeDiff also has several limitations. The pre-training and fine-tuning of the variational autoencoder in VegeDiff require substantial data and computation, and training VegeNet demands significant memory resources. These requirements pose challenges for efficiently transferring VegeDiff to other tasks. Additionally, as a latent diffusion model, forecasting large-scale and long-term future vegetation states with VegeDiff is time and resource-intensive, hindering its deployment and application in practical scenarios.

In the future, the study aim to explore efficient fine-tuning techniques for VegeDiff, enabling its low-cost and high-efficiency transfer to other tasks. The study also plan to investigate ways to reduce the denoising steps and accelerate the denoising process in VegeDiff, speeding up the vegetation state forecasting process and facilitating its deployment and application in practical scenarios.

VI. CONCLUSION

VegeDiff is proposed for the geospatial vegetation forecasting task, which is based on the latent diffusion model. Current

deterministic methods struggle with the inherent uncertainty in vegetation changes, often resulting in blurry and inaccurate forecasting results. VegeDiff employs a diffusion model to effectively capture this uncertainty and utilizes VegeNet to model the complex interactions between dynamic meteorological variables and static environmental variables, thereby accurately forecasting future vegetation states. The variational autoencoder within VegeDiff is pre-trained on 10M RGBN remote sensing data and fine-tuned on 20K vegetation remote sensing data. This extensive training process ensures a well-represented latent space, enhancing the model's understanding and predictive capability regarding vegetation states.

Comparative experiments on the EarthNet2021X dataset demonstrate the effectiveness of VegeDiff in geospatial vegetation forecasting tasks. Unlike various deterministic methods, VegeDiff probabilistically models the uncertainty in the vegetation change process and separately models the impact of dynamic meteorological and static environmental variables on vegetation states, enabling it to generate clear and accurate forecasting results. It is anticipated that VegeDiff will serve as a baseline in the field of vegetation forecasting, promoting the development of probabilistic models in this field and facilitating the deployment and application of VegeDiff in practical scenarios.

ACKNOWLEDGMENT

This work was done during the internship of Sijie Zhao at Shanghai Artificial Intelligence Laboratory, Shanghai, China. The authors would like to thank the editor and the anonymous reviewers for their constructive comments.

REFERENCES

[1] L. Miao, L. Ju, S. Sun, E. Agathokleous, Q. Wang, Z. Zhu, R. Liu, Y. Zou, Y. Lu, and Q. Liu, "Unveiling the dynamics of sequential extreme precipitation-heatwave compounds in china," *npj Climate and Atmospheric Science*, vol. 7, no. 1, p. 67, 2024.

[2] D. Wang, Y. Chen, M. Jarin, and X. Xie, "Increasingly frequent extreme weather events urge the development of point-of-use water treatment systems," *npj Clean Water*, vol. 5, no. 1, p. 36, 2022.

[3] O. Bellprat, V. Guemas, F. Doblas-Reyes, and M. G. Donat, "Towards reliable extreme weather and climate event attribution," *Nature communications*, vol. 10, no. 1, p. 1732, 2019.

[4] K. E. Trenberth, J. T. Fasullo, and T. G. Shepherd, "Attribution of climate extreme events," *Nature Climate Change*, vol. 5, no. 8, pp. 725–730, 2015.

[5] X.-P. Song, M. C. Hansen, S. V. Stehman, P. V. Potapov, A. Tyukavina, E. F. Vermote, and J. R. Townshend, "Global land change from 1982 to 2016," *Nature*, vol. 560, no. 7720, pp. 639–643, 2018.

[6] Q. Zhu, X. Guo, W. Deng, S. Shi, Q. Guan, Y. Zhong, L. Zhang, and D. Li, "Land-use/land-cover change detection based on a siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 63–78, 2022.

[7] E. J. Millet, W. Kruijer, A. Coupel-Ledru, S. Alvarez Prado, L. Cabrera-Bosquet, S. Lacube, A. Charcosset, C. Welcker, F. van Eeuwijk, and F. Tardieu, "Genomic prediction of maize yield across european environmental conditions," *Nature genetics*, vol. 51, no. 6, pp. 952–956, 2019.

[8] V. Sagan, M. Maimaitijiang, S. Bhadra, M. Maimaitiyiming, D. R. Brown, P. Sidike, and F. B. Fritschi, "Field-scale crop yield prediction using multi-temporal worldview-3 and planetscope satellite data and deep learning," *ISPRS journal of photogrammetry and remote sensing*, vol. 174, pp. 265–281, 2021.

[9] S. Tian, A. I. Van Dijk, P. Tregoning, and L. J. Renzullo, "Forecasting dryland vegetation condition months in advance through satellite data assimilation," *Nature Communications*, vol. 10, no. 1, p. 469, 2019.

[10] A. B. Barrett, S. Duivenvoorden, E. E. Salakpi, J. M. Muthoka, J. Mwangi, S. Oliver, and P. Rowhani, "Forecasting vegetation condition for drought early warning systems in pastoral communities in kenya," *Remote Sensing of Environment*, vol. 248, p. 111886, 2020.

[11] C. Deser, F. Lehner, K. B. Rodgers, T. Ault, T. L. Delworth, P. N. DiNezio, A. Fiore, C. Frankignoul, J. C. Fyfe, D. E. Horton *et al.*, "Insights from earth system model initial-condition large ensembles and future prospects," *Nature Climate Change*, vol. 10, no. 4, pp. 277–286, 2020.

[12] H. Yan, N. Sun, H. Eldardiry, T. B. Thurber, P. M. Reed, K. Malek, R. Gupta, D. Kennedy, S. C. Swenson, L. Wang *et al.*, "Characterizing uncertainty in community land model version 5 hydrological applications in the united states," *Scientific Data*, vol. 10, no. 1, p. 187, 2023.

[13] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and f. Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[14] X. Li, M. Feng, Y. Ran, Y. Su, F. Liu, C. Huang, H. Shen, Q. Xiao, J. Su, S. Yuan *et al.*, "Big data in earth system science and progress towards a digital twin," *Nature Reviews Earth & Environment*, vol. 4, no. 5, pp. 319–332, 2023.

[15] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.

[16] F. Yao, W. Lu, H. Yang, L. Xu, C. Liu, L. Hu, H. Yu, N. Liu, C. Deng, D. Tang *et al.*, "Ringmo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[17] D. Wu, X. Zhao, S. Liang, T. Zhou, K. Huang, B. Tang, and W. Zhao, "Time-lag effects of global vegetation responses to climate change," *Global change biology*, vol. 21, no. 9, pp. 3520–3531, 2015.

[18] Z. Zhou, S. Liu, Y. Ding, Q. Fu, Y. Wang, H. Cai, and

H. Shi, "Assessing the responses of vegetation to meteorological drought and its influencing factors with partial wavelet coherence analysis," *Journal of Environmental Management*, vol. 311, p. 114879, 2022.

[19] V. Benson, C. Requena-Mesa, C. Robin, L. Alonso, J. Cortés, Z. Gao, N. Linscheid, M. Weynants, and M. Reichstein, "Forecasting localized weather impacts on vegetation as seen from space with meteo-guided video prediction," *arXiv preprint arXiv:2303.16198*, 2023.

[20] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler, "Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task." in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1132–1142.

[21] C.-A. Diaconu, S. Saha, S. Günnemann, and X. X. Zhu, "Understanding the role of weather data for earth surface forecasting using a convlstm-based model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1362–1371.

[22] V. Benson, C. Robin, C. Requena-Mesa, L. Alonso, N. Carvalhais, J. Cortés, Z. Gao, N. Linscheid, M. Weynants, and M. Reichstein, "Multi-modal learning for geospatial vegetation forecasting," in *Conference on Computer Vision and Pattern Recognition 2024*, 2024.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[24] M. Fathi, M. Haghi Kashani, S. M. Jameii, and E. Mahdipour, "Big data analytics in weather forecasting: A systematic review," *Archives of Computational Methods in Engineering*, vol. 29, no. 2, pp. 1247–1275, 2022.

[25] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 270–294, 2022.

[26] C. Robin, C. Requena-Mesa, V. Benson, L. Alonso, J. Poehls, N. Carvalhais, and M. Reichstein, "Learning to forecast vegetation greenness at fine resolution over africa with convlstms," *arXiv preprint arXiv:2210.13648*, 2022.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[28] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 830–19 843.

[29] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[30] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[31] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 095–13 105.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[33] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 453–14 463.

[34] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *MICCAI Workshop on Deep Generative Models*. Springer, 2022, pp. 117–126.

[35] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 466–22 477.

[36] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 563–22 575.

[37] D. Danier, F. Zhang, and D. Bull, "Ldmvfi: Video frame interpolation with latent diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1472–1480.

[38] S. Yu, K. Sohn, S. Kim, and J. Shin, "Video probabilistic diffusion models in projected latent space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 456–18 466.

[39] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 444–18 455.

[40] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "Videofusion: Decomposed diffusion models for high-quality video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 209–10 218.

[41] A. Asperti, F. Merizzi, A. Paparella, G. Pedrazzi, M. Angelinelli, and S. Colamonaco, "Precipitation nowcasting with generative diffusion models," *arXiv preprint arXiv:2308.06733*, 2023.

[42] L. Li, R. Carver, I. Lopez-Gomez, F. Sha, and J. Anderson, "Generative emulation of weather forecast ensembles with diffusion models," *Science Advances*, vol. 10, no. 13, p. eadk4489, 2024.

[43] J. Leinonen, U. Hamann, D. Nerini, U. Germann, and G. Franch, "Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification," *arXiv preprint arXiv:2304.12891*, 2023.

[44] Z. Gao, X. Shi, B. Han, H. Wang, X. Jin, D. Maddix,

Y. Zhu, M. Li, and Y. B. Wang, "Prediff: Precipitation nowcasting with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[45] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[46] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[47] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[48] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlaspretrain: A large-scale dataset for remote sensing image understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 772–16 782.

[49] J. Louis, V. Debaecker, B. Pflug, M. Main-Knorn, J. Bieniarz, U. Mueller-Wilm, E. Cadau, and F. Gascon, "Sentinel-2 sen2cor: L2a processor for users," in *Proceedings living planet symposium 2016*. Spacebooks Online, 2016, pp. 1–8.

[50] R. C. Cornes, G. van der Schrier, E. J. van den Besselaar, and P. D. Jones, "An ensemble version of the e-obs temperature and precipitation data sets," *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 17, pp. 9391–9409, 2018.

[51] C. Tan, Z. Gao, S. Li, and S. Z. Li, "Simvp: Towards simple yet powerful spatiotemporal predictive learning," *arXiv preprint arXiv:2211.12509*, 2022.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[53] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, S. Y. Philip, and M. Long, "Predrnn: A recurrent neural network for spatiotemporal predictive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2208–2225, 2022.

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[55] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[56] Z. Gao, X. Shi, H. Wang, Y. Zhu, Y. B. Wang, M. Li, and D.-Y. Yeung, "Earthformer: Exploring space-time transformers for earth system forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 390–25 403, 2022.

[57] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[59] K.-R. Kladny, M. Milanta, O. Mraz, K. Hufkens, and B. D. Stocker, "Deep learning for satellite image forecasting of vegetation greenness," *bioRxiv*, pp. 2022–08, 2022.