EARTHMIND: PERSISTENCE-ANCHORED LATENT TRANSITIONS FOR ROBUST MULTIMODAL EARTH OBSERVATION FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting Earth observations is hampered by a practical fact: the inputs are routinely incomplete. In our globally distributed benchmark, only 90.3% of Sentinel-2 months and 82.6% of Sentinel-1 months are usable. We present EarthMind, a multimodal spatiotemporal world model that frames monthly forecasting as learning a transition from a persistence-anchored state rather than regenerating the scene from scratch. EarthMind fuses Sentinel-2 multispectral imagery, Sentinel-1 SAR, static terrain, and ten weather variables into latent Earth states. It predicts the next state with a Transformer and decodes that state into future Sentinel-2 and Sentinel-1 imagery and weather. Two extensions target real-world use. Structured modality masking (EarthMind-Robust) reproduces realistic sensor outages during training. Band-wise calibration toward persistence (EarthMind+) learns how much of the predicted change each spectral band can be trusted with. On 1,000 sites with site-disjoint splits (16,562 test sequences), EarthMind, trained from scratch, performs comparably to the pretrained Prithvi-EO-2.0 foundation model on clean inputs, with no significant difference on any metric, while additionally forecasting SAR and weather. EarthMind+ attains the lowest RMSE (0.0676), highest PSNR (23.40 dB) and SSIM (0.9704), and lowest SAM (4.68°) among the evaluated methods, while Prithvi retains the lowest MAE. When SAR and weather are both missing, structured masking reduces RMSE by 35.3%. In four-month closed-loop rollouts, EarthMind stays 42.8% below last-observation-carried-forward, and its frozen latent state improves land-cover balanced accuracy by 12.2 points over raw features without any land-cover supervision. The learned calibration coefficients further identify the shortwave-infrared bands as the least predictable.

1 INTRODUCTION

Earth observation (EO) data are multimodal, temporal, and incomplete (Wang et al., 2023b; Fuller et al., 2023). Optical imagery captures spectral surface appearance but is obscured by clouds, while synthetic aperture radar (SAR) can acquire observations largely independently of cloud cover and solar illumination (Ebel et al., 2022). Terrain supplies static geographic context, and weather provides information about environmental conditions that influence surface change (Hersbach et al., 2020; Muñoz-Sabater et al., 2021). Forecasting next month's observation therefore requires combining heterogeneous sensors into a shared state and predicting how that state evolves (Requena-Mesa et al., 2021; Benson et al., 2024; Gao et al., 2022b). In practice, the inputs needed to do so are often missing (Ma et al., 2022; Lee et al., 2023; Wang et al., 2023a): in our dataset, only 90.3% of site–months have usable Sentinel-2 observations and 82.6% have Sentinel-1 observations. Discarding every incomplete sequence would remove much of the record and would not reflect deployment conditions.

Two properties of this problem shape our design. First, **persistence is a strong prior**: at monthly scale, the latest observation is already a reasonable forecast, and a model that regenerates the full scene from a compressed code must spend capacity re-learning what is already visible. Second, **missingness is structured**: sensors can be unavailable in characteristic combinations, including cloud-driven optical gaps and SAR acquisition gaps. Missing weather records introduce another

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

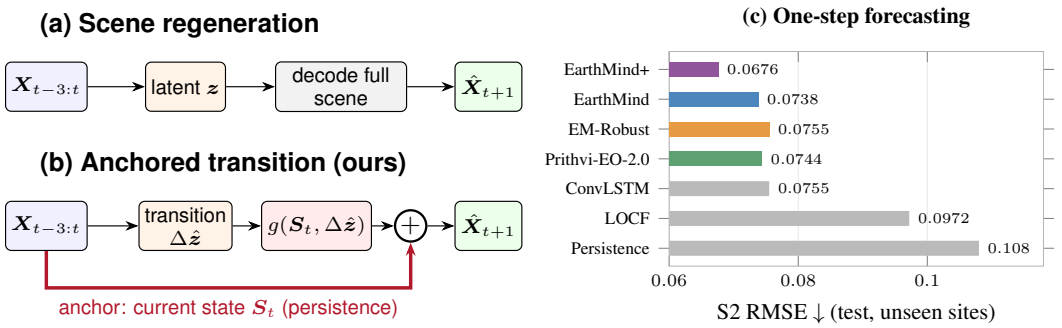

Figure 1: **Core idea and headline result.** (a) Conventional forecasters decode the next scene from a compressed code and must re-learn what is already visible. (b) EarthMind anchors on the current fused state and predicts only the transition. (c) Sentinel-2 RMSE on the held-out, site-disjoint test set (Table 1). Single EarthMind has comparable observed RMSE to pretrained Prithvi-EO-2.0 (Szwarcman et al., 2025), while calibrated EarthMind+ achieves the lowest RMSE among the evaluated configurations.

source of incomplete inputs. A zero-filled array must therefore be distinguished from a valid low-valued measurement (Ma et al., 2021; 2022).

We address both properties with EarthMind, a multimodal spatiotemporal world model, drawing on the broader idea of learning latent states and their transitions (Ha & Schmidhuber, 2018; Hafner et al., 2025). EarthMind *anchors* its forecast on the current state and *learns the transition*. Each month's Sentinel-2, Sentinel-1, terrain, and weather inputs are fused into a spatial Earth-state map and pooled into a latent state. A Transformer (Vaswani et al., 2017) predicts the next latent state, and the predicted latent transition $\hat{z}_{t+1} - \tilde{z}_t$ conditions a residual update of the most recent spatial map. The updated spatial representation is decoded into future Sentinel-2 and Sentinel-1 observations, while a separate head predicts future weather. The same anchor-plus-transition principle reappears at the output: EarthMind+ averages forecasts from three independently trained models, following the general deep-ensemble approach (Lakshminarayanan et al., 2017), and then calibrates, separately for each spectral band, how far the ensemble forecast moves away from persistence.

Contributions.

- **Persistence-anchored latent transitions.** A forecasting formulation in which a global latent transition, predicted jointly with a latent-consistency target, modulates a residual update of the last fused spatial state (Section 3.3). In the controlled, single-seed ablations, removing the temporal transition model increases RMSE by 6.9%, while removing weather increases RMSE by 15.4%.
- **Availability-aware state estimation.** Explicit availability masks with learned modality-specific missing embeddings, together with structured historical-input masking informed by missing-modality learning (Ma et al., 2021; 2022) (Section 3.2). Under joint SAR and weather loss, the robust training configuration reduces RMSE by approximately 35% relative to standard training and retains a latent R^2 of 0.588, compared with 0.081 for the standard configuration.
- **Band-wise transition calibration.** A validation-fitted shrinkage of the ensemble forecast transition toward persistence, separately for each spectral band, improves all five reported forecasting metrics relative to the standard EarthMind model without retraining (Section 3.4). Independently trained models are combined using the general deep-ensemble approach (Lakshminarayanan et al., 2017); the band-wise residual calibration is specific to EarthMind+.
- **A deployment-oriented evaluation protocol.** Site-disjoint splits under natural modality availability, a missing-modality stress suite, closed-loop rollouts, a representation probe, and paired site-level significance tests with multiple-comparison correction (Holm, 1979). The evaluation includes a pretrained Prithvi-EO-2.0 reference (Szwarcman et al., 2025) and reports metrics on which EarthMind does not obtain the lowest error.

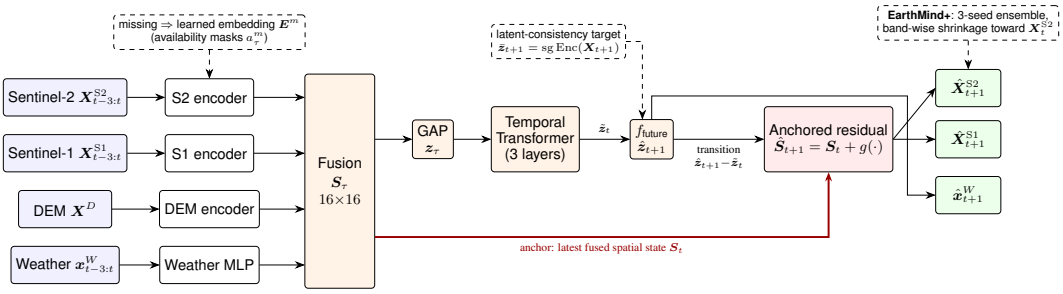

Figure 2: **EarthMind.** Four months of Sentinel-2, Sentinel-1, terrain, and weather are encoded per modality, with missing inputs replaced by learned embeddings, and fused into spatial Earth states $\boldsymbol{S}_\tau$. A Transformer over pooled states predicts the next latent state, supervised against a stop-gradient encoding of the true next observation. The predicted transition conditions a residual update of the most recent spatial state (red anchor path), which is decoded into future Sentinel-2 and Sentinel-1; weather is decoded from $\hat{\boldsymbol{z}}_{t+1}$. EarthMind+ applies band-wise shrinkage toward persistence.

2 Related Work

EO and spatiotemporal forecasting. Spatiotemporal predictive-learning architectures, including recurrent models such as ConvLSTM and PredRNN (Wang et al., 2017), SimVP (Gao et al., 2022a), and Earthformer (Gao et al., 2022b), forecast gridded fields directly in pixel or feature space and have been benchmarked systematically in OpenSTL (Tan et al., 2023). For Earth surface imagery, GreenEarthNet framed Sentinel-2 vegetation forecasting as a multimodal, meteorology-guided prediction task (Benson et al., 2024). Weather and climate foundation models such as ClimaX (Nguyen et al., 2023) and Aurora (Bodnar et al., 2025) forecast reanalysis variables rather than high-resolution surface imagery. EarthMind differs in two ways: its state is fused from optical, SAR, terrain, and weather before temporal modeling, and it forecasts a transition applied to the current spatial state rather than generating the future scene.

Remote sensing foundation models. Self-supervised EO encoders include SatMAE (Cong et al., 2022), SSL4EO-S12 (Wang et al., 2023b), CROMA (Fuller et al., 2023), Presto (Tseng et al., 2023), SkySense (Guo et al., 2024), DOFA (Xiong et al., 2024), Galileo (Tseng et al., 2025), and Prithvi (Jakubik et al., 2023; Szwarcman et al., 2025). They are evaluated mainly on classification, segmentation, and regression rather than forecasting future observations. We use Prithvi-EO-2.0 as a strong pretrained reference, not an architecture-matched comparison.

Multimodal EO and missing modalities. SAR has been used to compensate for cloud-contaminated optical time series (Ebel et al., 2022; 2023). General approaches to incomplete inputs include robustness analyses of multimodal Transformers (Ma et al., 2022), missing-aware prompting (Lee et al., 2023), and shared–specific feature modeling (Wang et al., 2023a). EarthMind instead samples missingness from a small set of structured outage patterns observed in EO practice and distinguishes absent from low-valued inputs through learned embeddings.

World models and latent prediction. Latent world models learn a compact state and its transition dynamics (Hafner et al., 2025), and joint-embedding predictive architectures supervise prediction in representation space (Assran et al., 2023; Bardes et al., 2024). EarthMind combines a latent-consistency target of this kind with pixel-level decoding, and uses the latent prediction as a transition signal for a spatial residual. We use "world model" in this limited sense and make no causal or interventional claim.

3 EARTHMIND

3.1 PROBLEM FORMULATION

At month t, the model receives Sentinel-2 imagery $\boldsymbol{X}_\tau^{\mathrm{S2}}$ (bands B2, B3, B4, B8, B11, B12), Sentinel-1 imagery $\boldsymbol{X}_\tau^{\mathrm{S1}}$ (VV, VH), a static DEM $\boldsymbol{X}^D$, and a ten-dimensional weather vector $\boldsymbol{x}_\tau^W$ for $\tau \in \{t-3, \ldots, t\}$, together with binary availability indicators a_τ^m for $m \in \{\mathrm{S2}, \mathrm{S1}, W\}$. The target is the next-month Sentinel-2, Sentinel-1, and weather observation. Availability is not assumed: at both training and test time, any historical modality may be absent.

3.2 AVAILABILITY-AWARE MULTIMODAL STATE

Sentinel-2, Sentinel-1, and DEM are encoded by separate convolutional encoders E_m, each with four stride-2 blocks (channels 64/128/192/256, group normalization, GELU), mapping a 256×256 input to $\mathbb{R}^{256\times16\times16}$. Weather is projected by a two-layer MLP ($10\to128\to256$, GELU, layer normalization). A zero-filled array is not a valid observation, so absent inputs are replaced *after* encoding by a learned embedding:

$$\tilde{\boldsymbol{H}}_\tau^m = a_\tau^m \, E_m(\boldsymbol{X}_\tau^m) + (1 - a_\tau^m) \, \boldsymbol{E}^m, \tag{1}$$

where $\boldsymbol{E}^m$ is a learned spatial map for Sentinel-2 and Sentinel-1 and a learned vector for weather. Terrain is always available and encoded once per sequence. Weather is broadcast over the 16×16 grid, the four modality features are concatenated, and a 1×1 then 3×3 convolution (group normalization, GELU) yields the fused spatial state $\boldsymbol{S}_\tau \in \mathbb{R}^{256\times16\times16}$. Global average pooling gives the latent state $\boldsymbol{z}_\tau = \mathrm{GAP}(\boldsymbol{S}_\tau) \in \mathbb{R}^{256}$.

Structured modality masking (EarthMind-Robust). Random per-modality dropout does not reflect how EO sensors fail. EarthMind-Robust uses the same architecture but samples each training sequence's historical availability from six outage patterns: full input, missing S1, missing S2, missing weather, missing S1+weather, and missing S2+weather, with probabilities 0.60, 0.10, 0.10, 0.10, 0.05, 0.05. Only historical inputs are masked; targets and terrain are unchanged.

3.3 PERSISTENCE-ANCHORED LATENT TRANSITIONS

The four latent states, with learned positional embeddings, are processed by a three-layer pre-norm Transformer encoder (eight heads, feed-forward width 4×256, GELU). Its output at the latest position, $\tilde{\boldsymbol{z}}_t$, is mapped to a next-month latent estimate $\hat{\boldsymbol{z}}_{t+1} = f_{\text{future}}(\tilde{\boldsymbol{z}}_t)$.

A global vector cannot carry the spatial layout needed for 256×256 forecasts, and decoding a future scene from scratch discards what is already known. EarthMind therefore treats the latest fused map as an anchor and predicts only a spatial correction, conditioned on the predicted state and the predicted latent transition:

$$\hat{\boldsymbol{S}}_{t+1} = \boldsymbol{S}_t + g(\boldsymbol{S}_t, \; \hat{\boldsymbol{z}}_{t+1}, \; \hat{\boldsymbol{z}}_{t+1} - \tilde{\boldsymbol{z}}_t), \tag{2}$$

where the two vectors are broadcast over the grid, concatenated with $\boldsymbol{S}_t$, and passed through a convolutional projection g. When the predicted transition is small, the forecast stays close to the current state; when it is large, g can rewrite local structure. Four transposed-convolution stages and a final convolution with sigmoid decode $\hat{\boldsymbol{S}}_{t+1}$ into six-band Sentinel-2 and two-band Sentinel-1 at 256×256; a layer-normalized MLP decodes future weather from $\hat{\boldsymbol{z}}_{t+1}$.

Latent-consistency training. In addition to pixel losses, the predicted state is supervised in representation space against $\bar{\boldsymbol{z}}_{t+1}$, the encoding of the true next observation by the same encoders under its own availability masks, with gradients stopped through the target (Assran et al., 2023). This encourages $\hat{\boldsymbol{z}}_{t+1} - \tilde{\boldsymbol{z}}_t$ to represent a genuine state transition rather than an arbitrary decoder input. The full objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{S2}} + 0.75\,\mathcal{L}_{\mathrm{S1}} + 0.5\,\mathcal{L}_W + 0.5\,\mathcal{L}_{\text{lat}} + 0.1\,\mathcal{L}_{\cos} + 0.2\,(\mathcal{L}_{\mathrm{S2}}^{\text{rec}} + \mathcal{L}_{\mathrm{S1}}^{\text{rec}}), \tag{3}$$

where $\mathcal{L}_{\mathrm{S2}}, \mathcal{L}_{\mathrm{S1}}$ are masked L_1 losses (unavailable targets contribute nothing), $\mathcal{L}_W$ is a masked Smooth-L_1 loss, $\mathcal{L}_{\text{lat}}$ and $\mathcal{L}_{\cos}$ are the mean-squared and cosine distances between $\hat{\boldsymbol{z}}_{t+1}$ and $\bar{\boldsymbol{z}}_{t+1}$, and the auxiliary L_1 terms reconstruct the current Sentinel-2 and Sentinel-1 inputs.

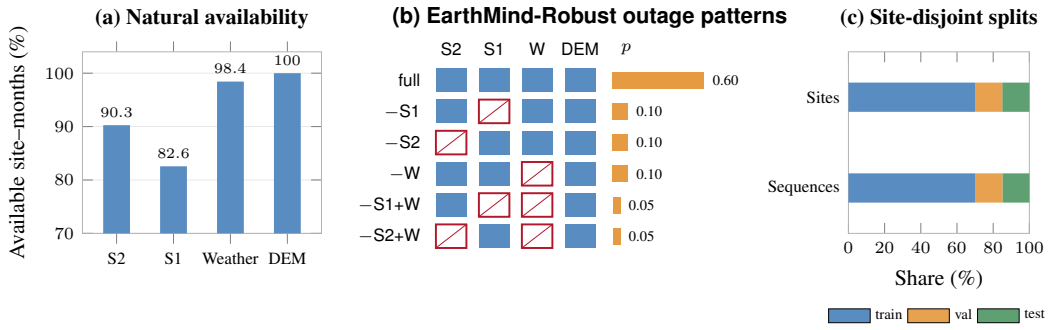

Figure 3: **Data and missingness.** (a) Modality availability across all 128,000 site–months. (b) The six historical outage patterns sampled by EarthMind-Robust (filled: available; crossed: removed) with their probabilities p. (c) Train/validation/test partition by site (700/150/150) and the resulting sequences (77,756/16,773/16,562).

3.4 Band-wise transition calibration (EarthMind+)

EarthMind+ is an inference procedure, not a new network. Sentinel-2 forecasts from three independently trained seeds (42, 123, 2026) are averaged into $\hat{\boldsymbol{X}}^{\text{ens}}_{t+1}$, and the forecast transition is shrunk toward persistence separately for each band b:

$$\hat{\boldsymbol{X}}^{+}_{t+1,b} = \boldsymbol{X}^{\text{S2}}_{t,b} + \alpha_b\big(\hat{\boldsymbol{X}}^{\text{ens}}_{t+1,b} - \boldsymbol{X}^{\text{S2}}_{t,b}\big). \tag{4}$$

The α_b are chosen to minimize RMSE on a fixed subset of 4,096 validation sequences and frozen before any test evaluation. Equation 4 extends the anchoring principle of Equation 2 from latent space to observation space. $\alpha_b < 1$ indicates that the ensemble's predicted change is, on average, larger than the change that minimizes squared error for that band, so α_b acts as an empirical, band-specific measure of how much one-month change the model can be trusted to predict.

4 Experimental Setup

Dataset. We constructed a globally distributed dataset of 1,000 fixed land and coastal sites (open ocean excluded; minimum separation 10 km). Each site is a 256×256 grid at 10 m ($\approx$2.56 km$\times$2.56 km) with 128 monthly slots from January 2016. A sample uses four consecutive months as input and the fifth as target; non-consecutive windows are rejected. Sentinel-2 monthly composites are produced from harmonized imagery with cloud filtering and median compositing and scaled to approximately $[0, 1]$; a patch with fewer than 5% valid pixels is marked unavailable. Sentinel-1 VV/VH is aggregated monthly in linear power, converted to dB, clipped to $[-30, 5]$, and normalized to $[0, 1]$. Weather comprises 2 m temperature, 2 m dew point, precipitation, surface pressure, 10 m u/v wind, wind speed, soil moisture, snow depth, and solar radiation, standardized with training-site statistics. Availability is 90.26% (Sentinel-2), 82.56% (Sentinel-1), 98.43% (weather), and 100% (terrain).

Site-disjoint splits. Sites, not samples, are partitioned into 700/150/150 train/validation/test, giving 77,756/16,773/16,562 sequences. All windows from a site stay in one split, so test performance measures forecasting at unseen locations under their natural availability.

Training. Latent width 256, context length 4, Transformer dropout 0.1. We use AdamW with learning rate 10^{-4} and weight decay 10^{-4}, physical batch size 2 with four-step gradient accumulation (effective 8), and mixed precision. Multi-seed experiments use seeds 42, 123, and 2026.

Baselines and metrics. Persistence copies month t's Sentinel-2 input as the forecast; last-observation-carried-forward (LOCF) uses the most recent available Sentinel-2 month in the context. We also compare with a ConvLSTM forecaster (see Tan et al., 2023) and a Prithvi-EO-2.0-100M-TL-based optical forecaster (Szwarcman et al., 2025). Because Prithvi differs from EarthMind in

Table 1: One-step Sentinel-2 forecasting on the site-disjoint test set (mean ± std over three seeds where applicable). EarthMind+ is a single calibrated three-seed ensemble. Relative RMSE is with respect to LOCF.

Method	MAE ↓	RMSE ↓	PSNR ↑	SSIM ↑	SAM (°) ↓	ΔRMSE
Persistence	0.0450	0.1080	19.33	0.9117	5.727	+11.1%
LOCF	0.0393	0.0972	20.25	0.9515	5.906	–
ConvLSTM	0.0357±.0016	0.0755±.0058	22.46±.67	0.9612±.0031	5.919±.272	−22.3%
Prithvi-EO-2.0 (pretrained)	**0.0303**±.0001	0.0744±.0005	22.57±.06	0.9687±.0001	4.941±.003	−23.5%
EarthMind-Robust	0.0354±.0002	0.0755±.0002	22.44±.02	0.9631±.0002	5.583±.044	−22.3%
EarthMind	0.0339±.0004	0.0738±.0008	22.64±.09	0.9664±.0006	5.100±.096	−24.1%
EarthMind+	0.0313	**0.0676**	**23.40**	**0.9704**	**4.681**	−30.5%

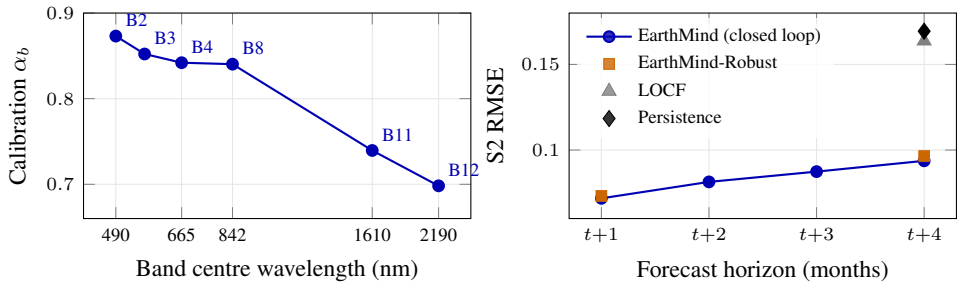

Figure 4: **Left:** validation-selected calibration coefficients (Eq. 4) fall with wavelength; the SWIR bands (B11, B12) retain the smallest fraction of the predicted change. **Right:** closed-loop rollout RMSE. Rolled forward from a one-step model, EarthMind at $t{+}4$ remains 43–45% below LOCF and persistence at the same horizon. All values are taken from Sections 3.4 and 5.

pretraining data, modality access, and inference design, it serves as a strong pretrained reference rather than a controlled ablation. We report MAE, RMSE, PSNR $= 10\log_{10}(1/\text{MSE})$ over normalized values, SSIM, and the spectral angle mapper (SAM).

5 Results

5.1 One-step forecasting

Table 1 reports clean one-step Sentinel-2 forecasting. Three observations stand out. (i) **A from-scratch multimodal model obtains comparable observed performance to a pretrained optical foundation model.** Single EarthMind has lower RMSE and higher PSNR than Prithvi-EO-2.0 (Szwarcman et al., 2025), while Prithvi has lower MAE, higher SSIM, and lower SAM. A paired site-level analysis over the 149 usable test sites, using a 10,000-resample bootstrap and paired t-tests with Holm multiple-comparison correction (Holm, 1979) across the five metrics, finds no significant difference on any metric after correction. We therefore interpret these results as comparable rather than evidence of statistical superiority. EarthMind additionally forecasts Sentinel-1 and weather and accepts incomplete historical inputs, whereas Prithvi-EO-2.0 is used as a pretrained optical reference rather than an architecture-matched baseline (Szwarcman et al., 2025).

(ii) **Calibration toward persistence helps on every reported metric.** EarthMind+ improves over single EarthMind on MAE (0.0339 → 0.0313), RMSE (−8.4%), PSNR (+0.75 dB), SSIM, and SAM (5.10° → 4.68°), obtaining the lowest RMSE, highest PSNR, highest SSIM, and lowest SAM among the evaluated methods; Prithvi retains the lowest MAE. EarthMind+ combines forecasts from independently trained models, following the general deep-ensemble setting (Lakshminarayanan et al., 2017), before applying the proposed band-wise calibration. Because EarthMind+ combines ensembling and calibration, its gain over Prithvi should not be attributed to the architecture alone.

(iii) **Naive persistence is weaker than LOCF.** Persistence obtains an RMSE of 0.1080 compared with 0.0972 for LOCF. Because LOCF uses the most recent available Sentinel-2 observation rather than assuming that the final context month contains a valid optical observation, the result shows that handling missing Sentinel-2 months matters when constructing a temporal baseline.

What the calibration coefficients reveal. The validation-selected coefficients are $\alpha = (0.873, 0.852, 0.842, 0.840, 0.740, 0.698)$ for B2, B3, B4, B8, B11, and B12. They decrease monotonically with wavelength in this experiment (Figure 4, left): visible and NIR bands retain 84–87% of the predicted change, while the two shortwave-infrared bands retain only 70–74%. B11 and B12 correspond to Sentinel-2's shortwave-infrared bands, which are sensitive to moisture-related surface and vegetation properties. The smaller validation-selected coefficients therefore show that, for our model and validation data, stronger shrinkage toward persistence minimizes squared error in the SWIR bands. This suggests that band-specific uncertainty or calibration, rather than a single global confidence parameter, is a useful direction for future EO forecasters.

5.2 Component ablations

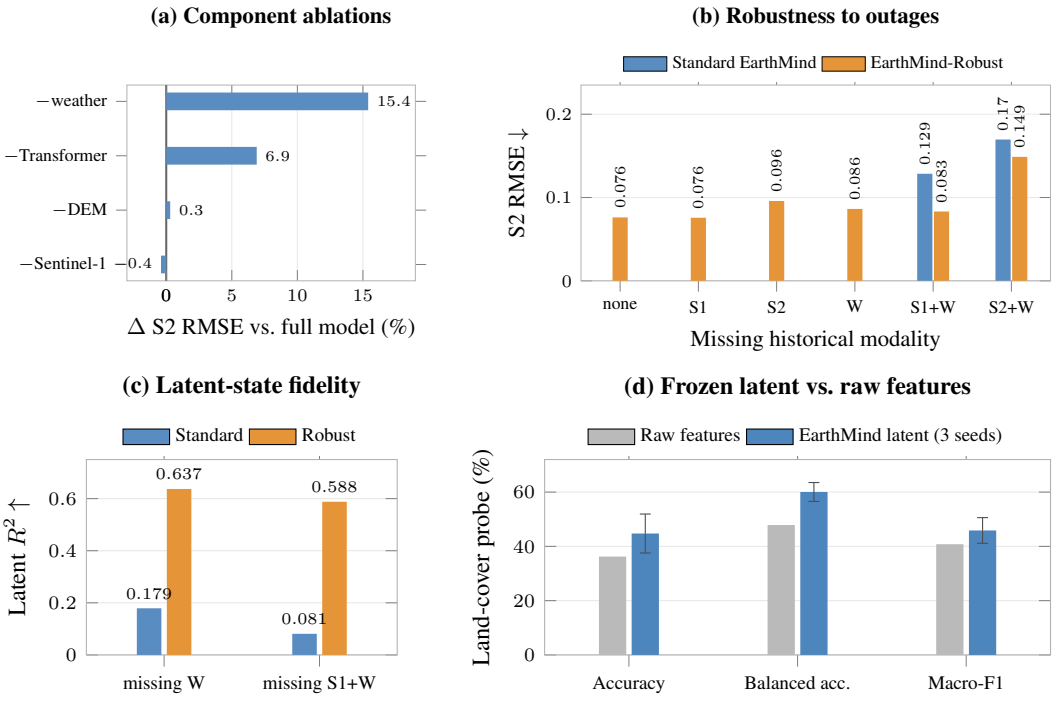

Figure 5: **What each component contributes and how the model degrades.** (a) Single-seed (seed 42) ablations; weather and the temporal transition model matter most. (b) Historical-input outages; structured masking cuts RMSE by 35% when SAR and weather are both missing. Standard-model values were evaluated for the two compound outages only. (c) Latent R^2 of the predicted state under outages. (d) Ten-class site-level land-cover linear probe on frozen 256-d states (mean $\pm$ std over three seeds) versus raw input features.

Figure 5a shows that removing weather produces the largest RMSE degradation among the tested components, increasing RMSE by 15.4%. This result is consistent with prior Earth-surface forecasting approaches that incorporate meteorological information (Requena-Mesa et al., 2021; Benson et al., 2024) and with weather and Earth-system forecasting models that explicitly model atmospheric variables (Nguyen et al., 2023; Bodnar et al., 2025). Removing the temporal transition model increases RMSE by 6.9% while also reducing SSIM and increasing SAM, supporting the contribution of explicit temporal modeling. Transformer-based temporal modeling has similarly been explored for Earth-system forecasting (Gao et al., 2022b). Removing the DEM has negligible effect in this single-seed benchmark. Removing Sentinel-1 slightly lowers RMSE but worsens

MAE, SSIM, and SAM; its contribution under complete inputs is therefore mixed. The complementary role of radar and optical information under incomplete or cloud-contaminated observations has been explored in multimodal EO (Ebel et al., 2022; 2023). In our outage experiment, missing S1 alone has little effect on EarthMind-Robust, whereas missing S2 produces a larger degradation (Figure 5b). These are single-seed configuration comparisons, not causal estimates.

5.3 Robustness to missing modalities

Figure 5b–c evaluates historical-input outages. Missing modalities are a recognized challenge in multimodal learning (Ma et al., 2021; 2022; Lee et al., 2023; Wang et al., 2023a). Structured masking matters most in our experiments when outages compound: with SAR and weather both missing, EarthMind-Robust reduces RMSE by 35.3% relative to standard training ($0.1285 \rightarrow 0.0832$); with optical and weather missing, RMSE is reduced by 12.3%.

The latent state also degrades substantially less under the tested missing-input conditions. Under missing weather, EarthMind-Robust retains a latent R^2 of 0.637 versus 0.179 for the standard model, and under missing S1+weather it retains 0.588 versus 0.081. These results indicate that structured missing-modality training makes the predicted latent state more robust to the tested outages. The cost is approximately 2.3% higher clean-input RMSE relative to standard EarthMind (Table 1). Absolute values in this subsection and the rollout experiment come from separate evaluation runs and differ slightly from Table 1; comparisons are made within each experiment.

5.4 Closed-loop rollouts

Autoregressive and recurrent spatiotemporal forecasting can accumulate prediction errors when model outputs are recursively fed back as inputs (Wang et al., 2017; Gao et al., 2022b; Tan et al., 2023). In our closed-loop evaluation, EarthMind's RMSE grows from 0.0718 at $t{+}1$ to 0.0814, 0.0874, and 0.0937 at $t{+}2$, $t{+}3$, and $t{+}4$, respectively, corresponding to a 30.5% increase across the evaluated horizon. EarthMind-Robust grows from 0.0732 at $t{+}1$ to 0.0965 at $t{+}4$.

Although trained only for one-step forecasting, EarthMind at $t{+}4$ remains 42.8% below LOCF (0.1637) and 44.7% below persistence (0.1694) in RMSE at the same horizon. Error increases monotonically with forecast horizon, showing that multi-step forecasting remains more difficult than one-step prediction, while the learned variants remain below LOCF and persistence at every evaluated horizon (Figure 4, right).

5.5 Is the latent state meaningful: A land-cover probe

Frozen-representation evaluation provides a lightweight way to examine whether learned features contain transferable semantic information and is commonly used in self-supervised representation learning, including EO representation models (Cong et al., 2022; Wang et al., 2023b; Fuller et al., 2023). We freeze the 256-d latent state, average it over three temporally distributed sequences per site, and train a ten-class site-level land-cover linear probe using 687/149/149 train/validation/test sites.

Across three seeds, EarthMind latent representations reach $44.7 \pm 7.2\%$ accuracy, $60.0 \pm 3.5\%$ balanced accuracy, and $45.9 \pm 4.7\%$ macro-F1, compared with 36.2%, 47.9%, and 40.8%, respectively, for raw input features. The observed improvements are +8.5, +12.2, and +5.1 percentage points. The forecasting-trained representation therefore contains land-cover-discriminative information despite receiving no land-cover supervision during forecasting training. We treat this experiment as a lightweight representation diagnostic rather than as a comprehensive downstream land-cover benchmark (Figure 5d).

6 Limitations

(1) The objective is regression-based, and Sentinel-1 forecasts are visibly smoother than their targets; high-frequency SAR speckle is not modeled. (2) EarthMind+'s gains combine ensembling and calibration; we did not apply the same calibration to Prithvi, so the EarthMind+ vs. Prithvi gap should not be read as an architectural comparison. (3) Ablations use a single seed, and standard-

model numbers are not reported for every outage pattern. (4) The benchmark covers 1,000 sites and does not establish generalization to every climate, land-cover type, or extreme event. (5) "World model" refers to a learned latent state and transition; we do not validate counterfactual or causal behavior.

7 Conclusion

EarthMind shows that monthly Earth-observation forecasting is better posed as learning a transition from an anchored multimodal state than as regenerating the scene. Trained from scratch, it performs comparably to a pretrained optical foundation model on clean inputs while also forecasting SAR and weather from incomplete histories. Structured modality masking makes the model substantially more robust when sensors fail together: it reduces RMSE by 35.3% when SAR and weather are both missing and preserves a far more informative latent state (latent R^2 of 0.588 versus 0.081). Band-wise calibration toward persistence improves every reported metric. It also shows that the shortwave-infrared bands retain only 70–74% of the predicted change, compared with 84–87% for the visible and NIR bands. Beyond one-step accuracy, the model remains more than 40% below persistence baselines at a four-month horizon, and its latent state encodes land-cover information it was never trained on. Taken together, these results support persistence-anchored latent transitions, combined with site-disjoint, availability-realistic evaluation, as a practical foundation for future Earth-observation world models.

Reproducibility Statement

Architecture, loss weights, masking probabilities, calibration coefficients and procedure, data splits, and optimization settings are specified in Sections 3–4. Code, the site list, preprocessing scripts, and trained checkpoints will be released upon acceptance.

Ethics Statement

This work uses publicly available Copernicus Sentinel-1 and Sentinel-2 imagery, gridded meteorological variables, and a public DEM. No personal data are involved. Forecasts are research outputs and are not intended for operational decisions without further validation.

LLM Usage Disclosure

Large language models were used solely to check this manuscript for grammatical errors, formatting inconsistencies, and unintended whitespace. They were not used to write or generate the manuscript's content.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 2025.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, 2022.

Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

Patrick Ebel, Vivien Sainte Fare Garnot, Michael Schmitt, Jan Dirk Wegner, and Xiao Xiang Zhu. UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

Anthony Fuller, Koreen Millard, and James Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Advances in Neural Information Processing Systems*, 2023.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

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. In *Advances in Neural Information Processing Systems*, 2022b.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 2025.

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

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *AAAI Conference on Artificial Intelligence*, 2021.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 2021.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *International Conference on Machine Learning*, 2023.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique De Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-EO-2.0: A versatile multitemporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

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 *Advances in Neural Information Processing Systems*, 2023.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R. Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In *International Conference on Machine Learning*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 2023b.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems*, 2017.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.