EO-WM: Toward Physically Informed World Modeling for Probabilistic Earth Observation Forecasting

Junwei Luo[1,2*] **Shuai Yuan**[1*] **Zhenya Yang**[1] **Yansheng Li**[2]
Zhe Liu[1] **Hengshuang Zhao**[1†]

[1]The University of Hong Kong [2]Wuhan University

https://github.com/hku-sail/EO-WM

Abstract

Earth Observation (EO) forecasting aims to predict future Earth surface dynamics from satellite observations under changing meteorological conditions. We study this task as probabilistic forecasting under partial observability, motivated by the broader goal of weather-driven world modeling. Weather provides a conditioning signal, while sparse observations and unobserved land-surface states leave future outcomes uncertain. However, existing methods do not fully capture this setting: deterministic models collapse uncertainty into a single future prediction, while diffusion-based methods typically treat weather variables as undifferentiated conditioning signals, and existing benchmarks focus mainly on reconstruction accuracy, with limited assessment of degradation severity and observed cross-year divergence. We introduce EO-WM, a weather-conditioned video diffusion transformer for probabilistic multispectral EO forecasting. EO-WM uses physically motivated conditioning features that represent meteorological forcing through a climatological baseline, weather anomalies, and cumulative stress signals. Specifically, it separates baseline and anomaly through distinct conditioning pathways, and accumulates anomalous forcing over time to capture sustained heat and drought stress. To evaluate forecasting behavior beyond standard metrics, we introduce two diagnostic benchmarks: an Extreme Summer Benchmark for severity-aware prediction of vegetation degradation under extreme weather, and a Seasonal Matched-Pair Benchmark that serves as an observational matched-pair forecast-divergence diagnostic. Experiments show that EO-WM reduces the error in predicted Normalized Difference Vegetation Index (NDVI) decline amplitude by a relative 5.63% and improves directional hit rate by a relative 7.80%, while remaining competitive on standard pixel-level metrics.

1 Introduction

Earth Observation (EO) forecasting predicts future satellite observations conditioned on weather information [33], and underpins important downstream applications such as extreme-event monitoring [45, 31], crop-yield prediction [46], and ecosystem or vegetation forecasting [43, 3]. Existing work [8, 36, 59] has achieved strong pixel reconstruction quality, yet current formulations and evaluations still capture only part of the EO forecasting problem.

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

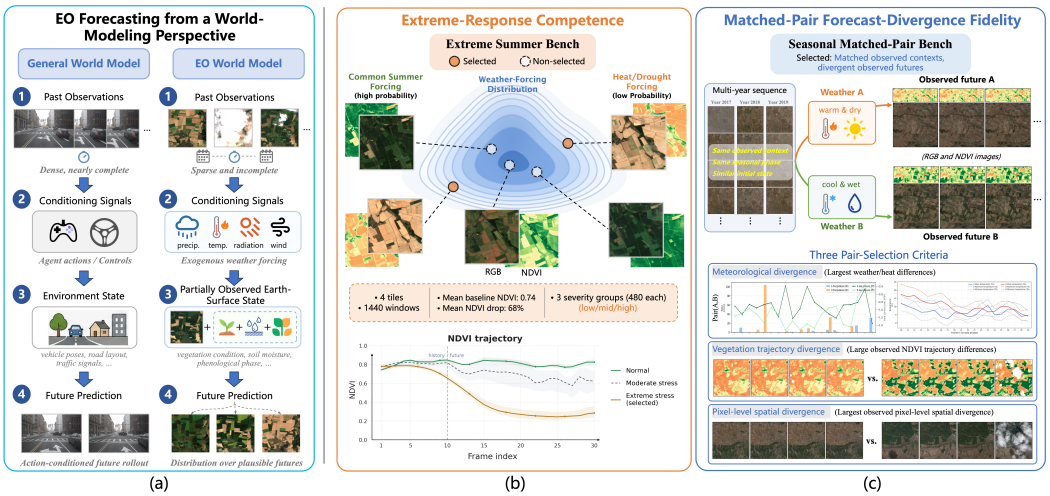


Figure 1: **A world-modeling perspective on EO forecasting and the proposed evaluation benchmarks.** (a): EO forecasting differs from standard action-conditioned world modeling: satellite observations are sparse and incomplete, and exogenous weather forcing drives future surface change in ways that depend on unobserved latent Earth-surface states. (b) and (c): We define two EO-specific evaluation dimensions beyond standard reconstruction: evaluating vegetation degradation under drought and heat stress, and evaluating observed future divergence across matched cross-year windows.

We draw on world modeling to study EO forecasting under partial observability. World models aim to learn predictive state dynamics from observations together with actions or other conditioning inputs [11, 13, 12], and have become a central paradigm in domains such as game-based visual control environments [2] and autonomous driving [29, 54]. EO forecasting likewise requires learning a dynamical process driven by exogenous forcings, with observed weather serving as the conditioning signal. Our focus is probabilistic forecasting of multispectral observations under given weather conditions, as a step toward Earth-surface world modeling.

As shown in Fig. 1 (a), two characteristics of EO forecasting shape this formulation: (1) EO observations are sparse and incomplete. Satellite revisit intervals are on the order of days, and cloud contamination further reduces the frequency of valid observations, leaving the dynamics between observations unobserved. (2) Although meteorology acts as an exogenous condition of surface state transitions, the forcing-response mapping remains stochastic under partial observability, as similar weather forcings can still lead to different surface outcomes because of unobserved internal variability and latent land-surface states (e.g., soil moisture) [5]. Together, these characteristics motivate probabilistic prediction and evaluation of vegetation dynamics beyond visual fidelity.

Existing EO forecasting methods address only parts of this picture. Deterministic predictors [8, 3, 36] provide strong point predictions, but cannot explicitly represent predictive uncertainty. More recent diffusion-based methods [59] move toward probabilistic forecasting. However, meteorological variables are still largely used as generic conditioning signals, without distinguishing among climatological background, anomalous weather events, and accumulated environmental stress. Meanwhile, existing benchmarks [33, 3] evaluate agreement with the realized future using pixel-level reconstruction and Normalized Difference Vegetation Index (NDVI) temporal consistency. They provide limited assessment of degradation severity and differences between observed futures under varying weather conditions. This motivates diagnostic evaluation beyond reconstruction, informed by the broader emphasis in world modeling on dynamics under different actions or external conditions [44, 35, 53].

To address these limitations, we present EO-WM, a weather-conditioned video diffusion transformer for probabilistic multispectral EO forecasting. EO-WM predicts future satellite imagery from sparse past observations and heterogeneous EO conditions. Its diffusion-based formulation generates multiple possible futures under partial observability. We organize weather inputs into physically motivated conditioning features that distinguish seasonal context from anomalous forcing. Specifically, we decompose weather forcing into a climatological baseline and an anomalous component, and inject them

through separate conditioning pathways. We further introduce physically motivated cumulative-stress conditioning by accumulating anomalous forcing over time. These features summarize the magnitude and duration of heat and rainfall-deficit anomalies, helping the model distinguish short-lived fluctuations from sustained exposure.

Furthermore, we introduce two benchmarks built on EarthNet2021 [33] to evaluate capabilities that are not captured by standard pixel-level metrics, as in Fig. 1 (b) and (c). The *Extreme Summer Benchmark* measures whether a model can predict vegetation degradation under drought and heat stress. The *Seasonal Matched-Pair Benchmark* is an observational matched-pair forecast-divergence diagnostic. It tests whether predictions reproduce the magnitude, direction, and ranking of observed future differences. Our contribution centers on weather-conditioning design and diagnostic evaluation as a step toward modeling the observed Earth. Our main contributions are as follows:

- We propose EO-WM, a weather-conditioned video diffusion transformer for probabilistic multispectral EO forecasting under partial observability. Its physically motivated conditioning design represents meteorological forcing through climatological baseline, weather anomaly, and cumulative stress features.
- We introduce two benchmarks: Extreme Summer and Seasonal Matched-Pair. The former evaluates vegetation degradation onset and severity under rare heat and drought forcing, and the latter evaluates the direction and magnitude of observed future divergence between matched cross-year windows.
- Experiments show that EO-WM improves vegetation degradation prediction and matched-pair forecast-divergence fidelity relative to the evaluated deterministic and diffusion-based baselines, while remaining competitive on standard reconstruction metrics.

2 Related Work

2.1 Video Prediction, Controllable Generation, and World Models

Diffusion-based video prediction provides a natural way to model uncertain future dynamics [16, 25, 58, 47, 55, 30]. Recent controllable video generation models have further improved how external conditions are injected into large video backbones [28]. Methods such as STIV [23] and ATI [49] incorporate text, image, trajectory, or motion controls into diffusion transformers, while large-scale systems such as Wan [48] and Open-Sora [60] demonstrate strong controllable generation quality with flexible conditioning interfaces. These conditions are mainly used to specify semantic content, motion patterns, or camera behavior, and are typically evaluated by alignment with the given control. In EO forecasting, meteorological inputs are observed exogenous forcings, and the key question is whether the predicted surface change responds to these forcings in a physically consistent way.

A recent line of video-based world modeling work has moved beyond passive generation toward interactive simulation. One direction develops large world foundation models or open interactive simulators [4, 1, 42]. Another direction adapts pretrained video diffusion models into action-conditioned or interactive world models [34, 18]. These works show that video generation models can be extended toward world simulation when paired with actions or controls. However, most of their conditioning signals are typically an agent action, camera motion, or user control. EO forecasting differs from this setting: the signal is observed meteorological forcing rather than a controllable action. This motivates EO-specific probabilistic modeling and diagnostic evaluation of vegetation dynamics beyond visual reconstruction.

2.2 Earth Observation Forecasting and Generative Models

EO forecasting has been formalized by EarthNet2021 [33] as predicting future satellite observations conditioned on the weather. Early methods mainly adopt deterministic prediction. ConvLSTM-based models [6] show that explicit weather conditioning improves forecasting performance, Earthformer [8] provides a strong spatiotemporal transformer backbone, and vegetation forecasting studies highlight the importance of multi-modal context and weather signals [3, 21, 19]. These methods establish EO forecasting as a weather-guided prediction task, but they typically produce a single forecast and therefore cannot explicitly represent predictive uncertainty under partial observability.

Generative EO models have recently started to address this limitation. Some methods [38, 39, 59, 37] study diffusion-based satellite forecasting and reconstruction, while UniTS [57] explores a unified framework for remote-sensing time-series tasks. Meanwhile, related generative models [20, 61, 56, 9] show the potential of metadata-conditioned and climate-aware generation. Recently, Remote-BAGEL [26] and RS-WorldModel [52] extend foundation-model and world-modeling concepts to remote sensing. However, these methods mainly focus on generic generative quality. They do not explicitly structure weather into climatological baseline, anomaly, and accumulated stress, nor do they evaluate whether generated futures respond faithfully to changes in meteorological forcing.

3 Method

We formalize weather-conditioned EO forecasting under partial observability (Sec.3.1), then describe EO-WM's architecture and generative formulation (Sec.3.2). We then present two physically motivated conditioning designs: climatology-anomaly decomposition of weather forcing (Sec.3.3) and cumulative-stress conditioning (Sec.3.4).

3.1 Problem Formulation

We consider an EO forecasting task defined over multispectral satellite image sequences. Let $\mathbf{o}_i \in \mathbb{R}^{C\times H\times W}$ denote the satellite observation at sparse timestamp u_i, where $i = 1, \ldots, T$ indexes satellite frames and C is the number of spectral bands. Let $\mathbf{a}_\ell \in \mathbb{R}^{C_a\times H_a\times W_a}$ denote dense meteorological forcing at time step $\ell = 1, \ldots, L$ on the weather grid, where C_a is the number of meteorological channels (five in EarthNet2021). We use $\pi(i) \in \{1, \ldots, L\}$ to denote the dense-weather time index aligned with satellite frame i. Given past sparse observations $\mathbf{o}_{1:T_{\text{in}}}$, the goal is to predict future observations $\mathbf{o}_{T_{\text{in}}+1:T}$, conditioned on dense meteorological forcing $\mathbf{a}_{1:L}$, static geographic context $\mathbf{s}$ (e.g., Digital Elevation Model, DEM), and spatiotemporal metadata $\mathbf{m}$ (e.g., location and calendar time). The generative objective is:

$$p_\theta(\mathbf{o}_{T_{\text{in}}+1:T} \mid \mathbf{o}_{1:T_{\text{in}}},\ \mathbf{a}_{1:L},\ \mathbf{s},\ \mathbf{m}). \tag{1}$$

Text captions are used only as auxiliary backbone context and are not part of the core task definition. This formulation treats weather as an exogenous driver of surface-state transitions, analogous to the conditioning signal in action-conditioned world models [11, 13]. At the same time, the mapping from $(\mathbf{o}_{1:T_{\text{in}}}, \mathbf{a}_{1:L})$ to $\mathbf{o}_{T_{\text{in}}+1:T}$ is not deterministic: the same weather forcing can produce different surface outcomes depending on unobserved land-surface conditions. This motivates a probabilistic generative model that can preserve observed context while representing multiple plausible futures.

3.2 Model Architecture

To model stochastic futures under sparse observations and exogenous conditions, EO-WM is built on a latent diffusion architecture [60]. We use standard generative components to study how weather conditioning helps predict future surface changes. As shown in Fig. 2, an EO-specific variational autoencoder (VAE) [22] encodes the multispectral input $\mathbf{o}_{1:T} \in \mathbb{R}^{C\times T\times H\times W}$ into a clean latent representation $\mathbf{z}_0 \in \mathbb{R}^{D\times T\times H'\times W'}$, where D is the latent dimension and H', W' are the downsampled spatial sizes. The core generative model is a Multimodal Diffusion Transformer (MMDiT) trained with flow matching [24]. Let $\mathbf{c}$ collect all conditioning inputs, including mask-aware observations, meteorological forcing, static geographic context, metadata, and auxiliary text embeddings. Given $\mathbf{z}_0$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we sample a shifted flow time $r \in (0, 1)$ and form the noisy latent $\mathbf{z}_r$ with target velocity $\mathbf{v}^\star$:

$$\mathbf{z}_r = (1 - r)\mathbf{z}_0 + [\sigma_{\min} + (1 - \sigma_{\min})r]\,\boldsymbol{\epsilon}, \tag{2}$$

$$\mathbf{v}^\star = \frac{d\mathbf{z}_r}{dr} = (1 - \sigma_{\min})\boldsymbol{\epsilon} - \mathbf{z}_0, \tag{3}$$

where $\sigma_{\min}$ is a small minimum-noise constant. The MMDiT predicts $\mathbf{v}_\theta(\mathbf{z}_r, r; \mathbf{c})$ and is trained to match $\mathbf{v}^\star$ with mean-squared error, excluding conditioned context frames and invalid pixels.

Multi-source condition routing. As summarized in Fig. 2, EO-WM routes heterogeneous conditions by their roles. The noisy latent video $\mathbf{z}_r$ forms the main video-token stream. Timestep,

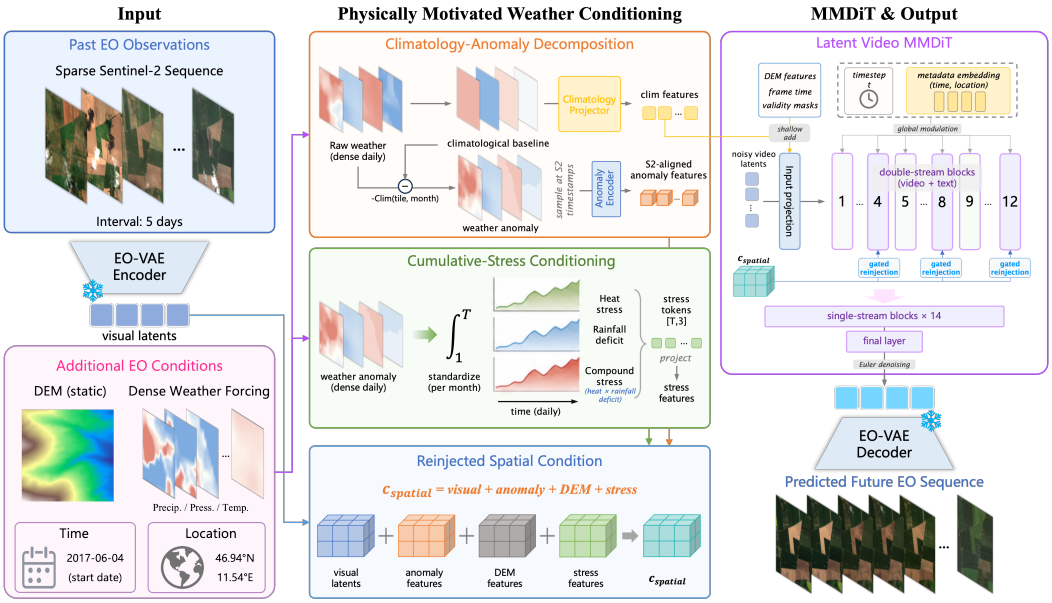


Figure 2: **Overview of EO-WM.** Sparse EO observations are encoded into visual latents, while dense daily weather forcing is decomposed using $\mathrm{Clim}(\text{tile}, \text{month})$, a precomputed monthly climatology for each geographic tile. The climatological features are injected through a shallow conditioning path, whereas anomaly, DEM, visual, and cumulative-stress features are combined into a spatial condition.

geospatial metadata, and auxiliary captions use the inherited MMDiT conditioning interfaces. Climatology features are added once at the input token layer as a seasonal reference, while spatially aligned conditions are collected into a reinjected condition $\mathbf{c}_{\text{spatial}}$:

$$\mathbf{c}_{\text{spatial}} = \mathbf{c}_{\text{vis}} + \mathbf{c}_{\text{dem}} + \mathbf{c}_{\text{time}} + \mathbf{c}_{\text{anom}} + \mathbf{c}_{\text{stress}}, \tag{4}$$

where $\mathbf{c}_{\text{vis}}$ is the projected mask-aware visual context, $\mathbf{c}_{\text{dem}}$ is the DEM features, $\mathbf{c}_{\text{time}}$ is the frame-time embedding, and $\mathbf{c}_{\text{anom}}$ and $\mathbf{c}_{\text{stress}}$ are produced by the weather modules below. All terms share the packed video-token layout, so $\mathbf{c}_{\text{spatial}}$ can be added directly to video-token hidden states during reinjection. Spatial conditions injected only at the input may weaken as features propagate through the transformer. Therefore, EO-WM periodically reinjects $\mathbf{c}_{\text{spatial}}$ into the video-token stream after selected double-stream blocks using zero-initialized learned gates. In the final model, reinjection occurs every four double-stream blocks. This lightweight mechanism keeps observation-aware and forcing-aware spatial signals available at depth and supports the anomaly and cumulative-stress conditioning designs introduced next.

3.3 Climatology–Anomaly Decomposition of Weather Forcing

Absolute meteorological values mix two different sources of physical information: a slowly varying seasonal background and a departure from that background. Thus, we decompose weather forcing into a climatological baseline and a residual anomaly, then route the two components through pathways that match their physical roles. For each geographic tile q and calendar month m, we precompute a monthly climatological mean $\bar{\mathbf{a}}_{q,m} \in \mathbb{R}^{C_a}$. For satellite frame i, with aligned dense-weather index $\ell_i = \pi(i)$ and calendar month m_i, we define:

$$\mathbf{a}_i^{\text{clim}} = \bar{\mathbf{a}}_{q,m_i}, \qquad \mathbf{a}_i^{\text{anom}} = \mathbf{a}_{\ell_i} - \bar{\mathbf{a}}_{q,m_i}, \tag{5}$$

where $\mathbf{a}_i^{\text{clim}}$ is the satellite-aligned climatological baseline and $\mathbf{a}_i^{\text{anom}}$ is the residual anomaly. The baseline is a compact seasonal reference and is added once to the input token stream, while the anomaly retains its spatial field structure, is sampled to align Sentinel-2 timestamps, and enters the reinjected spatial pathway as $\mathbf{c}_{\text{anom}}$. This separation lets the model condition on the expected seasonal regime without treating normal seasonal variation as active forcing.

This design also supports anomaly-targeted classifier-free guidance (CFG) [15]: during training, we randomly drop the anomaly tensor while retaining climatology, and at inference, we can compare full and no-anomaly predictions to amplify sensitivity to unusual forcing, as shown in Tab. 5.

3.4 Physically Motivated Cumulative-Stress Conditioning

The anomaly decomposition above captures instantaneous departures from climatology, but land-surface response often depends on the integrated history of forcing. For example, the vegetation degradation under heat or drought typically emerges after sustained exposure rather than isolated spikes. Therefore, EO-WM derives cumulative stress features on the dense meteorological timeline and samples them at the sparse satellite observation dates. Here, physically motivated describes the design of these conditioning features, instead of imposing a physical constraint on the learned dynamics.

We first standardize anomaly fields by the per-tile monthly climatological standard deviation, using $\tilde{a}$ to denote the resulting standardized anomaly. Let $\tilde{a}_\ell^{\text{temp}}(\mathbf{x})$ and $\tilde{a}_\ell^{\text{precip}}(\mathbf{x})$ denote the temperature and precipitation components at dense weather time ℓ and weather-grid location $\mathbf{x}$. Specifically, we accumulate three stress fields:

$$\text{Heat stress:} \quad S_\ell^{\text{heat}}(\mathbf{x}) = \sum_{\tau=1}^{\ell} \mathrm{ReLU}\big(\tilde{a}_\tau^{\text{temp}}(\mathbf{x})\big)\,, \tag{6}$$

$$\text{Water deficit:} \quad S_\ell^{\text{water}}(\mathbf{x}) = \sum_{\tau=1}^{\ell} \mathrm{ReLU}\Big(-\tilde{a}_\tau^{\text{precip}}(\mathbf{x})\Big)\,, \tag{7}$$

$$\text{Compound stress:} \quad S_\ell^{\text{comp}}(\mathbf{x}) = S_\ell^{\text{heat}}(\mathbf{x}) \cdot S_\ell^{\text{water}}(\mathbf{x}). \tag{8}$$

The ReLU gates retain the anomaly directions used to represent heat and rainfall-deficit exposure: positive temperature anomalies and negative precipitation anomalies, respectively. At each satellite-aligned time, we spatially average and log-compress the three stress values, project the resulting stress tokens, and add them to the reinjected spatial condition as $\mathbf{c}_{\text{stress}}$. Thus, EO-WM receives both instantaneous anomalies and cumulative exposure features without changing the diffusion objective.

4 Benchmarks and Metrics

Standard EarthNet2021 metrics [33] assess agreement with one realized future, but provide limited assessment of degradation severity and observed cross-year forecast divergence. We therefore build two diagnostic benchmarks from the EarthNet2021 test splits. Tab. 1 summarizes how each benchmark is constructed, what forecasting behavior it probes, and which metrics we report. Detailed pipelines and metrics are given in the supplementary material.

Table 1: **Benchmark construction pipeline and metric design.** Both benchmarks retain the EarthNet2021 10-context/20-target forecasting protocol, but focus on model behaviors that standard reconstruction metrics do not capture.

Benchmark	Construction pipeline	Diagnostic question	Primary metrics
Extreme Summer	`extreme_test` → NDVI curve smoothing → stable-to-decline anchor at the context/target boundary → cloud-quality and baseline-relative event verification → low/mid/high severity bins.	Tests whether the model forecasts the *onset and severity* of vegetation degradation from a healthy context under heat/drought forcing.	**Standard:** ENS/P-MAE/N-MAE **Diagnostic:** TN-MAE and DAE by severity bin
Seasonal Matched-Pair	`seasonal_test` → same-location/same-season cross-year pairs → quality filtering and initial-state matching → stratified selection over meteorological, vegetation-trajectory, and pixel-spatial divergence.	Tests whether forecasts reproduce the observed *magnitude, direction, and ranking* of future vegetation divergence in matched cross-year pairs.	**Standard:** ENS/P-MAE/N-MAE **Diagnostic:** DRR, DHR, and PDC

Extreme Summer Benchmark. The Extreme Summer Benchmark contains 1,440 verified windows selected from 4,000 sequences covering the 2018 European summer heat event. We first require a mean target-period valid-pixel ratio above 0.30 and a sequence-level NDVI decline of at least 0.35.

These thresholds ensure sufficient observable data and select substantial declines. From the 1,447 candidates, we place each 30-frame window so that the 10-frame context ends near the onset of decline. We then apply cloud-quality filtering and verify the event within the 20-frame target period. The verification threshold is 0.10 below the mean NDVI of the last four valid context frames. At least two consecutive target frames must fall below this threshold, each with at least 25% valid pixels. This baseline-relative criterion identifies sustained declines within the localized window. Verified windows are divided into low-, mid-, and high-severity bins using a composite score of decline amplitude, persistence, and observation quality. Full construction details are given in Sec. A.2.1. Selection uses observed target trajectories to form an event-focused diagnostic set. Results therefore characterize forecasting on verified declines and do not estimate event prevalence or false-positive rates in unselected sequences.

Seasonal Matched-Pair Benchmark. The Seasonal Matched-Pair Benchmark is an observational matched-pair forecast-divergence diagnostic with 422 pairs from 380 locations. We form 36,000 candidate pairs from the same geographic cube and seasonal phase in different years. After initial quality filtering, we retain pairs whose observed-context distance is at or below the 40th percentile within each seasonal phase. This distance combines vegetation and spectral histories, recent weather, and observation quality to reduce measured initial-state differences. Both windows must then have a mean valid-pixel ratio of at least 0.30, with additional cloud and shared-observation gates. These gates ensure enough observations for comparison. We select pairs through meteorological, vegetation-trajectory, and pixel-level spatial divergence tracks. Within each track and seasonal phase, we select the 50 most divergent pairs with a per-cube cap of three, then merge and deduplicate the selections. Sec. A.2.2 gives the full procedure. Future-outcome selection enriches the benchmark for divergent futures, so results describe this selected diagnostic set. Weather is a model input and one basis for sample selection. The metrics compare predicted and observed NDVI divergence without directly using meteorological variables. Unobserved land-surface states may still differ across years, so matching does not identify the independent causal effect of weather.

Metrics. We report standard reconstruction metrics for context: EarthNetScore (ENS), the official EarthNet2021 aggregate score combining MAD, OLS, EMD, and SSIM sub-scores; Pixel-MAE (P-MAE); and NDVI-MAE (N-MAE). For Extreme Summer, Trough NDVI-MAE (TN-MAE) measures NDVI error at the ground-truth trough, and Drop Amplitude Error (DAE) measures error in baseline-to-trough NDVI decline amplitude. For Seasonal Matched-Pair, Divergence Reproduction Ratio (DRR) compares predicted and observed absolute divergence magnitudes and is best when close to 1. Directional Hit Rate (DHR) measures the sign accuracy of pairwise NDVI differences on sufficiently divergent target frames. Paired Divergence Correlation (PDC) is the Spearman correlation between predicted and observed per-pair total absolute divergence. It measures whether forecasts preserve the ranking of observed future divergence across pairs. Together, these metrics assess divergence magnitude, direction, and ranking.

5 Experiments

5.1 Implementation Details

We follow the standard EarthNet2021 forecasting setting: all methods receive 10 context frames and the same weather conditions, and predict 20 future 4-channel Sentinel-2 frames at 128×128 resolution. For EO-WM, the EO-VAE tokenizer [22] is finetuned on the EarthNet2021 training split. The diffusion backbone is trained from scratch with 387M parameters. Unless otherwise specified, EO-WM does not use CFG during inference. More details are in the supplementary material.

For comparison methods, all methods except Wan2.1-Fun-V1.1-1.3B-InP (abbreviated as Wan2.1) are trained on the same EarthNet2021 training split. Earthformer follows the official implementation. Latte and the OpenSTL [41] baselines (SimVP, TAU, PredRNN, PredRNNv2, and PhyDNet) are adapted to 4-channel EO input/output protocol and receive the same weather variables through cross-attention and FiLM-based [32] conditioning, respectively. Wan2.1 is initialized from the official 1.3B checkpoint and is adapted through a carefully designed four-stage fine-tuning procedure. For stochastic generative models, we draw five predictions and evaluate the ensemble mean unless otherwise specified. More details are provided in the supplementary material.

Table 2: **Comparison of deterministic and generative models on the Extreme Summer Benchmark.** We report TN-MAE and DAE under low-, mid-, and high-intensity extreme event bins. "*" indicates that the model is finetuned from its pretrained weight.

Model	ENS↑	P-MAE↓	N-MAE↓	Low bin		Mid bin		High bin	
				TN-MAE↓	DAE↓	TN-MAE↓	DAE↓	TN-MAE↓	DAE↓
Deterministic Models									
SimVP [7]	0.1177	0.0686	0.1846	0.2142	0.2857	0.2275	0.2844	0.2247	0.2896
PhyDNet [10]	0.1157	0.0918	0.2639	0.2733	0.3317	0.2761	0.3464	0.2458	0.3458
PredRNN [50]	0.0589	0.1357	0.3757	0.3782	0.3169	0.3812	0.3226	0.3721	0.3154
PredRNN v2 [51]	0.1435	0.0655	0.2358	0.2643	0.2961	0.2874	0.3098	0.2749	0.3073
TAU [40]	0.1348	0.0781	0.2293	0.2456	0.3229	0.2553	0.3525	0.2441	0.3488
Earthformer [8]	0.2539	0.0355	**0.1099**	0.1271	0.2227	0.1348	0.2801	0.1338	0.3084
Generative Models									
Wan2.1-Fun-InP* [48]	0.2028	0.0434	0.1345	0.1371	**0.2176**	0.1447	0.2538	0.1426	0.2694
Latte [27]	0.1480	0.0581	0.1983	0.1934	0.2342	0.1972	0.2868	0.2020	0.3017
EO-WM (Ours)	**0.2543**	**0.0332**	0.1106	**0.1266**	0.2192	**0.1296**	**0.2427**	**0.1281**	**0.2372**

Table 3: **Comparison of deterministic and generative models on the Seasonal Matched-Pair Benchmark.** "*" indicates that the model is finetuned from its pretrained weight. $\mathrm{DRR}_{\mathrm{mean}}$: mean value of DRR. $\mathrm{PDC}_{\mathrm{sp}}$: Spearman correlation of predicted and observed per-pair divergence.

Model	ENS↑	Pixel-MAE↓	NDVI-MAE↓	$\mathrm{DRR}_{\mathrm{mean}} \to 1$	DHR↑	$\mathrm{PDC}_{\mathrm{sp}}$↑
Deterministic Models						
SimVP [7]	0.2098	0.0475	0.1586	1.4467	0.4938	0.1606
PhyDNet [10]	0.0626	0.1124	0.3233	1.5278	0.5049	0.0362
PredRNN [50]	0.0312	0.1325	0.3936	0.1619	0.4436	-0.1503
PredRNN v2 [51]	0.1868	0.1070	0.3614	0.3961	0.5049	0.1227
TAU [40]	0.1735	0.0482	0.1512	0.5183	0.4948	0.1386
Earthformer [8]	0.2879	0.0280	0.1109	0.6302	0.5551	0.1814
Generative Models						
Wan2.1-Fun-Inp* [48]	0.2537	0.0325	0.1208	**0.7066**	0.6050	0.2809
Latte [27]	0.2050	0.0455	0.1680	1.4021	0.5613	0.1882
EO-WM (Ours)	**0.3022**	**0.0222**	**0.1012**	0.6927	**0.6522**	**0.2942**

5.2 Comparisons

Tables 2 and 3 show that standard reconstruction quality alone does not fully characterize degradation severity and matched-pair forecasting performance. Earthformer remains a strong deterministic baseline and gives the lowest overall NDVI-MAE on Extreme Summer, but its drop-amplitude error increases with event severity, indicating conservative forecasts that under-reproduce large vegetation declines. Generative baselines provide stochastic futures. Wan2.1 achieves the DRR closest to 1 on Seasonal Matched-Pair, but has lower DHR and PDC than EO-WM. EO-WM achieves the best TN-MAE in all severity bins on Extreme Summer and the best DHR/PDC on Seasonal Matched-Pair. These results show improved degradation prediction and more accurate reproduction of the direction and ranking of observed cross-year divergence.

Additionally, we add $K = 19$ ensemble diagnostics to further support stochastic probabilistic forecasting with EO-WM. Appendix A.2.6 presents the full results and examines uncertainty calibration of the raw ensemble and exploratory scalar recalibration.

5.3 Ablation Studies

The influence of physically motivated weather representation. Tab. 4 compares the two weather-conditioning components with raw weather under the same backbone and injection pathway. Climatology–anomaly decomposition reduces DAE from 0.2459 to 0.2367 and increases PDC from 0.2704 to 0.2919. The changes in ENS are small on both benchmarks. This suggests that separating

Table 4: **Ablation of physically motivated weather conditioning.** "Weather repr." denotes the representation of weather conditions. The first row feeds raw weather conditions through the same learned spatial encoder and reinjection pathway. *Decomp*: climatology–anomaly decomposition. *CumSt*: physically motivated cumulative stress conditioning. The evaluation uses no inference-time CFG.

Weather repr.	Extreme Summer				Seasonal Matched-Pair			
	ENS↑	N-MAE↓	TN-MAE↓	DAE↓	ENS↑	$\mathrm{DRR}_{\mathrm{mean}} \to 1$	DHR↑	$\mathrm{PDC}_{\mathrm{sp}}$↑
Raw	0.2538	0.1114	0.1187	0.2459	0.3007	0.6515	0.6127	0.2704
Decomp	0.2541	0.1118	**0.1171**	0.2367	0.3012	**0.6939**	0.6247	0.2919
Decomp + CumSt	**0.2543**	**0.1106**	0.1179	**0.2330**	**0.3022**	0.6927	**0.6522**	**0.2942**

Table 5: **Ablation of inference strategy.** The same EO-WM checkpoint is evaluated with different ensemble sizes and different anomaly CFG guidance scales (λ_{anom}).

N	λ_{anom}	Extreme Summer				Seasonal Matched-Pair			
		ENS↑	N-MAE↓	TN-MAE↓	DAE↓	ENS↑	$\mathrm{DRR}_{\mathrm{mean}} \to 1$	DHR↑	$\mathrm{PDC}_{\mathrm{sp}}$↑
1	1.0	0.2397	0.1185	0.1264	0.2304	0.2955	0.7085	0.6461	**0.3059**
5	1.0	**0.2543**	0.1106	0.1179	0.2330	**0.3022**	0.6927	0.6522	0.2942
5	2.0	0.2539	**0.1104**	**0.1175**	0.2360	**0.3022**	0.7169	0.6538	0.2992
5	5.0	0.2499	0.1116	0.1198	0.2352	0.2993	0.8173	**0.6629**	0.3035
5	10.0	0.2401	0.1163	0.1266	**0.2277**	0.2922	**0.9954**	0.6558	0.2972

seasonal context from weather anomalies mainly benefits degradation-amplitude prediction and the reproduction of observed cross-year divergence. Adding cumulative-stress features further reduces DAE to 0.2330 and improves DHR from 0.6247 to 0.6522. PDC also increases slightly to 0.2942. These gains support the use of accumulated heat and rainfall-deficit exposure alongside instantaneous anomalies. The improvement is not uniform across metrics. Compared with decomposition alone, TN-MAE increases from 0.1171 to 0.1179, and DRR remains nearly unchanged at about 0.69. The cumulative features therefore improve direction and ranking agreement while leaving substantial underestimation of divergence magnitude. These results provide empirical support for the feature design. Appendix A.1.5 further shows that generic accumulation is useful, while directional CumSt improves DAE, DHR, and PDC over this control.

The influence of inference strategy. Tab. 5 evaluates ensemble averaging and anomaly CFG using the same checkpoint. Increasing N from 1 to 5 at $\lambda_{\mathrm{anom}} = 1.0$ improves ENS on both benchmarks and lowers Extreme N-MAE and TN-MAE. However, DAE increases from 0.2304 to 0.2330, while Seasonal DRR and PDC decrease. Thus, better reconstruction from ensemble averaging does not ensure better degradation-amplitude or divergence-ranking predictions. At $N = 5$, increasing guidance brings DRR closer to 1, but DHR and PDC do not improve monotonically. DHR peaks at $\lambda_{\mathrm{anom}} = 5.0$. At $\lambda_{\mathrm{anom}} = 10.0$, DRR reaches 0.9954 and DAE falls to 0.2277, while ENS on both benchmarks and Extreme N-MAE and TN-MAE worsen relative to unguided inference. PDC also falls relative to scale 5.0. This trade-off shows why DRR should be read alongside direction, ranking, and reconstruction metrics. A near-unit average divergence ratio alone does not establish accurate paired forecasts. We use unguided inference for the main architectural comparisons and report anomaly CFG as an inference-time sensitivity analysis.

5.4 Diagnostic Visualization

Fig. 3 further illustrates the quantitative results by visualizing the behavioral differences behind the benchmark metrics. In Fig. 3 (a), EO-WM has the steepest fitted slope and the highest auxiliary DRA, indicating better severity calibration than others. The detection bars (Fig. 3 (b)) show a complementary advantage of probabilistic forecasting, both generative models detect extreme events much more often than the deterministic EarthFormer, especially in the low- and mid-severity events. Together, the amplitude and detection plots assess whether forecasts identify vegetation decline and reproduce its severity. Both aspects matter when evaluating extreme-event predictions. The paired

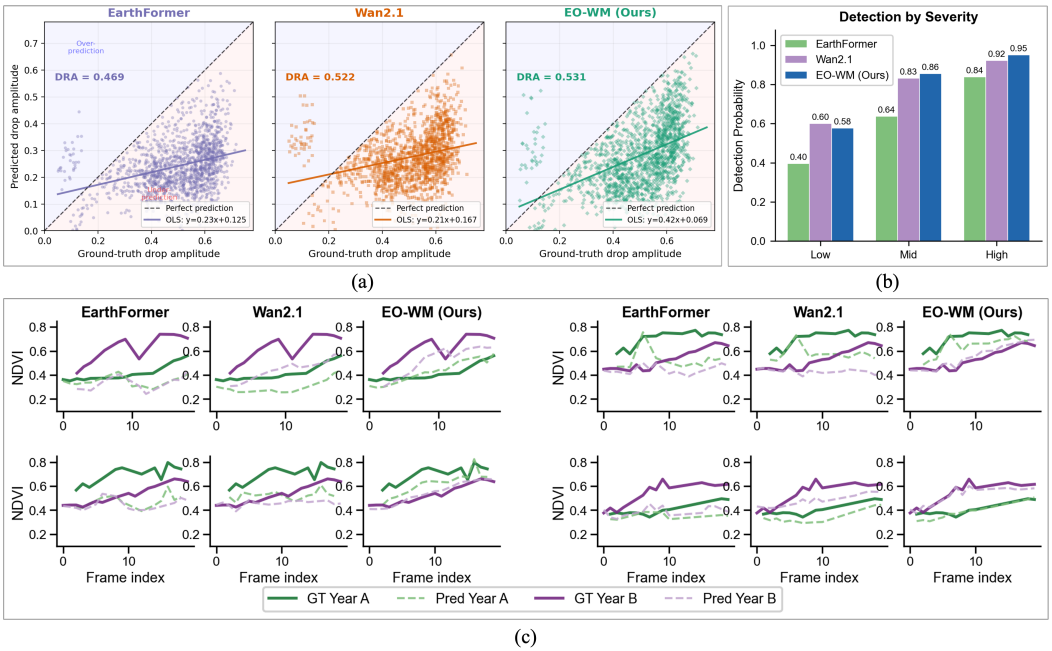


Figure 3: **Visual diagnostics apart from benchmark metrics.** (a) Predicted versus ground-truth NDVI drop amplitude on the Extreme Summer Benchmark, where the dashed line is perfect severity reproduction. DRA (Drop Reproduction Accuracy) measures relative agreement between predicted and ground-truth drop amplitudes. (b) Extreme-event detection rate by severity bin, measured as the fraction of forecasts whose target-period mean NDVI over valid vegetated pixels falls below the benchmark event threshold (0.3). (c) Seasonal Matched-Pair NDVI trajectories for same-location cross-year examples. Solid lines show ground truth and dashed lines show predictions. The examples illustrate how forecasts reproduce observed future differences.

NDVI trajectories show that EO-WM more often preserves the observed ordering and separation between the two cross-year futures. The other models reproduce these differences less accurately.

6 Conclusion

We presented EO-WM, a weather-conditioned video diffusion transformer for probabilistic multispectral Earth Observation forecasting under sparse and partial observations. Its physically motivated conditioning design separates climatological context, weather anomalies, and cumulative stress features. We also introduced two diagnostic benchmarks. Extreme Summer evaluates vegetation-degradation severity. Seasonal Matched-Pair provides an observational diagnostic of forecast divergence across years. Across the evaluated deterministic and generative baselines, EO-WM improves degradation-severity prediction and matched-pair forecast-divergence fidelity while remaining competitive on standard pixel-level metrics. These results support weather-conditioned probabilistic EO forecasting as a step toward Earth-surface world modeling.

Limitations and broader impact. Our current evaluation covers selected European regions and years in EarthNet2021, and the scale of the two diagnostic benchmarks are relatively small. Moreover, it is still hard to isolate weather effects from unobserved interannual changes. Longer-horizon forecasting and downstream benefits also remain to be validated.

Future work could combine satellite imagery with ground measurements to better observe land-surface states and assess their role in forecasting. Evaluating these forecasts in ecosystem monitoring, crop-growth prediction, and climate-risk assessment would clarify their practical value. This could improve forecasting accuracy and support ecosystem monitoring, crop-growth prediction, and climate-risk assessment, with positive impacts on society and the environment. Potential negative impacts include overreliance on forecasts for high-stakes agricultural, insurance, or disaster-response decisions.

7 Acknowledgments and Disclosure of Funding

This work is supported by the Hong Kong Research Grant Council General Research Fund (No. 17213925), National Natural Science Foundation of China (No. 62422606, No. 42371321 and No. 424B2011), Shenzhen Loop Area Institute Fund (No. FPF10120250006), and HKU Shanghai Intelligent Computing Research Center. This work is also partly supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM111, and the Key Research and Development Program of Hubei Province under Grant 2025BAB024.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

[3] Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. Multi-modal learning for geospatial vegetation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27788–27799, 2024.

[4] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Clara Deser, Flavio Lehner, Keith B Rodgers, Toby Ault, Thomas L Delworth, Pedro N DiNezio, Arlene Fiore, Claude Frankignoul, John C Fyfe, Daniel E Horton, et al. Insights from earth system model initial-condition large ensembles and future prospects. *Nature climate change*, 10 (4):277–286, 2020.

[6] Codruț-Andrei Diaconu, Sudipan Saha, Stephan Günnemann, and Xiao Xiang 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*, pages 1362–1371, 2022.

[7] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3170–3180, 2022.

[8] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

[9] Muhammed Goktepe, Amir hossein Shamseddin, Erencan Uysal, Javier Muinelo Monteagudo, Lukas Drees, Aysim Toker, Senthold Asseng, and Malte Von Bloh. Ecomapper: Generative modeling for climate-aware satellite imagery. In *Forty-second International Conference on Machine Learning*, 2025.

[10] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11474–11484, 2020.

[11] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

[12] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[13] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[14] George H Hargreaves and Zohrab A Samani. Reference crop evapotranspiration from temperature. *Applied engineering in agriculture*, 1(2):96–99, 1985.

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022.

[16] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *Transactions on Machine Learning Research*, 2022, 2022.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

[18] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

[19] Pascal Janetzky, Florian Gallusser, Simon Hentschel, Andreas Hotho, and Anna Krause. Global vegetation modeling with pre-trained weather transformers. *arXiv preprint arXiv:2403.18438*, 2024.

[20] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2023.

[21] Klaus-Rudolf Kladny, Marco Milanta, Oto Mraz, Koen Hufkens, and Benjamin D Stocker. Enhanced prediction of vegetation responses to extreme drought using deep learning and earth observation data. *Ecological Informatics*, 80:102474, 2024.

[22] Nils Lehmann, Yi Wang, Zhitong Xiong, and Xiaoxiang Zhu. Eo-vae: Towards a multi-sensor tokenizer for earth observation data. *arXiv preprint arXiv:2602.12177*, 2026.

[23] Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, et al. Stiv: Scalable text and image conditioned video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16249–16259, 2025.

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[25] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In *The Twelfth International Conference on Learning Representations*, 2023.

[26] Yuxi Lu, Biao Wu, Zhidong Li, Kunqi Li, Chenya Huang, Huacan Wang, Qizhen Lan, Ronghao Chen, Ling Chen, and Bin Liang. Remote sensing-oriented world model. *arXiv preprint arXiv:2509.17808*, 2025.

[27] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[28] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Bingyuan Wang, Qinghe Wang, Xuanhua He, Hongfa Wang, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

[29] Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15522–15533, 2024.

[30] Enrico Pallotta, Sina Mokhtarzadeh Azar, Shuai Li, Olga Zatsarynna, and Juergen Gall. Syncvp: joint diffusion for synchronous multi-modal video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13787–13797, 2025.

[31] Oscar J Pellicer-Valero, Miguel-Ángel Fernández-Torres, Chaonan Ji, Miguel D Mahecha, and Gustau Camps-Valls. Explainable earth surface forecasting under extreme events. *Earth's Future*, 13(9):e2024EF005446, 2025.

[32] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[33] Christian Requena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim 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*, pages 1132–1142, 2021.

[34] Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024.

[35] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

[36] Takayuki Shinohara. Vit-koop: Vision-transformer-koopman operators for efficient time-series forecasting of earth-observation data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2835–2844, October 2025.

[37] Qidi Shu, Xiaolin Zhu, Shuai Xu, Yan Wang, and Denghong Liu. Restore-dit: Reliable satellite image time series reconstruction by multimodal sequential diffusion transformer. *Remote Sensing of Environment*, 328:114872, 2025.

[38] Michael Smith, Luke Fleming, and James Geach. Earthpt: a foundation model for earth observation. *European Geosciences Union General Assembly 2024 (EGU24)*, page 1760, 2024.

[39] Jason Stock, Jaideep Pathak, Yair Cohen, Mike Pritchard, Piyush Garg, Dale Durran, Morteza Mardani, and Noah Brenowitz. Diffobs: Generative diffusion for global forecasting of satellite observations. *arXiv preprint arXiv:2404.06517*, 2024.

[40] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18770–18782, 2023.

[41] Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z Li. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[42] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

[43] Siyuan Tian, Albert IJM Van Dijk, Paul Tregoning, and Luigi J Renzullo. Forecasting dryland vegetation condition months in advance through satellite data assimilation. *Nature Communications*, 10(1):469, 2019.

[44] Stephen Tian, Chelsea Finn, and Jiajun Wu. A control-centric benchmark for video prediction. In *International Conference on Learning Representations*, 2023.

[45] Kevin E Trenberth, John T Fasullo, and Theodore G Shepherd. Attribution of climate extreme events. *Nature climate change*, 5(8):725–730, 2015.

[46] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and electronics in agriculture*, 177:105709, 2020.

[47] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

[48] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[49] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025.

[50] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017.

[51] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022.

[52] Linrui Xu, Zhongan Wang, Fei Shen, Gang Xu, Huiping Zhuang, Ming Li, and Haifeng Li. Rs-worldmodel: a unified model for remote sensing understanding and future sense forecasting. *arXiv preprint arXiv:2603.14941*, 2026.

[53] Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. Worldmark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026.

[54] Zhenya Yang, Zhe Liu, Yuxiang Lu, Liping Hou, Chenxuan Miao, Siyi Peng, Bailan Feng, Xiang Bai, and Hengshuang Zhao. Geniedrive: Towards physics-aware driving world model with 4d occupancy guided video generation. *arXiv preprint arXiv:2512.12751*, 2025.

[55] Xi Ye and Guillaume-Alexandre Bilodeau. Stdiff: Spatio-temporal diffusion for continuous stochastic video prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6666–6674, 2024.

[56] Ziqi Ye, Shuran Ma, Jie Yang, Xiaoyi Yang, Yi Yang, Ziyang Gong, Xue Yang, and Haipeng Wang. Object fidelity diffusion for remote sensing image generation, 2026. URL `https://arxiv.org/abs/2508.10801`.

[57] Yuxiang Zhang, Shunlin Liang, Wenyuan Li, Han Ma, Jianglei Xu, Yichuan Ma, Jiangwei Xie, Wei Li, Mengmeng Zhang, Ran Tao, et al. Units: Unified time series generative model for remote sensing. *arXiv preprint arXiv:2512.04461*, 2025.

[58] Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19310–19320, 2024.

[59] Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, and Lei Bai. Vegediff: Latent diffusion model for geospatial vegetation forecasting. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[60] Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025.

[61] Zhuo Zheng, Stefano Ermon, Dongjun Kim, Liangpei Zhang, and Yanfei Zhong. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):725–741, 2024.

A Technical appendices and supplementary material

We organize our supplementary material as follows. Section A.1 provides additional EO-WM implementation details, tokenizer reconstruction diagnostics, ablation results, and qualitative examples. Section A.2 describes the construction of the Extreme Summer and Seasonal Matched-Pair benchmarks, including filtering criteria, sample statistics, and evaluation metrics. Section A.3 summarizes the adaptation and training details for the comparison methods, including Wan2.1, Latte and the OpenSTL deterministic baselines. Section A.4 describes the data and asset availability.

A.1 EO-WM Training and Ablations

A.1.1 More Training Details

EO-WM follows the EarthNet2021 10-to-20 forecasting protocol used in the main paper. Each training example is a 30-frame Sentinel-2 sequence at 128×128 resolution with four optical channels (B/G/R/NIR), where the first 10 frames are provided as visual context and the remaining 20 frames are predicted. We first train the EO-VAE on the EarthNet2021 training split, holding out 2% of the training samples for validation. The VAE is trained with base learning rate 2×10^{-5} until convergence. We train the MMDiT diffusion backbone from scratch. The backbone uses hidden size 768, 12 attention heads, 12 double-stream blocks, 14 single-stream blocks, patch size 2, and spatial-condition reinjection after every four double-stream blocks.

For data quality control, pixels whose cloud probability is at or above 0.2 are treated as invalid in the training quality mask. The pixel-space quality mask is downsampled to the latent grid with area averaging: a latent cell is kept in the diffusion loss only when at least 50% of the corresponding pixels are valid. The loss is also masked to exclude the conditioned visual context frames, so optimization is applied to valid target-frame latents rather than to copied reference frames. During training, we randomly drop only the ERA5 anomaly branch with probability 0.15, while retaining the climatological baseline and other conditioning inputs. At inference time, the no-anomaly prediction used for guidance is obtained by zeroing the anomaly branch.

Auxiliary text captions are generated from a small set of fixed templates using sample metadata and observed sequence descriptors, such as the EarthNet2021 tile, date range, season, frame counts, cloud-condition summary, spectral bands, and the presence of daily meteorological drivers. These captions are used only to populate the inherited text-conditioning interface of the video backbone, they do not include target image content and are not part of the core EO forecasting task definition.

We train on 4 GPUs with bfloat16 mixed precision, ZeRO-2 data-parallel optimization, gradient checkpointing, and no gradient accumulation. For 30-frame clips at 128×128 resolution, we use batch size 64 per GPU, giving an effective batch size of 256. We optimize with AdamW using learning rate 2×10^{-4}, weight decay 0.01, and cosine learning-rate scheduling with 500 warm-up steps. Flow-matching timestep sampling uses shift parameter $\alpha = 2.0$. All EO-WM results reported in this paper are evaluated using the 18,000-step checkpoint selected by validation performance.

A.1.2 Tokenizer Reconstruction Analysis

We first assess whether the tokenizer reconstruction quality could be a limiting factor for downstream forecasting performance. Table A.1 compares two fine-tuned tokenizers, EO-VAE and Wan-VAE, against trivial reconstruction references on the two proposed benchmark sets. We report both masked pointwise reconstruction errors (MAE/MSE) and EarthNetScore-style components. This table therefore serves two purposes: it checks whether the tokenizer preserves multispectral observations faithfully, and it illustrates why the main paper reports standard EarthNet-style reconstruction scores together with task-targeted diagnostic metrics.

EarthNetScore components. Let S_{MAD}, S_{OLS}, S_{EMD}, and S_{SSIM} denote the benchmark-level mean subscores after excluding undefined values. Following the EarthNetScore aggregation, the overall score is the harmonic mean of the four components:

$$\mathrm{ENS} = H(S_{\mathrm{MAD}}, S_{\mathrm{OLS}}, S_{\mathrm{EMD}}, S_{\mathrm{SSIM}}) = \frac{4}{\frac{1}{S_{\mathrm{MAD}}+\epsilon} + \frac{1}{S_{\mathrm{OLS}}+\epsilon} + \frac{1}{S_{\mathrm{EMD}}+\epsilon} + \frac{1}{S_{\mathrm{SSIM}}+\epsilon}}, \quad \text{(A.1)}$$

where $\epsilon = 10^{-8}$. All four components are scaled to $[0, 1]$ with higher values indicating better agreement. MAD measures median absolute deviation between predicted and target reflectance values over non-masked pixels and all spectral channels. OLS compares ordinary-least-squares slopes of pixelwise NDVI time series, using non-masked target observations and the corresponding prediction interval. EMD computes a Wasserstein-1 distance between the predicted and observed pixelwise NDVI value distributions, with target distributions formed only from non-masked observations. SSIM computes structural similarity over frames and channels with sufficient valid target pixels, while masked target pixels are filled from the prediction to avoid penalizing unobserved regions. In Table A.1, the ENS values in this diagnostic table are the harmonic mean of the EMD, MAD, OLS, and SSIM. We report the unexponentiated SSIM component as Raw-SSIM for interpretability.

Relation to the proposed metrics. EarthNetScore is a useful and widely adopted aggregate for EarthNet2021-style forecasting. At the same time, it provides limited assessment of the diagnostic tasks studied here: reproducing vegetation-degradation severity under heat and drought, and reproducing observed future divergence between matched cross-year windows. This motivates our evaluation protocol: ENS, P-MAE, and N-MAE are reported as standard reconstruction-oriented context, while TN-MAE and DAE measure event-severity fidelity on Extreme Summer, and DRR, DHR, and PDC measure the reproduction of observed divergence magnitude, direction, and ranking on Seasonal Matched-Pair. The reconstruction experiment below provides a controlled example of this complementarity. Even when the target sequence itself is used as a reference, the masked EMD/OLS components need not reach their formal maximum, whereas masked MAE/MSE directly reflect pixelwise reconstruction fidelity. We therefore interpret EarthNet-style scores and the proposed diagnostic metrics as complementary rather than substitutive.

Table A.1: **The performance of tokenizer reconstruction on the two benchmarks.** GT is included as a diagnostic reference, not as a theoretical upper bound for EMD or OLS under masked EarthNetScore-style evaluation. Raw-SSIM denotes the unexponentiated SSIM component used for the harmonic aggregation in this table.

Method	ENS↑	EMD↑	MAD↑	OLS↑	Raw-SSIM↑	MAE↓	MSE↓
Extreme Summer Benchmark							
Ground Truth	0.3610	0.1661	1.0000	0.3267	1.0000	0.0000	0.0000
Copy last clear frame	0.1259	0.1361	0.1881	0.2902	0.7660	0.0257	0.0035
Persistence (cloud-free mean)	0.1079	0.1247	0.1849	0.2902	0.7473	0.0263	0.0033
EO-VAE (fine-tuned)	0.2955	0.1621	0.3127	0.3257	0.9125	0.0022	0.0000
Wan-VAE (fine-tuned)	0.3039	0.1999	0.2866	0.3465	0.9454	0.0037	0.0001
Seasonal Matched-Pair Benchmark							
Ground Truth	0.3809	0.1879	1.0000	0.3332	1.0000	0.0000	0.0000
Copy last clear frame	0.2790	0.2250	0.2354	0.3333	0.9100	0.0161	0.0013
Persistence (cloud-free mean)	0.2626	0.2137	0.2268	0.3333	0.8950	0.0183	0.0014
EO-VAE (fine-tuned)	0.3005	0.1735	0.3068	0.3187	0.8694	0.0034	0.0000
Wan-VAE (fine-tuned)	0.3048	0.2278	0.2863	0.3407	0.9223	0.0050	0.0001

Reconstruction fidelity. On masked pointwise metrics, EO-VAE reconstruction is near-lossless and clearly stronger than the trivial baselines. On Extreme Summer, EO-VAE reduces MAE from 0.0257 (copy-last) and 0.0263 (persistence) to 0.0022; on Seasonal Matched-Pair, it reduces MAE from 0.0161 and 0.0183 to 0.0034. EO-VAE also achieves lower MAE/MSE than Wan-VAE on both benchmark sets, indicating that the EO-adapted tokenizer preserves EarthNet multispectral observations more faithfully at the pixel level. These results suggest that tokenizer reconstruction is already strong enough that the main bottlenecks in downstream forecasting lie in future-dynamics modeling rather than in the latent autoencoding stage.

Why GT is not the EMD/OLS upper bound. A potentially confusing pattern in Table A.1 is that **GT does not achieve EMD or OLS equal to** 1, and some simple or smoothed predictors can even exceed GT on these two submetrics. This is not a computation error, but a consequence of the official EarthNet2021 metric definitions under pixel-level spatiotemporal masking. In the official implementation, EMD is computed for each pixelwise NDVI time series by comparing the prediction distribution over the *full predicted sequence* against the target NDVI distribution formed only from

non-masked target values. OLS is similarly asymmetric: for each pixelwise NDVI time series, the target slope is fitted only on non-masked target values, whereas the prediction slope is fitted over the contiguous interval between the first and last non-masked target observations. Therefore, exact recovery of the full ground-truth trajectory is not, in general, the optimum for either metric. If the cloud-masked target values differ systematically from the statistics of the non-masked subset, then even the raw GT sequence can score below 1.

Implication for interpreting the table. The table itself shows this asymmetry clearly. On Seasonal Matched-Pair, both Copy last clear frame and Persistence exceed GT in EMD/OLS, even though their MAE/MSE are much worse. Likewise, Wan-VAE achieves higher EMD/OLS than EO-VAE on both benchmark sets despite having consistently worse masked MAE/MSE. These cases indicate that EMD and OLS reward agreement with the *non-masked target-subset statistics of each pixelwise time series* rather than exact pointwise reconstruction of the complete target trajectory. We therefore treat EMD/OLS here as diagnostic references for the behavior of the official EarthNet metrics under sparse, cloud-masked observations, not as tokenizer-fidelity upper bounds. For assessing the reconstruction quality of the tokenizer itself, masked MAE/MSE and Raw-SSIM are the more reliable indicators; under those metrics, EO-VAE reconstruction is substantially stronger than the trivial baselines and slightly better than Wan-VAE.

A.1.3 Condition-Injection Ablation

We isolate two design choices: whether all the EO conditions are included in addition to mask-aware visual context, and whether spatial condition features are reinjected inside the MMDiT blocks. The results are shown in Table A.2, the first row uses only visual context and quality masks. The second row adds EO side conditions but injects the resulting features once only at the input. The third row is the default EO-WM setting, with spatial-condition reinjection after every four double-stream blocks. All rows are evaluated without the inference-time CFG.

Table A.2: **Ablation of condition injection.** "EO cond." denotes meteorological, static geographic, and spatiotemporal conditions. "Deep reinj." denotes repeated injection of spatial condition features inside the MMDiT blocks. N-MAE denotes NDVI-MAE, TN-MAE denotes trough NDVI-MAE, and DAE denotes drop amplitude error. $\mathrm{PDC}_{\mathrm{sp}}$ is the Spearman correlation of predicted and observed per-pair divergence.

EO cond.	Deep reinj.	Extreme Summer				Seasonal Matched-Pair			
		ENS↑	N-MAE↓	TN-MAE↓	DAE↓	ENS↑	$\mathrm{DRR}_{\mathrm{mean}} \to 1$	DHR↑	$\mathrm{PDC}_{\mathrm{sp}}$↑
✗	✗	0.1458	0.2153	0.2140	0.2722	0.1508	0.4524	0.4302	0.1621
✓	✗	0.2385	0.1223	0.1286	0.2395	0.2918	0.6612	0.6186	0.2785
✓	✓	0.2543	0.1106	0.1179	0.2330	0.3022	0.6927	0.6522	0.2942

As depicted in Table A.2, using only visual context and masks performs poorly on both benchmarks, indicating that visual history alone is insufficient for weather-conditioned EO forecasting. Adding EO side conditions at the input recovers most of the performance, improving Extreme ENS from 0.1458 to 0.2385 and Seasonal DHR from 0.4302 to 0.6186. Enabling deep spatial-condition reinjection further improves reconstruction and matched-pair forecast-divergence metrics, suggesting that anomaly and stress signals are more effective when they remain accessible throughout the MMDiT transition layers rather than only at the initial tokenization stage.

A.1.4 Reinjection, Fusion, and Text Conditioning

We examine three conditioning choices in Table A.3. Each comparison changes one choice and keeps the other settings fixed. We use five-sample ensemble means without inference-time CFG.

Reinjection interval. We compare reinjection after every 2, 4, or 6 double-stream blocks. Interval 0 uses conditions only at the input and disables reinjection. Interval 2 gives the lowest DAE and the DRR closest to 1. Interval 6 gives the highest Extreme ENS and lowest N-MAE. The four-block setting achieves the highest DHR and PDC and ties for the highest Seasonal ENS. We retain it as a strong balanced default across the two benchmarks.

Condition fusion. Each modality has its own encoder and is projected into the same packed token space before fusion. Summation forms a residual condition that is reinjected through learned gates. We compare this parameter-free fusion operator with gated fusion and modality attention. Gated fusion improves several Seasonal metrics but lowers Seasonal ENS and increases Extreme TN-MAE and DAE. Modality attention lowers Extreme N-MAE but increases DAE and lowers PDC. Summation provides a simple and competitive default. No tested operator is best on every metric.

Auxiliary text. The no-text variant removes text conditioning throughout both training and inference. The template captions contain no target image information and provide optional backbone context. Removing them causes a modest decline across the reported metrics. The explicit spatial conditions supply the core task information, while text provides limited auxiliary context.

Table A.3: **Conditioning design ablations.** Each panel varies one choice from the default configuration, which uses four-block reinjection, sum fusion, and auxiliary text. DRR denotes $\mathrm{DRR}_{\mathrm{mean}}$ and PDC denotes $\mathrm{PDC}_{\mathrm{sp}}$.

	Extreme Summer				Seasonal Matched-Pair			
Setting	ENS ↑	N-MAE ↓	TN-MAE ↓	DAE ↓	ENS ↑	DRR → 1	DHR ↑	PDC ↑
Reinjection interval								
0 (input only)	0.2385	0.1223	0.1286	0.2395	0.2918	0.6612	0.6186	0.2785
2	0.2481	0.1182	0.1223	0.2231	0.2861	0.7060	0.6304	0.2833
4 (default)	0.2543	0.1106	0.1179	0.2330	0.3022	0.6927	0.6522	0.2942
6	0.2575	0.1100	0.1188	0.2299	0.3022	0.6857	0.6404	0.2841
Fusion operator								
Sum (default)	0.2543	0.1106	0.1179	0.2330	0.3022	0.6927	0.6522	0.2942
Gated fusion	0.2548	0.1104	0.1184	0.2346	0.3010	0.7167	0.6589	0.3167
Modality attention	0.2541	0.1092	0.1185	0.2437	0.3020	0.6939	0.6512	0.2845
Text conditioning								
Text (default)	0.2543	0.1106	0.1179	0.2330	0.3022	0.6927	0.6522	0.2942
No text	0.2529	0.1122	0.1196	0.2451	0.3012	0.6751	0.6418	0.2785

A.1.5 Cumulative-Stress Feature Ablations

The heat, rainfall-deficit, and product channels in Sec. 3.4 are simple summaries of the available meteorological variables. We compare them with generic temporal accumulation and an alternative climatic water balance feature. All variants use the same training and inference settings. We report five-sample ensemble-mean results without inference-time CFG.

Generic accumulation. We use standardized daily temperature and precipitation anomalies, denoted by z_T and z_P. The generic control contains three channels:

$$A_T(t) = \sum_{\tau \le t} |z_T(\tau)|, \qquad A_P(t) = \sum_{\tau \le t} |z_P(\tau)|, \qquad A_J(t) = A_T(t) + A_P(t). \tag{A.2}$$

This control retains temporal accumulation but removes directional filtering and replaces the heat–water product with a sum. It uses the same daily validity handling, window start, satellite-date sampling, and log compression as CumSt. The backbone and feature-injection pathway are unchanged.

Table A.4: **Generic accumulation and directional cumulative-stress features.** All rows use climatology–anomaly decomposition. Directional CumSt is the paper setting. DRR denotes $\mathrm{DRR}_{\mathrm{mean}}$ and PDC denotes $\mathrm{PDC}_{\mathrm{sp}}$. Bold marks the best value in each column.

Setting	Extreme Summer				Seasonal Matched-Pair			
	ENS↑	N-MAE↓	TN-MAE↓	DAE↓	ENS↑	DRR→ 1	DHR↑	PDC↑
Decomp	0.2541	0.1118	**0.1171**	0.2367	0.3012	**0.6939**	0.6247	0.2919
+ Generic accumulation	**0.2577**	**0.1092**	0.1188	0.2338	0.3014	0.6843	0.6323	0.2831
+ Directional CumSt	0.2543	0.1106	0.1179	**0.2330**	**0.3022**	0.6927	**0.6522**	**0.2942**

Table A.4 shows that generic accumulation improves Extreme ENS, N-MAE, and DAE over Decomp. Temporal aggregation itself is therefore useful. Directional CumSt further improves DAE, DHR, and PDC over generic accumulation, while Extreme ENS and N-MAE are slightly worse. These results provide empirical support for the directional feature design on several diagnostic metrics. This comparison changes both directional filtering and the interaction term, so it does not separate their individual effects.

Evapotranspiration-aware climatic water balance. We also replace the precipitation-only deficit channel with a climatic water balance (CWB) feature. We estimate daily reference evapotranspiration using the Hargreaves–Samani formulation [14]:

$$ET_0 = \max\Big\{0,\ 0.0023(T_{\mathrm{mean}} + 17.8)\sqrt{\max(T_{\max} - T_{\min}, 0)}\,(0.408R_a)\Big\}. \tag{A.3}$$

Here T_{mean}, $T_{\min}$, and $T_{\max}$ are daily temperatures. Extraterrestrial radiation R_a is derived from latitude and day of year. The estimate represents reference evapotranspiration. Actual evapotranspiration is not observed in this setting.

With daily precipitation P, we define deficit-positive CWB as $ET_0 - P$. The accumulated deficit is

$$D^+(t) = \left[\sum_{\tau \le t}\big(ET_0(\tau) - P(\tau)\big)\right]_+, \qquad [x]_+ = \max(x, 0). \tag{A.4}$$

The balance is summed before rectification, which allows wet days to offset deficits. We retain the original heat channel and form its product with $D^+(t)$. All other settings remain unchanged.

Table A.5: **Sensitivity to an ET_0-aware CWB feature.** The alternative replaces the precipitation-only deficit channel. Directional CumSt is the paper setting. DRR denotes $\mathrm{DRR}_{\mathrm{mean}}$ and PDC denotes $\mathrm{PDC}_{\mathrm{sp}}$. Bold marks the best value in each column.

Setting	Extreme Summer				Seasonal Matched-Pair			
	ENS↑	N-MAE↓	TN-MAE↓	DAE↓	ENS↑	DRR→ 1	DHR↑	PDC↑
Directional CumSt	0.2543	**0.1106**	**0.1179**	0.2330	**0.3022**	**0.6927**	**0.6522**	0.2942
ET_0-aware CWB	**0.2562**	0.1110	0.1193	**0.2298**	0.2996	0.6716	0.6304	**0.3263**

The CWB variant improves Extreme ENS and DAE and Seasonal PDC (Table A.5). It gives weaker results on the other reported metrics. The two formulations offer different trade-offs. We retain directional CumSt as a simple, effective feature design. These experiments support cumulative-stress conditioning without establishing a uniquely optimal stress formula or a mechanistic land-surface model.

A.1.6 Qualitative Forecasting Results

Figure A.1 provides representative qualitative comparisons under the 10-context/20-target forecasting protocol. The examples show that the task is highly partially observed: even the ground-truth Sentinel-2 sequence contains frequent cloud-contaminated or missing frames, shown as white regions in the visualization. Despite this sparse supervision, the forecasts should still respond to the dense future meteorological forcing and predict the onset and progression of vegetation degradation.

The qualitative results are consistent with the quantitative trends in the experiments. Earthformer tends to produce conservative forecasts whose vegetation decline appears delayed relative to the observed target sequence, especially when heat and drought stress cause a rapid transition from green vegetation to dry or senescent surfaces. The diffusion-based models, Wan2.1 and EO-WM, better capture this forcing-response behavior and predict the degradation earlier. EO-WM further preserves more coherent spatial patterns and a more faithful timing of the response, which is consistent with the benefit of conditioning on climatological baseline, weather anomaly, and physically motivated cumulative-stress features.

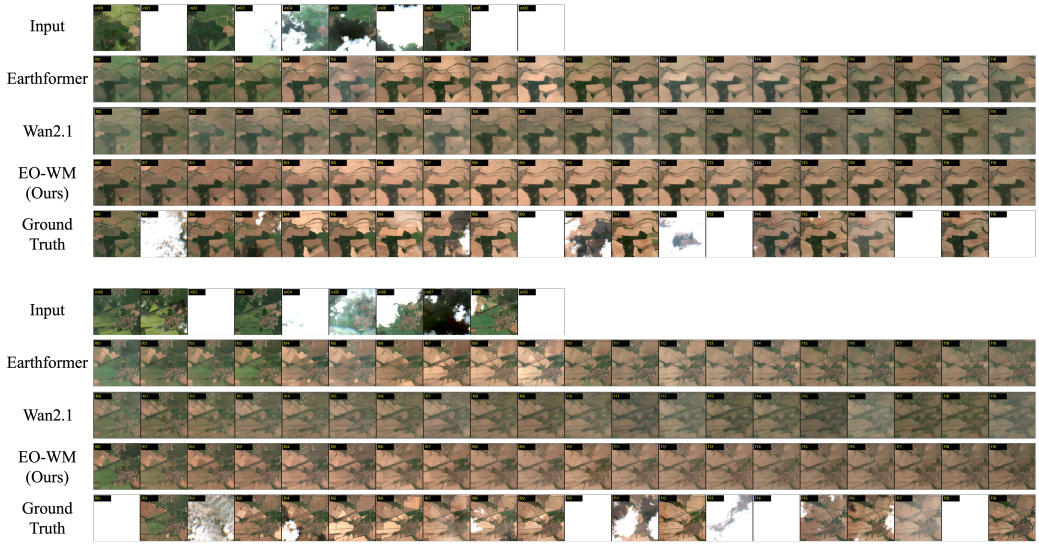


Figure A.1: **Qualitative visualization of 10-to-20 EO forecasting.** Each example uses 10 past Sentinel-2 frames as visual context and predicts the next 20 future frames. White regions indicate cloud-contaminated or missing observations in the real Sentinel-2 sequence, illustrating the sparse-observation setting under which the models must infer future surface dynamics.

A.2 Benchmark Construction and Evaluation Protocols

We describe the full construction pipelines for the two proposed benchmarks. Both are derived from existing EarthNet2021 [33] test splits via multi-stage quality filtering and event verification, producing curated evaluation sets that target specific capabilities beyond pixel-level fidelity.

A.2.1 Extreme Summer Benchmark Construction

Data source. The benchmark is constructed from the EarthNet2021 `extreme_test` split, which targets the 2018 European heatwave and drought, one of the most severe compound climate events on record in Central Europe. This split contains 4,000 sequences of 60 frames (5-day cadence, $\sim$300 days) across four Sentinel-2 tiles (32UPC, 32UNC, 32UMC, 32UQC) covering parts of France and Germany. Each sequence provides four spectral bands (blue, green, red, NIR) at 20 m resolution with an associated per-pixel quality mask indicating cloud and shadow contamination.

Our goal is to extract 30-frame evaluation windows (10 context + 20 target frames) that each contain a **verified vegetation degradation event** in the target period, i.e., a transition from a stable, vegetated baseline to a significant Normalized Difference Vegetation Index (NDVI) decline driven by the extreme weather conditions.

Stage 1: Sequence-level NDVI curve analysis. Guided by the aggregate behavior in Fig. A.2, we first analyze the full 60-frame NDVI trajectory of each sample to characterize its temporal dynamics and identify candidate sequences. For each frame, we compute the median NDVI over valid (cloud-free) pixels and apply a validity filter: frames with fewer than 30% valid pixels are masked as unobservable. Short gaps ($\leq$3 consecutive masked frames) are filled by linear interpolation, while longer gaps are preserved as missing to avoid hallucinating trends. A 5-frame moving average is applied to obtain a smoothed trajectory for robust peak–trough detection.

From each trajectory, we extract statistics including the peak and trough NDVI values, the drop amplitude (peak minus trough), the number of local minima (with prominence ≥ 0.04), the mean valid-pixel ratio in the context and target periods, and the temporal location of the global trough.

Stage 2: Window localization and extreme-event verification. This stage identifies the optimal 30-frame window within each sequence and verifies that it contains a true extreme event. The procedure operates in two sub-stages.

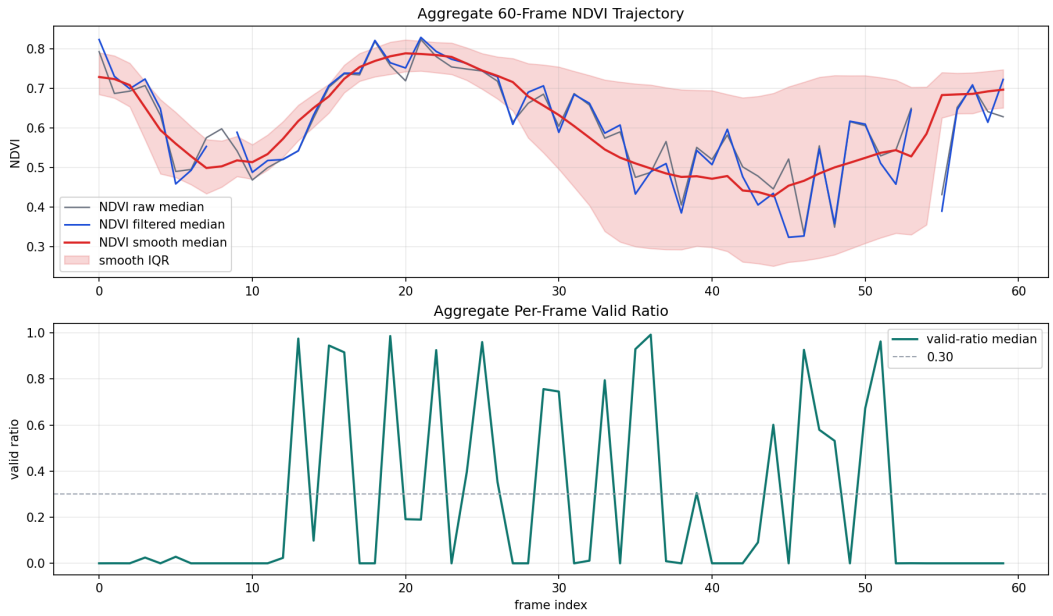


Figure A.2: **Aggregate 60-frame diagnostics for Extreme Summer window construction.** Top: aggregate NDVI trajectory over the original 60-frame sequences. The gray curve shows the raw median NDVI computed from valid pixels, the blue curve shows the median after frame-level validity filtering and short-gap interpolation, and the red curve shows the smoothed median used for robust peak–trough detection. Bottom: aggregate per-frame valid-pixel ratio, with the dashed line marking the 0.30 validity threshold. The figure illustrates both the vegetation decline signal targeted by the benchmark and the sparse-observation regime caused by clouds and missing frames.

Sub-stage 2a: Sample-level prefiltering. We apply three hard conditions to identify candidate sequences: (i) the mean valid-pixel ratio in the target period must exceed 30%, ensuring sufficient observable data; (ii) the 60-frame sequence-level NDVI drop from context-end to target-minimum must be at least 0.35, indicating a substantial decline; and (iii) the global trough NDVI must be non-negative, excluding sensor artifacts. This prefilter is used only to select candidate sequences; the final-window drop amplitude Δ reported below is the peak-to-trough amplitude measured after window localization. Of 4,000 sequences, 1,447 pass all hard conditions and proceed to window-level analysis.

Sub-stage 2b: Anchor-based window placement. For each prefiltered sequence, we locate the transition point t_{stable} where vegetation begins to decline, and place the 30-frame window so that t_{stable} lies at or near the end of the context period. This is done via a *global trough method*: we find the global NDVI minimum t_{trough} and the preceding peak t_{peak}, then search backward from t_{trough} for the latest frame where NDVI remains within 30% of the peak value and local variability (standard deviation over a 5-frame neighborhood) is below 0.05. For sequences with complex multi-trough structure, we additionally apply a *multi-trough method* that considers up to 8 local minima as candidate anchors and selects the one yielding the highest-confidence window.

Window quality filtering. Windows are discarded if they contain more than 6 invalid context frames (out of 10), more than 12 invalid target frames (out of 20), or more than 15 half-cloudy frames (valid ratio between 20% and 80%).

Extreme event verification. For each surviving window, we compute a baseline NDVI $\bar{y}_{\text{base}}$ as the mean NDVI of the last 4 valid context frames (requiring $\geq$20% valid pixels per frame). We then set an extreme threshold $\theta = \bar{y}_{\text{base}} - 0.10$. A window is confirmed as containing an extreme event if at least 2 consecutive target frames and at least 2 total target frames have spatially-averaged NDVI below θ (each requiring $\geq$25% valid pixels). Windows that fail this verification are rejected.

Each verified window receives a composite extreme score:

$$s_{\text{extreme}} = \Delta \cdot (0.55 + 0.45\, r_{\text{persist}}) \cdot (0.70 + 0.30\, r_{\text{consec}}) \cdot (0.40 + 0.60\, q), \tag{A.5}$$

where Δ is the NDVI drop amplitude, r_{persist} is the fraction of valid target frames below threshold, $r_{\text{consec}} = \min(n_{\text{consec}}/6, 1)$ rewards sustained events, and q is the mean valid-pixel ratio over the full window. The complete procedure is summarized in Algorithm 1.

Algorithm 1 Extreme Summer Benchmark Construction

Require: EarthNet2021 `extreme_test` split: 4,000 sequences $\times$ 60 frames
Ensure: Benchmark set $\mathcal{B}_{\text{ext}}$ of verified 30-frame extreme windows

Stage 1: Curve analysis
for each sequence $i = 1, \ldots, 4000$ **do**
    Compute per-frame median NDVI; mask frames with $<$30% valid pixels
    Interpolate NaN gaps $\leq$3 frames; smooth with 5-frame moving average
    Extract: peak/trough NDVI, drop amplitude, local minima count, valid ratios
end for
Stage 2a: Sample prefiltering
Apply hard conditions: target valid ratio $\geq$0.30, sequence-level target drop $\geq$0.35, trough $\geq$0
$\mathcal{C} \leftarrow$ 1,447 sequences passing all conditions
Stage 2b: Window localization & verification
for each sequence $i \in \mathcal{C}$ **do**
    Locate anchor t_{stable} via global-trough and/or multi-trough method
    Place 30-frame window with t_{stable} near the context boundary
    Compute anchor confidence c_{anchor}
    Filter: reject if too many invalid/half-cloudy frames
    Compute baseline $\bar{y}_{\text{base}}$ from last 4 valid context frames
    Verify: confirm $\geq$2 consecutive and $\geq$2 total frames with NDVI $< \bar{y}_{\text{base}} - 0.10$
    if verified **then**
        Compute extreme score s_{extreme}; add to $\mathcal{B}_{\text{ext}}$
    end if
end for
return $\mathcal{B}_{\text{ext}}$ (1,440 windows), sorted by s_{extreme}

Final benchmark statistics. Of 1,447 prefiltered candidates, 1,440 pass all verification criteria (7 rejected: 3 due to excessive context cloud cover, 3 failing extreme verification, 1 due to target cloud cover). Table A.6 summarizes the benchmark statistics.

Table A.6: **Extreme Summer Benchmark statistics.** Distribution of key attributes across the 1,440 verified windows. Q1 denotes the 25th percentile.

Attribute	Mean	Std	Min	Q1	Median	Max
Drop amplitude (Δ)	0.504	0.131	0.052	0.428	0.534	0.716
Baseline NDVI ($\bar{y}_{\text{base}}$)	0.743	0.055	0.434	0.713	0.746	0.890
Target minimum NDVI	0.287	0.091	0.150	0.213	0.260	0.676
Extreme score (s_{extreme})	0.248	0.064	0.057	0.201	0.251	0.445
Anchor confidence (c_{anchor})	0.898	0.062	0.544	—	—	0.999
Valid-pixel ratio (mean)	0.417	0.052	0.298	0.381	0.405	0.629

Anchor mode distribution: global-trough 62.0%, multi-trough 27.2%, exhaustive scan 10.7%
Tile distribution: 32UPC: 491, 32UNC: 416, 32UMC: 313, 32UQC: 220

A.2.2 Seasonal Matched-Pair Benchmark Construction

Data source and design rationale. The Seasonal Matched-Pair Benchmark is constructed from the EarthNet2021 `seasonal_test` split. It provides 3-year merged Sentinel-2 sequences for each geographic location (210 frames at 5-day cadence, covering 2017–2019). We pair windows from the *same location* and *same seasonal phase* across *different years*, then match their observed contexts. This yields an *observational matched-pair forecast-divergence diagnostic*. It tests whether forecasts reproduce the magnitude, direction, and ranking of observed future differences. Matching reduces some observable differences between windows, but does not identify the independent causal effect of weather. Unobserved soil moisture, irrigation, crop management, and land-cover changes may

also affect the observed outcomes. Weather is a model input and contributes to sample selection. The DRR, DHR, and PDC formulas use predicted and observed NDVI divergence, without directly using meteorological variables.

Window extraction. From each 210-frame sequence, we extract 30-frame windows (10 context + 20 target frames, matching the model's operational setting). Each year spans frames 0–69, and we extract three windows per year at start offsets 0, 20, and 40 within the year (corresponding to early, mid, and late growing season phases), yielding up to 9 windows per location. For each window, we compute:

- Per-frame NDVI trajectory: median NDVI over valid pixels, with frames below 30% validity masked, short gaps ($\leq$3 frames) interpolated, and a 5-frame moving average applied.
- Baseline NDVI: median of the last 4 input frames.
- Quality metrics: mean valid-pixel ratio for input and target periods, count of cloudy frames.

A window is marked as *benchmark-usable* if both mean valid ratios exceed 20% and the median valid pixel count exceeds 400 pixels per frame. From the full seasonal test split, 12,992 windows pass this soft quality gate.

Stage 1: Pair construction and divergence scoring. A pair (A, B) is formed whenever two windows share the same geographic location (cube) and seasonal phase (window start offset within year) but come from different calendar years (e.g., 2017 vs. 2019). This yields 36,000 candidate pairs. For each pair, we compute multi-dimensional divergence scores capturing different aspects of how the two windows differ:

(a) Initial state distance (D_{init}). We measure how similar the two windows are at the start of the prediction period, combining: L2 distance of input NDVI trajectories (normalized by per-group standard deviation), NIR and Red trajectory distances, absolute difference of the last input NDVI value, spectral summary distance (mean band values), recent weather distance (last 3 input frames), and valid-ratio difference. These components are aggregated via equal-weighted summation after groupwise robust normalization (IQR-based standardization within each seasonal phase) to prevent cross-phase scale differences from dominating.

(b) Meteorological divergence (D_{meteo}). Captures how differently the weather evolves in the target period, combining: absolute differences in cumulative rainfall, mean/minimum/maximum temperature, peak daily maximum temperature, and a composite heat stress index. This score drives the *Meteorological Divergence* track.

(c) Vegetation trajectory divergence (D_{veg}). Measures how differently the NDVI outcomes evolve, using a shift-tolerant ($\pm$3 frame) L2 trajectory distance that accounts for phenological timing differences, plus absolute differences in baseline-relative drop, recovery amplitude, and temporal volatility. This drives the *Vegetation Trajectory Divergence* track.

(d) Pixel-level spatial divergence (D_{pixel}). The per-frame L1 distance of spatially co-registered NDVI values, averaged over frames with $\geq$30% shared valid pixels (requiring $\geq$6 clean frames per pair). This drives the *Pixel-level Spatial Divergence* track.

Stage 2: Quality filtering and initial-state matching. We apply a series of gates to improve data quality and reduce differences in observed initial conditions:

Benchmark eligibility. Both windows in a pair must independently pass the soft quality gate. This reduces the candidate pool from 36,000 to 8,444 pairs.

Initial-state matching. We retain only pairs whose D_{init} falls at or below the 40th percentile within their seasonal phase group. This yields 3,379 pairs with similar observed initial conditions. The matching step reduces measured context differences. It does not ensure that unobserved land-surface states are identical.

Hard quality gates. Both windows must have a mean valid-pixel ratio $\geq$30% over the full 30-frame window and $\leq$20 cloudy frames. The vegetation trajectory track requires $\geq$8 clean frames with shared valid pixels. The pixel-level track requires $\geq$6 clean frames. After these gates, the eligible pool sizes are: meteorological track 3,372, vegetation trajectory track 2,047, and pixel-level track 2,675.

Stage 3: Stratified selection and convergence. To produce a manageable and balanced benchmark, we apply stratified top-N selection:

1. Within each track, we rank pairs by their respective divergence score (descending).
2. Per seasonal phase (3 phases: offsets 0, 20, 40), we select the top 50 highest-divergence pairs, yielding 150 pairs per track.
3. A per-cube cap of 3 pairs prevents any single location from dominating the benchmark.
4. The three track selections are merged via union and deduplicated by pair identity. Each pair retains metadata indicating which track(s) selected it and its rank within each track.

Selection by meteorological, vegetation-trajectory, and pixel-level divergence emphasizes pairs with measurable differences. In particular, future-outcome selection enriches the benchmark for divergent futures. Results therefore describe this selected diagnostic set. The final benchmark contains **422 unique pairs** (844 inference windows). Multi-track membership provides complementary evaluation perspectives: 394 pairs belong to a single paper-facing track, and 28 to two tracks. The complete procedure is summarized in Algorithm 2, and Table A.7 summarizes the properties of the final 422-pair benchmark.

Algorithm 2 Seasonal Matched-Pair Benchmark Construction

Require: EarthNet2021 `seasonal_test` split: 3-year sequences (210 frames), 5-day cadence
Ensure: Benchmark set $\mathcal{B}_{\text{sea}}$ of matched cross-year pairs with verified divergence
Window extraction: Extract 30-frame windows at 3 seasonal offsets $\times$ 3 years per location
Mark windows as usable if mean valid ratio $>$20% and pixel count $>$400
Pair construction: Form all same-location, same-phase, cross-year pairs $\rightarrow$ 36,000 candidates
for each pair (A, B) **do**
    Compute D_{init}, D_{meteo}, D_{veg}, D_{pixel} with groupwise normalization
end for
Stage 2: Filtering
Require both windows benchmark-eligible $\rightarrow$ 8,444 pairs
Require $D_{\text{init}} \leq$ 40th percentile (per seasonal phase) $\rightarrow$ 3,379 pairs
Apply hard quality gates (valid ratio $\geq$0.30, cloudy $\leq$20, overlap frames) per track
Stage 3: Stratified selection
for each track $\in$ {Meteorological, Vegetation Trajectory, Pixel-level} **do**
    for each seasonal phase $\in \{0, 20, 40\}$ **do**
        Select top-50 pairs by track score, with per-cube cap of 3
    end for
end for
$\mathcal{B}_{\text{sea}} \leftarrow$ union of all track selections, deduplicated $\rightarrow$ 422 pairs
return $\mathcal{B}_{\text{sea}}$, with per-pair track membership and divergence metadata

A.2.3 Extreme Summer Benchmark: Evaluation Metrics

In addition to the standard EarthNetScore (ENS), we report reconstruction and event-severity metrics that are computed on the 20-frame target period of each benchmark window. When a method produces multiple stochastic samples, we follow the main-paper protocol and evaluate the ensemble-mean prediction. Deterministic methods have $N = 1$.

Notation. Let $\hat{\mathbf{o}}^{(k)} \in \mathbb{R}^{C\times T\times H\times W}$, $k = 1, \ldots, N$, denote the N generated target sequences and $\mathbf{o} \in \mathbb{R}^{C\times T\times H\times W}$ the ground truth, where C is the number of spectral channels, H and W are the spatial dimensions, and $T = 20$ is the target length. We use c for channel index, t for target-frame index, and p for a spatial pixel index. The evaluated prediction is the ensemble mean $\bar{\hat{\mathbf{o}}} = N^{-1} \sum_{k=1}^{N} \hat{\mathbf{o}}^{(k)}$. Let $M_{t,p} \in \{0, 1\}$ be the target validity mask, where 1 denotes an observable pixel. We write $y_{t,p}$ and $\hat{y}_{t,p}$ for NDVI computed from the ground truth and from $\bar{\hat{\mathbf{o}}}$, respectively. Let $\text{Red}_{t,p}$ and $\text{NIR}_{t,p}$ denote the ground-truth red and near-infrared channel values at frame t and pixel p. Then:

$$y_{t,p} = \frac{\text{NIR}_{t,p} - \text{Red}_{t,p}}{\text{NIR}_{t,p} + \text{Red}_{t,p} + \epsilon}, \qquad \epsilon = 10^{-8}. \tag{A.6}$$

Table A.7: **Seasonal Matched-Pair Benchmark statistics.** (a) Overall composition. (b) Divergence score distributions across the selected pairs (scores are in normalized units after groupwise robust standardization).

(a) Composition

Attribute	Value
Total pairs	422
Unique geographic locations	380
Unique Sentinel-2 tiles	11
Track membership	
Meteorological Divergence	150
Vegetation Trajectory Divergence	150
Pixel-level Spatial Divergence	150
Seasonal phase distribution	
Phase 0 (early season)	145
Phase 20 (mid season)	142
Phase 40 (late season)	135

(b) Divergence score distributions

Score	Mean	Median	Max
D_{init}	-0.505	-0.466	-0.220
D_{meteo}	0.237	0.256	1.077
D_{veg}	0.212	0.142	4.530
D_{pixel}	0.478	0.259	3.951

D_{init}: negative values indicate close initial-state matching (lower = more similar). D_{meteo}, D_{veg}, D_{pixel}: higher values indicate greater divergence. All scores are in robust-normalized units (0 = group median, 1 $\approx$ 1 IQR above median).

The target-period vegetation mask is defined from the ground truth as

$$\mathcal{V} = \{(t,p) : y_{t,p} \geq 0.3,\ M_{t,p} = 1\}, \tag{A.7}$$

and $\mathcal{V}_t = \{p : (t,p) \in \mathcal{V}\}$ denotes the valid vegetated pixels at frame t.

Pixel MAE (P-MAE). Pixel-MAE is the mean absolute error over all valid pixels and spectral channels:

$$\text{P-MAE} = \frac{1}{C \sum_{t,p} M_{t,p}} \sum_{c=1}^{C} \sum_{t,p} M_{t,p} \left| \bar{\hat{o}}_{c,t,p} - o_{c,t,p} \right|. \tag{A.8}$$

NDVI MAE. NDVI-MAE evaluates vegetation-state accuracy over valid vegetated pixels:

$$\text{N-MAE} = \frac{1}{|\mathcal{V}|} \sum_{(t,p)\in\mathcal{V}} \left| \hat{y}_{t,p} - y_{t,p} \right|, \tag{A.9}$$

where $\hat{y}_{t,p}$ is computed from $\bar{\hat{\mathbf{o}}}$.

Trough NDVI MAE (TN-MAE). TN-MAE is the NDVI error at the ground-truth trough frame t^* recorded in the benchmark metadata:

$$\text{TN-MAE} = \frac{1}{|\mathcal{V}_{t^*}|} \sum_{p\in\mathcal{V}_{t^*}} \left| \hat{y}_{t^*,p} - y_{t^*,p} \right|. \tag{A.10}$$

This isolates prediction accuracy at the moment of maximum observed stress.

Drop Amplitude Error (DAE). For each window, the benchmark metadata provides a baseline NDVI $\bar{y}_{\text{base}}$ and a verified ground-truth drop amplitude Δ_{gt} from the construction stage (the stored `drop_amplitude` field). The prediction-side target trajectory is computed as the frame-wise mean NDVI of the prediction over valid ground-truth vegetation pixels:

$$\bar{\hat{y}}(t) = \frac{1}{|\mathcal{V}_t|} \sum_{p\in\mathcal{V}_t} \hat{y}_{t,p}. \tag{A.11}$$

The predicted decline amplitude is

$$\Delta_{\text{pred}} = \bar{y}_{\text{base}} - \min_{t:|\mathcal{V}_t|>0} \bar{\hat{y}}(t), \tag{A.12}$$

and the Drop Amplitude Error is

$$\text{DAE} = \left| \Delta_{\text{pred}} - \Delta_{\text{gt}} \right|. \tag{A.13}$$

This measures how accurately the model reproduces the magnitude of the vegetation decline, regardless of exact timing.

Drop Reproduction Accuracy (DRA). A naive threshold-crossing detection metric can reward models that systematically under-predict NDVI. We therefore use DRA as an auxiliary severity-calibration score in the visualization analysis. Let N_{samples} be the number of evaluated benchmark windows, and let Δ^i_{pred} and Δ^i_{gt} be the predicted and benchmark drop amplitudes for window i. Then:

$$\text{DRA} = \frac{1}{N_{\text{samples}}} \sum_{i=1}^{N_{\text{samples}}} \max\left(0,\ 1 - \frac{|\Delta^i_{\text{pred}} - \Delta^i_{\text{gt}}|}{\Delta^i_{\text{gt}}}\right). \tag{A.14}$$

DRA $\in [0, 1]$ (higher is better). A score of 1.0 means the predicted decline amplitude exactly matches the benchmark drop amplitude, while predictions whose absolute error exceeds Δ_{gt} receive zero credit.

Severity-bin aggregation. The low-, mid-, and high-severity results in the main paper are obtained by splitting benchmark windows into three bins using the 33.3rd and 66.7th percentiles of the composite extreme score s_{extreme}, rather than by the drop amplitude alone. Within each bin, TN-MAE and DAE are averaged over per-window metric values.

A.2.4 Seasonal Matched-Pair Benchmark: Evaluation Metrics

The Seasonal Matched-Pair metrics measure how well forecasts reproduce observed future differences between matched cross-year windows. Each pair (A, B) shares the same geographic location and seasonal phase and has similar observed contexts. We compare predicted NDVI divergence with observed NDVI divergence. The formulas do not directly use meteorological variables. As above, stochastic methods are evaluated with the ensemble-mean prediction for each window. All metrics operate on the spatially-averaged NDVI trajectory over vegetation pixels (NDVI ≥ 0.3 and valid).

Notation. For each window $w \in \{A, B\}$ in a pair, let $\bar{y}^w_{\text{gt}}(t)$ denote the ground-truth spatially-averaged NDVI at target frame t, and $\bar{\hat{y}}^w(t)$ the predicted NDVI from the ensemble-mean forecast, spatially averaged over vegetation pixels. Define the per-frame ground-truth divergence $d^{\text{gt}}_t = \bar{y}^A_{\text{gt}}(t) - \bar{y}^B_{\text{gt}}(t)$ and predicted divergence $d^{\text{pred}}_t = \bar{\hat{y}}^A(t) - \bar{\hat{y}}^B(t)$. Let $\mathcal{T}$ denote the set of target frames where both sides have valid observations (finite NDVI values). Here $t \in \{1, \ldots, T_{\text{out}}\}$ indexes the target frames, i indexes benchmark pairs, $\mathbf{1}\{\cdot\}$ denotes the indicator function, $\text{sign}(\cdot)$ returns the sign of its argument, and $\rho_S(\cdot, \cdot)$ denotes Spearman's rank correlation.

Divergence Reproduction Ratio (DRR). DRR measures whether the model reproduces the correct *magnitude* of vegetation divergence between paired windows. We compute absolute divergences and filter by a noise threshold $\tau = 0.02$ to exclude frames where the ground-truth difference is negligible:

$$\text{DRR} = \frac{\overline{|d^{\text{pred}}_t|}_{t \in \mathcal{T}_\tau}}{\overline{|d^{\text{gt}}_t|}_{t \in \mathcal{T}_\tau}}, \quad \text{where } \mathcal{T}_\tau = \{t \in \mathcal{T} : |d^{\text{gt}}_t| > \tau\}, \tag{A.15}$$

and the overline denotes temporal averaging. Following the main-paper tables, we report DRR_{mean}, the mean of the per-pair DRR values. The median DRR is also computed as a robust diagnostic. DRR $= 1.0$ is ideal. Values below 1 indicate underestimation of observed divergence, while values above 1 indicate overestimation.

Directional Hit Rate (DHR). DHR measures whether the model correctly predicts *which window has lower NDVI* at each timestep, conditioned on the ground-truth difference exceeding the noise floor:

$$\text{DHR} = \frac{1}{|\mathcal{T}_\tau|} \sum_{t \in \mathcal{T}_\tau} \mathbf{1}\Big\{\text{sign}(d^{\text{pred}}_t) = \text{sign}(d^{\text{gt}}_t)\Big\}. \tag{A.16}$$

DHR $\in [0, 1]$ (higher is better). A score of 0.5 corresponds to uniform random sign prediction. Higher scores indicate more accurate prediction of which cross-year window has lower observed NDVI. We aggregate DHR across all pairs by pooling the hit counts: $\text{DHR}_{\text{agg}} = \sum_i n^i_{\text{hits}} \,/\, \sum_i |\mathcal{T}^i_\tau|$, where n^i_{hits} is the number of correct-sign timesteps for pair i and $\mathcal{T}^i_\tau$ is its thresholded valid-frame set. This weights each pair proportionally to its number of valid comparison frames.

Paired Divergence Correlation (PDC). PDC measures the agreement between the predicted and observed *rankings* of future divergence across pairs. For each pair i, let $\mathcal{T}_i$ denote its valid comparison-frame set. We compute the total (time-summed) absolute divergence:

$$D_i^{\text{gt}} = \sum_{t \in \mathcal{T}_i} |d_{i,t}^{\text{gt}}|, \qquad D_i^{\text{pred}} = \sum_{t \in \mathcal{T}_i} |d_{i,t}^{\text{pred}}|. \tag{A.17}$$

PDC is the Spearman rank correlation between the two vectors across all N_{pairs} evaluated pairs:

$$\text{PDC} = \rho_S\left(\{D_i^{\text{gt}}\}_{i=1}^{N_{\text{pairs}}},\ \{D_i^{\text{pred}}\}_{i=1}^{N_{\text{pairs}}}\right). \tag{A.18}$$

PDC $\in [-1, 1]$ (higher is better). A high PDC indicates that pairs with larger observed future divergence also tend to have larger predicted divergence. It measures ranking agreement across pairs, not calibration against the magnitude of weather differences.

Complementarity of the three metrics. The three metrics describe complementary aspects of matched-pair forecast-divergence fidelity:

- **DRR**: reproduction of observed divergence magnitude.
- **DHR**: accuracy of the predicted direction of observed divergence.
- **PDC**: agreement between predicted and observed divergence rankings across pairs.

A model can predict the correct sign while underestimating divergence magnitude. It can also preserve the ranking across pairs while systematically underestimating absolute divergence. The three metrics jointly describe how forecasts reproduce observed cross-year differences. They do not establish a causal weather-response relationship.

Track-level evaluation. All metrics are computed both globally (across all 422 pairs) and per track (Meteorological Divergence, Vegetation Trajectory Divergence, Pixel-level Spatial Divergence).

A.2.5 Benchmark Selection Sensitivity

We evaluate five Extreme Summer configurations (E0–E4) and seven Seasonal Matched-Pair configurations (S0–S6). These include the two default configurations and ten variants. Each variant changes the indicated selection rule while retaining the remaining construction settings. The forecasting models and evaluation protocol are unchanged.

Extreme Summer. E0 uses a sequence-level decline threshold of 0.35 and a mean target-period valid-pixel ratio threshold of 0.30. E1 and E2 change the decline threshold to 0.30 and 0.40. E3 and E4 change the target-period validity threshold to 0.25 and 0.35. These changes apply to sample prefiltering. The baseline-relative event verification threshold remains 0.10 in all five settings.

Seasonal Matched-Pair. S0 uses the 40th percentile for initial-state matching and a frame-level valid-pixel ratio threshold of 0.30. S1 and S2 change the matching quantile to 0.30 and 0.50. S3 and S4 change the frame-level validity threshold to 0.25 and 0.35 during trajectory construction. S5 and S6 change the minimum clean-frame counts for vegetation-trajectory and pixel-level comparisons. The default counts are 8 and 6. S5 lowers them to 6 and 4, while S6 raises them to 10 and 8. The full-window mean validity threshold of 0.30 and the maximum of 20 cloudy frames remain unchanged.

Tables A.8 and A.9 show stable comparisons on the main severity and matched-pair metrics. Some individual metrics favor other models. Changing a threshold also changes the selected subset. The results therefore describe robustness to these selection choices within the two diagnostic benchmarks.

A.2.6 Probabilistic Forecasting and Calibration Diagnostics

The main comparison tables use five-sample ensemble means. Here, we evaluate $K = 19$ independent forecasts per input to examine predictive distributions and uncertainty calibration. We first report raw-ensemble results, followed by an exploratory scalar recalibration on separate IID data.

Table A.8: **Extreme Summer selection sensitivity.** E0 is the default. E1/E2 vary the sequence-level decline threshold. E3/E4 vary the target-period validity threshold. Bold marks the best result within each configuration. EO-WM achieves the best ENS and DAE in all five configurations.

Variant	Model	ENS↑	N-MAE↓	TN-MAE↓	DAE↓
E0	Earthformer	0.2539	**0.1099**	0.1197	0.2704
	Wan2.1	0.2028	0.1345	0.1381	0.2483
	EO-WM	**0.2543**	0.1106	**0.1179**	**0.2330**
E1	Earthformer	0.2554	**0.1085**	**0.1171**	0.2539
	Wan2.1	0.2041	0.1327	0.1379	0.2330
	EO-WM	**0.2560**	0.1100	0.1185	**0.2246**
E2	Earthformer	0.2523	**0.1114**	0.1225	0.2885
	Wan2.1	0.2013	0.1341	0.1392	0.2648
	EO-WM	**0.2550**	0.1117	**0.1210**	**0.2493**
E3	Earthformer	0.2508	0.1122	0.1229	0.2759
	Wan2.1	0.2055	0.1333	0.1356	0.2504
	EO-WM	**0.2558**	**0.1117**	**0.1220**	**0.2358**
E4	Earthformer	0.2584	**0.1058**	0.1162	0.2529
	Wan2.1	0.1995	0.1337	0.1411	0.2389
	EO-WM	**0.2587**	0.1067	**0.1131**	**0.2347**

Table A.9: **Seasonal Matched-Pair selection sensitivity.** S0 is the default. S1/S2 vary initial-state matching, S3/S4 vary frame-level validity, and S5/S6 vary clean-frame requirements. Bold marks the best result within each configuration. EO-WM achieves the best ENS and DHR in all seven configurations. Wan2.1 has the best PDC in S5.

Variant	Model	ENS↑	$\mathrm{DRR}_{\mathrm{mean}}\to 1$	DHR↑	$\mathrm{PDC}_{\mathrm{sp}}$↑
S0	Earthformer	0.2879	0.6302	0.5551	0.1814
	Wan2.1	0.2537	**0.7066**	0.6050	0.2809
	EO-WM	**0.3022**	0.6927	**0.6522**	**0.2942**
S1	Earthformer	0.2916	0.6423	0.5495	0.1664
	Wan2.1	0.2574	0.7782	0.6045	0.2991
	EO-WM	**0.3057**	**0.7983**	**0.6686**	**0.3517**
S2	Earthformer	0.2856	0.6290	0.5680	0.1728
	Wan2.1	0.2540	0.7269	0.6164	0.2797
	EO-WM	**0.2995**	**0.7525**	**0.6458**	**0.3477**
S3	Earthformer	0.2879	0.6325	0.5585	0.1822
	Wan2.1	0.2541	0.7462	0.6058	0.3193
	EO-WM	**0.3013**	**0.7590**	**0.6484**	**0.3530**
S4	Earthformer	0.2880	0.6440	0.5612	0.1685
	Wan2.1	0.2543	0.7615	0.5991	0.2556
	EO-WM	**0.3020**	**0.7866**	**0.6421**	**0.3337**
S5	Earthformer	0.2854	0.6411	0.5541	0.2265
	Wan2.1	0.2503	0.7473	0.5944	**0.3233**
	EO-WM	**0.2960**	**0.7568**	**0.6528**	0.3136
S6	Earthformer	0.2935	0.5966	0.5738	0.2484
	Wan2.1	0.2574	0.7269	0.6028	0.2782
	EO-WM	**0.3065**	**0.7738**	**0.6544**	**0.3907**

Pixel-level ensemble diagnostics. We evaluate target-period NDVI over valid vegetation pixels. Let z_i be the observed NDVI and $\hat{z}_i^{(k)}$ the prediction from member k. The finite-ensemble-corrected fair CRPS is

$$\text{F-CRPS}_i = \frac{1}{K}\sum_{k=1}^{K}|\hat{z}_i^{(k)} - z_i| - \frac{1}{2K(K-1)}\sum_{k=1}^{K}\sum_{k'\neq k}|\hat{z}_i^{(k)} - \hat{z}_i^{(k')}|. \tag{A.19}$$

We also report the spread-skill ratio (SSR), which compares ensemble spread with the RMSE of the ensemble mean, and empirical 90% coverage (Cov90). Their reference values are 1 and 0.90. ENS and N-MAE evaluate the ensemble mean.

Table A.10: **Raw $K = 19$ ensemble diagnostics.** F-CRPS and coverage are pixel-level NDVI diagnostics over valid target vegetation pixels.

Benchmark	Model	ENS ↑	N-MAE ↓	F-CRPS ↓	SSR → 1	Cov90 → 0.90
Extreme	Wan2.1	0.2071	0.1276	0.0879	0.824	0.724
	EO-WM	0.2565	0.1088	0.0861	0.737	0.654
Seasonal	Wan2.1	0.2588	0.1052	0.0762	0.741	0.774
	EO-WM	0.3044	0.0883	0.0730	0.685	0.711

EO-WM achieves lower F-CRPS and better ensemble-mean accuracy on both benchmarks. Its nonzero spread supports stochastic ensemble forecasting. Both models remain under-dispersed, with SSR below 1 and Cov90 below 0.90. Wan2.1 has greater spread and coverage than EO-WM. These results do not establish calibrated predictive uncertainty for the raw ensemble.

Spatial event probabilities. For each case, the 19 members define an empirical probability that a valid vegetation pixel experiences an NDVI drop of at least 0.10. We compute AUPRC and Brier score across pixels within each case, then macro-average across cases. Table A.11 reports paired differences as EO-WM minus Wan2.1, with 95% confidence intervals.

Table A.11: **Pixel-level vegetation-decline probabilities.** Model scores are listed as EO-WM / Wan2.1.

Benchmark	Metric	EO-WM / Wan2.1	Paired difference [95% CI]
Extreme	AUPRC ↑	0.9588 / 0.9390	$+0.0198$ $[+0.0175, +0.0221]$
	Brier ↓	0.0912 / 0.0957	-0.0045 $[-0.0063, -0.0027]$
Seasonal	AUPRC ↑	0.6319 / 0.6103	$+0.0216$ $[+0.0165, +0.0269]$
	Brier ↓	0.2123 / 0.2633	-0.0510 $[-0.0624, -0.0398]$

EO-WM improves both metrics on both benchmarks. Extreme cases have a high event prevalence, so their AUPRC mainly reflects within-scene localization. These results support informative spatial event probabilities. They do not establish fully calibrated event probabilities or prediction intervals.

Scalar NDVI-drop prediction. We reduce each member to one target-period NDVI-drop amplitude per case. We evaluate the ensemble center with MAE and the scalar ensemble with fair CRPS. These case-level metrics are distinct from the pixel-level diagnostics above. EO-WM obtains lower errors on both benchmarks (Table A.12).

Table A.12: **Raw $K = 19$ scalar NDVI-drop predictions.** Both metrics are better when lower.

Benchmark	Model	Center MAE	Scalar fair CRPS
Extreme	EO-WM	0.07241	0.05549
	Wan2.1	0.07848	0.06510
Seasonal	EO-WM	0.06482	0.05654
	Wan2.1	0.08197	0.05981

Exploratory IID-only scalar recalibration. We select 512 tile-balanced cases from the EarthNet2021 IID set with no overlap with either diagnostic benchmark. Both models use the same calibration cases. For each model, we fit one scalar factor α and rescale the NDVI-drop members $d^{(k)}$ around their mean $\bar{d}$:

$$\tilde{d}^{(k)} = \bar{d} + \alpha(d^{(k)} - \bar{d}). \tag{A.20}$$

This transformation leaves the ensemble center unchanged. We search $\alpha \in [0.50, 4.00]$ in steps of 0.01 to match 90% empirical coverage on the IID subset. Ties favor the value closest to 1. For

Table A.13: **EO-WM scalar recalibration with a frozen IID-fitted factor.** Coverage is best near 0.90. Lower fair CRPS and interval score are better. The predictive center remains unchanged.

Benchmark	Variant	Fair CRPS	Cov90	90% interval score	Center MAE
Extreme	Raw	0.05549	0.623	0.803	0.07241
	IID-recalibrated	0.05335	0.928	0.506	0.07241
Seasonal	Raw	0.05654	0.694	0.836	0.06482
	IID-recalibrated	0.04767	0.871	0.444	0.06482

$K = 19$, scalar Cov90 uses the ensemble min–max interval. Each fitted factor is frozen and applied to both benchmarks. The fitted values are 1.89 for EO-WM and 1.22 for Wan2.1.

The IID-fitted factor improves EO-WM's scalar coverage and distributional scores on both benchmarks (Table A.13). Under the same protocol, Wan2.1's scalar Cov90 changes from 0.677 to 0.915 on Extreme and from 0.748 to 0.887 on Seasonal. This exploratory correction applies only to scalar NDVI-drop amplitudes. It does not calibrate pixel-event probabilities or the full multispectral predictive distribution. We retain the raw ensemble as the primary probabilistic evaluation.

A.3 Comparison Method Adaptations

A.3.1 Model Size and Inference Cost

We profile the models on one A100-80GB GPU under the EarthNet2021 forecasting setting. For this controlled efficiency comparison, all diffusion models use bfloat16 and a common budget of 30 sampling steps. Deterministic models use float32 and one forward pass. Latency is the CUDA-synchronized median after warm-up. Table A.14 reports latency per window for one and five diffusion samples, together with total parameter counts and peak GPU memory. The main accuracy comparisons use five samples for diffusion models with model-specific sampling settings. In particular, Latte uses 250 DDIM steps for accuracy evaluation; its efficiency measurements here use 30 steps.

Table A.14: **Model size and inference cost.** A deterministic forecast uses one forward pass. Diffusion latency is reported for one / five samples at a common 30-step profiling budget. Latte's accuracy evaluation uses 250 DDIM steps.

Model	Total params (M)	Inference	Latency (s/window)	Peak memory (GB)
PredRNNv2	6.61	1 pass	0.084	0.47
TAU	15.22	1 pass	0.012	0.38
Earthformer	60.62	1 pass	0.060	1.48
Latte	757.45	30 steps	1.876 / 8.712	8.96
Wan2.1	1,696.39	30 steps	1.982 / 9.214	15.03
EO-WM	482.55	30 steps	1.823 / 8.206	14.64

Table A.15: **Diffusion-model parameter and computation breakdown.** MACs are measured with fvcore for one DiT step at batch size 1 using the full latent sequence. They exclude VAE computation.

Model	DiT params (M)	VAE params (M)	DiT MACs/step (G)
Latte	673.79	83.66	873.13
Wan2.1	1,569.49	126.90	798.05
EO-WM	387.04	95.51	379.19

EO-WM has the smallest DiT and lowest DiT MACs per step among the evaluated diffusion models (Table A.15). Its 387.04M DiT parameters and 95.51M VAE parameters sum to 482.55M. DiT computation scales with denoising steps and sample count. The lower DiT cost does not imply the lowest peak memory, as Latte uses less memory in this profiling setup. Deterministic models are substantially faster under their one-pass inference protocol.

A.3.2 Wan2.1 Adaptation

To evaluate whether a strong general-purpose video generation model can serve as an effective baseline for Earth surface forecasting, we adapt Wan2.1-Fun-V1.1-1.3B-InP [48] (abbreviated as Wan2.1-Inp), a 1.3B-parameter latent video diffusion transformer originally designed for video inpainting and prediction, to the EarthNet2021 task. This adaptation requires non-trivial architectural modifications to handle 4-channel multi-spectral input (B, G, R, NIR) and to inject geospatial and meteorological conditioning, followed by a carefully staged fine-tuning procedure to preserve the pretrained generative capabilities while specializing to the satellite domain.

We treat Wan2.1 as a pretrained reference. Its pretraining data and total training budget differ from those of EO-WM, so this comparison does not control for data and compute. Accuracy comparisons use five-sample ensemble means for diffusion models, with model-specific sampling settings. Task-specific baselines trained under the same EarthNet2021 protocol provide more controlled comparisons.

Base model architecture. Wan2.1-Inp is built on a 3D Diffusion Transformer (DiT) with flow matching training [24]. The architecture consists of 30 transformer blocks with hidden dimension 1536, 12 attention heads, and a feed-forward dimension of 8960. Video inputs are patchified with a 3D patch size of $(1, 2, 2)$ (temporal, height, width). The model uses a causal 3D VAE that compresses video spatially by $8\times$ and temporally by $4\times$, producing 16-channel latent representations. As an inpainting model, it accepts both noisy latents and condition-frame latents concatenated along the channel dimension (input dimension $= 16 + 16 + 1_{\text{mask}} + 3_{\text{pad}} = 36$). Conditioning is injected via AdaLN modulation from the diffusion timestep embedding, and text guidance is provided through cross-attention from a T5-based text encoder.

Architectural modifications for EO adaptation. We introduce two categories of modifications to adapt the model for the EarthNet2021 satellite prediction task:

(1) VAE channel expansion. The original VAE operates on 3-channel RGB video. We extend it to 4-channel input (B, G, R, NIR) by expanding the encoder's first convolutional layer from CausalConv3d$(3 \to d)$ to CausalConv3d$(4 \to d)$, and the decoder's output layer from CausalConv3d$(d \to 3)$ to CausalConv3d$(d \to 4)$. The weights for the new 4th channel (NIR) are initialized as the average of the Red and Green channel weights, providing a reasonable starting point that leverages the pretrained spectral representations.

(2) Earth observation conditioning modules. We add four conditioning pathways to the DiT to inject geospatial and meteorological information:

- *Global geospatial embedding.* A 6-dimensional vector encoding the tile's geographic location (3D spherical coordinates from latitude/longitude) and temporal position (cyclical day-of-year encoding and normalized year) is projected through a 2-layer MLP and added to the diffusion timestep embedding before AdaLN modulation, providing location- and season-aware generation.
- *Per-frame temporal embedding.* A 3-dimensional per-frame vector (cyclical day-of-year, normalized year) is projected to the hidden dimension and added to each token according to its temporal position, enabling frame-level temporal awareness.
- *DEM spatial encoder.* A 3-layer stride-2 convolutional encoder processes the digital elevation model (1 channel + validity mask) from 128×128 to 16×16 feature maps, which are broadcast across the temporal dimension, patchified to match the transformer's token grid, linearly projected, and added to the token embeddings.
- *ERA5 meteorological encoder.* A 3-layer stride-2 convolutional encoder processes 5-channel ERA5 reanalysis data (precipitation, pressure, mean/min/max temperature, plus a validity mask) per frame. The resulting spatiotemporal features are patchified, projected, and added to the token stream alongside the DEM features.

All new modules are initialized with near-zero output (zero bias, small-norm weights) to ensure that the pretrained generation behavior is preserved at initialization.

Four-stage fine-tuning procedure. To transfer the strong video generation prior of Wan2.1 to the satellite domain without catastrophic forgetting, we design a progressive four-stage fine-tuning strategy that gradually unfreezes model capacity:

Stage 1: VAE warm-up. Only the newly expanded input and output convolutional layers of the VAE are trained (approximately 2M parameters), while all intermediate encoder and decoder layers remain frozen. This allows the NIR channel weights to adapt to the VAE's internal feature space without disrupting the pretrained reconstruction capability for the RGB channels. The model is trained for 5 epochs with learning rate 5×10^{-5} using MSE reconstruction loss with KL regularization (weight 10^{-6}), masked by per-pixel quality flags to exclude cloud-contaminated observations.

Stage 2: VAE full fine-tuning. Starting from the Stage 1 checkpoint, all VAE parameters (approximately 100M) are unfrozen and trained for 30 epochs at a reduced learning rate of 1×10^{-5} with cosine annealing. This allows the intermediate layers to fully adapt to the statistical properties of 4-channel satellite imagery. After this stage, we recompute the per-channel latent normalization statistics (mean and standard deviation across 2,000 training samples) to replace the original RGB-video statistics, ensuring stable DiT training.

Stage 3: DiT LoRA warm-up. The pretrained DiT weights are frozen, and we train only the newly added EO conditioning modules (approximately 2M parameters) together with rank-32 LoRA adapters [17] attached to the query, key, value, and feed-forward projections of all transformer blocks (approximately 15M parameters). This stage runs for 20,000 steps with learning rate 1×10^{-4}, effective batch size 32, and cosine scheduling with 500 warm-up steps. The quality mask is applied to the diffusion loss to exclude invalid pixels.

Stage 4: DiT full fine-tuning. Starting from the merged LoRA checkpoint, all 1.3B DiT parameters are unfrozen and trained for 50,000 steps at a lower learning rate of 5×10^{-6} to refine the full model. An exponential moving average (EMA) with decay 0.9999 is maintained for stable evaluation. Gradient checkpointing and bfloat16 mixed precision are used throughout to fit within GPU memory.

Training hyperparameters. Table A.16 summarizes the key hyperparameters across all four training stages.

Table A.16: **Wan2.1-Inp fine-tuning hyperparameters** across four progressive stages.

	Stage 1 (VAE warm-up)	Stage 2 (VAE full)	Stage 3 (DiT LoRA)	Stage 4 (DiT full)
Trainable params	2M	100M	17M	1.3B
Optimizer	AdamW			
Learning rate	5×10^{-5}	1×10^{-5}	1×10^{-4}	5×10^{-6}
Effective batch size	16	16	32	32
Training duration	5 epochs	30 epochs	20k steps	50k steps
LR scheduler	Cosine annealing			
Warmup steps	—	—	500	500
Gradient norm clip	1.0			
Mixed precision	bfloat16			
EMA decay	—	—	—	0.9999
LoRA rank	—	—	32	—
Gradient checkpointing	—	—	✓	✓

Inference configuration. At inference time, we use 30-step Euler ODE integration with a flow-matching scheduler (shift $= 5.0$). The model receives 10 context frames and generates 20 future frames at 128×128 resolution with 4 spectral channels. Classifier-free guidance is disabled for all Wan2.1 evaluations (guidance scale $= 1.0$).

A.3.3 Latte Adaptation

To make Latte [27] condition-consistent with the EO forecasting setting, we adapt it with an explicit cross-attention pathway for non-visual EO conditions. We keep Latte's video diffusion transformer backbone and repaint-style video-to-video prediction, but augment each transformer block with a lightweight cross-attention module whose keys and values are produced from future meteorological and geospatial condition tokens. This turns Latte from a visual-context-only video diffusion baseline into a weather-conditioned latent video diffusion model.

The adaptation follows a two-stage training procedure. First, we train a 4-channel VAE from scratch on EarthNet2021 B/G/R/NIR frames, using masked reconstruction loss over valid pixels. The VAE uses 4 latent channels and is trained with learning rate 10^{-4} and KL weight 10^{-6} until convergence, which occurs at approximately 50k steps. Second, we train a Latte-XL/2 DiT in the learned latent space for 30-frame EarthNet clips, matching the 10-context/20-target protocol used by the other methods. The DiT uses the Latte-XL/2 configuration with patch size 2, 28 transformer blocks, hidden size 1152, 16 attention heads, and learned variance prediction. Accuracy evaluation uses 250 DDIM sampling steps. For the controlled efficiency comparison in Sec. A.3.1, Latte is profiled with 30 sampling steps.

For condition injection, ERA5 variables are sampled at the Sentinel-2 frame times and encoded by a small spatiotemporal convolutional encoder at the latent resolution. DEM is encoded by a separate spatial convolutional encoder and broadcast across time. Geospatial and seasonal metadata are projected by an MLP and appended as global condition tokens. The resulting condition sequence is projected to the Latte hidden dimension and used as the encoder context for cross-attention after the self-attention sublayer in each spatial transformer block. The cross-attention output is zero-initialized through a gated residual projection.

At inference time, the 10 observed context frames are encoded into latents and reinserted during the reverse diffusion process, while the remaining 20 latent frames are sampled under the same future ERA5, DEM, and metadata conditions used by the other EO forecasting baselines. We train the conditioned Latte variant for 200k steps with AdamW, learning rate 10^{-4}, local batch size 4, and 1k warm-up steps.

A.3.4 OpenSTL Deterministic Baselines

We evaluate five deterministic spatiotemporal prediction baselines from OpenSTL [41]: SimVP [7], TAU [40], PredRNN [50], PredRNNv2 [51], and PhyDNet [10]. These methods cover CNN-based, recurrent, and physics-informed prediction paradigms and are adapted to the EarthNet2021 setting with 10 observed frames and 20 future frames at 128×128 resolution.

EarthNet input/output adaptation. Each input frame contains 10 channels: 4 Sentinel-2 satellite bands (B, G, R, NIR), 5 ERA5 mesodynamic variables, and a static elevation channel. The ERA5 variables are temporally aligned to the Sentinel-2 timestamps and spatially upsampled to 128×128; the static elevation channels are repeated over time. All models are trained to predict the 4 satellite bands. For training, the target tensor additionally carries a cloud/quality mask and the 6-channel condition sequence.

FiLM conditioning. In the film OpenSTL adaptation, meteorological and static conditions are explicitly provided during the prediction horizon, instead of conditioning only on the observed-context weather. For SimVP, the encoder and decoder operate on the 4 satellite channels, while the 6 auxiliary channels are encoded by a small convolutional condition encoder and injected into the temporal processor through FiLM [32] modulation. For recurrent models, including PredRNN and PredRNNv2, we apply per-frame FiLM before the recurrent backbone. The full 30-frame condition sequence is formed by concatenating the 10 observed-context conditions with the 20 future conditions, and each satellite frame is modulated with the condition at the corresponding time step. PhyDNet uses the same per-frame FiLM mechanism for both context frames and teacher-forced target frames during training; at inference, the future FiLM modulation still injects weather information through the generated sequence. TAU follows the OpenSTL SimVP-style implementation and receives the 6 condition channels by concatenation at each rollout.

Prediction rollout. For SimVP and TAU, which naturally predict an output sequence with the same length as the input sequence, the 20-frame horizon is generated by two 10-frame auto-regressive rollouts. The first rollout predicts frames 11–20 from the observed frames 1–10 using future conditions for frames 11–20. The second rollout predicts frames 21–30 from the previously predicted satellite frames, with the future conditions for frames 21–30 reattached. Recurrent and physics-informed methods generate the full 20-frame horizon using their native OpenSTL training and inference procedures, with the same future-condition sequence supplied through FiLM.

Training details. All OpenSTL baselines are trained from scratch on the EarthNet2021 training set for 200 epochs. We use cloud-masked MSE loss on the 4 satellite output channels, where pixels marked invalid by the quality mask are excluded from the effective loss. The film training recipe uses AdamW, weight decay 10^{-5}, cosine learning-rate scheduling, 10 warm-up epochs, minimum learning rate 10^{-5}, and gradient clipping at 1.0. The best checkpoint is selected by validation loss and used for all evaluations.

A.4 Data and Asset Availability

Released benchmark files. The proposed Extreme Summer and Seasonal Matched-Pair benchmarks are derived from the public EarthNet2021 test splits. At submission time, we provide the inference CSV files used in our experiments, which specify the selected benchmark windows, pair identities, split membership, track labels, and metadata needed to run evaluation on the corresponding EarthNet2021 samples. The CSV files do not redistribute the raw EarthNet2021 imagery or weather data, users should obtain the raw data from the official EarthNet2021 source and apply the terms of the original dataset.

Existing assets and licenses. We use and cite publicly available datasets, model weights, and codebases. EarthNet2021 [33] is distributed under the CC-BY-NC-SA 4.0 license according to its official dataset page (`https://earthnet.tech/resources/datasets/earthnet2021`). EarthNet2021 includes Copernicus Sentinel data, whose access and use are governed by the Copernicus free, full, and open data policy as documented by the EarthNet2021 dataset page. The EO-VAE tokenizer [22] is listed as Apache-2.0 on its official Hugging Face release (`https://huggingface.co/nilsleh/eo-vae`). Open-Sora [60] is released under Apache-2.0 on its official GitHub repository (`https://github.com/hpcaitech/Open-Sora`). Wan2.1-Fun-V1.1-1.3B-InP [48] is listed as Apache-2.0 on the official Alibaba-PAI Hugging Face release (`https://huggingface.co/alibaba-pai/Wan2.1-Fun-V1.1-1.3B-InP`). Latte [27], OpenSTL [41], and Earthformer [8] are also released under Apache-2.0 according to their official repositories (`https://github.com/Vchitect/Latte`, `https://github.com/chengtan9907/OpenSTL`, and `https://github.com/amazon-science/earth-forecasting-transformer`). We use these assets only for research benchmarking and adaptation, and we respect the corresponding attribution and license terms.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe weather-conditioned probabilistic EO forecasting, physically motivated conditioning features, and two diagnostic benchmarks. The claims are supported by the reported comparisons and ablations. Seasonal Matched-Pair is presented as an observational diagnostic.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion discusses limited geographic coverage, benchmark size, unobserved interannual factors, and unvalidated longer-horizon and downstream performance. Appendix A.2.6 discusses raw-ensemble under-dispersion and the scope of scalar recalibration.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results requiring proofs. Its equations define the forecasting task, training objective, conditioning features, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper and appendices describe the model architecture, forecasting protocol, benchmark construction, metrics, training settings, and baseline adaptations. EO-WM code and model weights are publicly available at `https://github.com/hku-sail/EO-WM`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: EO-WM code and model weights are publicly available at `https://github.com/hku-sail/EO-WM`. The raw imagery and weather data are available through EarthNet2021. The provided benchmark CSV files specify the evaluation windows and pairs, as described in Appendix A.4.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main paper specifies the 10-context/20-target forecasting protocol. Appendices A.1, A.2, and A.3 provide training settings, masking rules, checkpoint selection, benchmark filtering criteria, and baseline adaptation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Appendix Table A.11 reports paired 95% confidence intervals for EO-WM minus Wan2.1 differences in pixel-event AUPRC and Brier score on both benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix A.3.1 reports the profiling GPU, numerical precision, sampling budget, inference latency, peak memory, parameter counts, and DiT MACs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public Earth-observation and meteorological data for scientific forecasting and evaluation. It does not involve private personal data or human-subject research and conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The conclusion discusses potential uses in ecosystem monitoring, crop-growth prediction, and climate-risk assessment. It also identifies the risk of overreliance on forecasts in agricultural, insurance, and disaster-response decisions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released model forecasts multispectral Earth-observation sequences for environmental research. The benchmark files reference public satellite data. These assets do not present the high-risk misuse scenarios addressed by this question.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper credits the datasets, models, and codebases used. Appendix A.4 lists the licenses and terms of use for EarthNet2021, Copernicus Sentinel data, EO-VAE, Open-Sora, Wan2.1, Latte, OpenSTL, and Earthformer.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The public release provides EO-WM code and model weights. The paper and appendices document the model, training settings, and benchmark construction, filtering criteria, statistics, and metrics. The provided benchmark CSV files identify the evaluation windows and pairs.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing, participant studies, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human-subject research, so IRB or equivalent approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not an important, original, or non-standard component of the core method. Auxiliary captions use fixed templates and provide optional backbone context.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.