Multi-modal learning for geospatial vegetation forecasting

Vitus Benson[1,2,3,*] Claire Robin[1,2] Christian Requena-Mesa[1,2] Lazaro Alonso[1]
Nuno Carvalhais[1,2] José Cortés[1] Zhihan Gao[4] Nora Linscheid[1] Mélanie Weynants[1]
Markus Reichstein[1,2]
[1] Max-Planck-Institute for Biogeochemistry [2] ELLIS Unit Jena [3] ETH Zürich
[4] Hong Kong University of Science and Technology [*] vbenson@bgc-jena.mpg.de

Abstract

Precise geospatial vegetation forecasting holds potential across diverse sectors, including agriculture, forestry, humanitarian aid, and carbon accounting. To leverage the vast availability of satellite imagery for this task, various works have applied deep neural networks for predicting multispectral images in photorealistic quality. However, the important area of vegetation dynamics has not been thoroughly explored. Our study introduces GreenEarthNet, the first dataset specifically designed for high-resolution vegetation forecasting, and Contextformer, a novel deep learning approach for predicting vegetation greenness from Sentinel 2 satellite images with fine resolution across Europe. Our multi-modal transformer model Contextformer leverages spatial context through a vision backbone and predicts the temporal dynamics on local context patches incorporating meteorological time series in a parameter-efficient manner. The GreenEarthNet dataset features a learned cloud mask and an appropriate evaluation scheme for vegetation modeling. It also maintains compatibility with the existing satellite imagery forecasting dataset EarthNet2021, enabling cross-dataset model comparisons. Our extensive qualitative and quantitative analyses reveal that our methods outperform a broad range of baseline techniques. This includes surpassing previous state-of-the-art models on EarthNet2021, as well as adapted models from time series forecasting and video prediction. To the best of our knowledge, this work presents the first models for continental-scale vegetation modeling at fine resolution able to capture anomalies beyond the seasonal cycle, thereby paving the way for predicting vegetation health and behaviour in response to climate variability and extremes. We provide open source code and pre-trained weights to reproduce our experimental results under `https://github.com/vitusbenson/greenearthnet` *[10].*

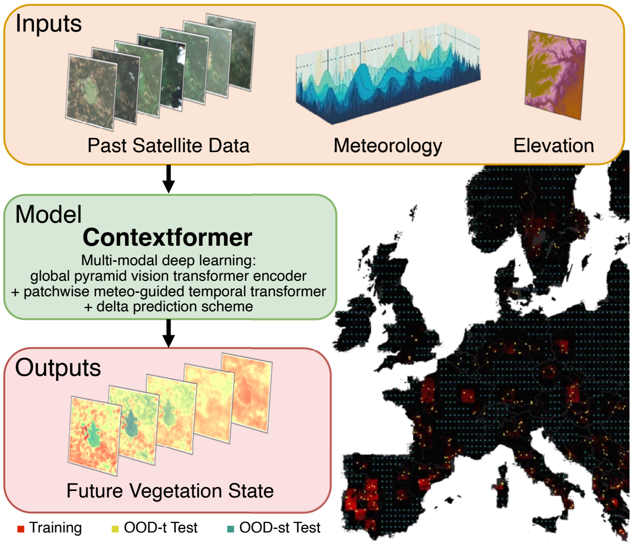
Inputs
Past Satellite Data
Meteorology
Elevation
Model
Contextformer
Multi-modal deep learning:
global pyramid vision transformer encoder
+ patchwise meteo-guided temporal transformer
+ delta prediction scheme
Outputs
Future Vegetation State
🟥 Training
🟨 OOD-t Test
🟩 OOD-st Test

Figure 1. GreenEarthNet approach (map shows sample locations).

1. Introduction

Optical satellite images have been proven useful for monitoring vegetation status. This is essential for a variety of applications in agricultural planning, forestry advisory, humanitarian assistance or carbon monitoring. In all these cases, prognostic information is relevant: Farmers want to know how their farmland may react to a given weather scenario [83]. Humanitarian organisations need to understand the localized impact of droughts on pastoral communities for mitigation of famine with anticipatory action [49]. Afforestation efforts need to consider how their forests react to future climate [71]. However, providing such prognostic information through fine resolution vegetation forecasts is challenging due to ecological memory effects [35], spatial interactions and the influence of weather variations. Deep neural networks can model relationships in space, time or across modalities. Hence, their application to vegetation forecasting given a sufficiently large dataset seems natural.

So far, deep learning for vegetation forecasting can be roughly grouped into two categories. 1) global forecasting, using low-resolution data from satellites like AVHRR and MODIS [8, 30, 35, 37, 46, 65, 92], which overlooks pixel-level heterogeneity. 2) Local forecasting, utilizing high-resolution imagery from satellites like Sentinel 2 and Landsat [17, 33, 60, 61, 68, 75], which aims to capture field-scale heterogeneity: a grassland will react different to a forest, two neighbouring fields can have almost opposite dynamics depending on the type of crops (*e.g.* winter wheat vs. maize), and vegetation on a north-facing slope close to a river is more resilient to drought stress than on a rocky south-facing slope. However, aforementioned local forecasting approaches have focused mainly on perceptual image quality instead of vegetation dynamics, which renders their suitability for vegetation forecasting uncertain.

In this study, we tackle continental-scale vegetation forecasting by predicting vegetation greenness at 20m resolution based on coarse-scale weather data (Fig. 1). For this, we introduce the *GreenEarthNet* dataset, comprising Sentinel 2 satellite image time series and a high quality deep learning-based cloud mask, which allows to distinguish between anomalies due to data corruption and those due to meteorological and anthropogenic influence. GreenEarthNet maintains consistency with the EarthNet2021 dataset in training locations and spatiotemporal dimensions, facilitating the reuse of leading models such as ConvLSTM [17], SGED-ConvLSTM [33] and Earthformer [22] for vegetation forecasting. Essentially, GreenEarthNet represents an enhanced version of EarthNet2021, addressing its shortcomings and enabling multi-modal learning for geospatial vegetation forecasting. To advance state-of-the-art on the new dataset, we introduce the *Contextformer*, a lightweight transformer model incorporating spatial interactions through a Pyramid Vision Transformer [78, 79] and temporal dynamics modeling using a temporal transformer encoder. Additionally, it incorporates a delta-prediction scheme to prioritize persistence from an initial observation.

Our major **contributions** can be summarized as follows.

(1) We present the GreenEarthNet dataset, the first large-scale dataset suitable for prediction of within-year vegetation dynamics, including a learned cloud mask and a new evaluation scheme.

(2) We introduce the Contextformer, a novel multi-modal transformer model suitable for vegetation forecasting, leveraging spatial context through its vision backbone, and forecasting the temporal evolution of small context patches with a temporal transformer.

(3) We compare the Contextformer against a previously unseen variety of state-of-the-art models from related tasks and find it outperforms all of them across metrics.

2. Related Work

Vegetation forecasting There is a growing interest in vegetation growth forecasting driven by the democratization of machine learning techniques, the availability of remote sensing data, and the urgency to address climate change [15, 21, 31, 38]. Numerous studies in vegetation modeling use coarse resolution data from satellites like AVHRR or MODIS [30, 35, 37, 46, 92]. Since 2015, Sentinel-2 has provided high-resolution satellite imagery (up to 10m), enabling more localized modeling. The introduction of EarthNet2021 [60] marked the first dataset for self-supervised Earth surface forecasting, which contains predicting satellite imagery and derived vegetation state with a focus on perceptual quality. Subsequently, the ConvLSTM model [66] has been widely used for satellite imagery prediction [1, 17, 33, 45, 61], hence we are including it as a baseline.

Spatio-temporal learning Learning spatio-temporal dynamics (as in the case of vegetation forecasting) is a challenge across many disciplines. Often, temporal dynamics dominate, so local time series models can be effective. For instance in traffic, weather or electricity forecasting, time series models such as LSTM [28], Prophet [74], Autoformer [84] or NBeats [91] yield useful performance. Still, often spatial interactions are important or at least offer additional predictive capacity. For instance in video prediction, ConvNets [6, 23], ConvLSTM [66], ConvLSTM successors [81, 85], PredRNN [80, 81], SimVP [73] and transformers [22, 24, 50] have been found skillful. Often, the necessity of modeling the spatial component translates to Earth science: spatio-temporal deep learning is being applied for precipitation nowcasting [57, 67], weather forecasting [11, 36, 53], climate projection [52], and wildfire modeling [34]. Hence, when evaluating our Contextformer model, we need to do so against strong baselines from video prediction [22, 66, 73, 81], as a priori one might expect them to outperform also on vegetation forecasting. However, vegetation forecasting does present some unique challenges, it builds upon multi-modal data fusion and requires capturing across-scale relationships (in time and space), which may prove challenging for existing video prediction models and thus interesting to the computer vision community.

Multi-modal transformers for data fusion Levering remote sensing data often means multi-modal data fusion. Recently, machine learning methods have shown significant advancements in fusing different satellite sensors compared to traditional approaches [3, 16, 39, 69, 82]. This includes recent work on combining Sentinel 2 and SAR data to impute cloudy Sentinel 2 images [48, 77, 88]. Gapfilling vegetation time series could also be done with the models presented in this study, as they leverage meteorology to inform the imputation [70]. However, as gapfilling is just done in retrospective, one should rather resort to complementary satellite data like SAR.

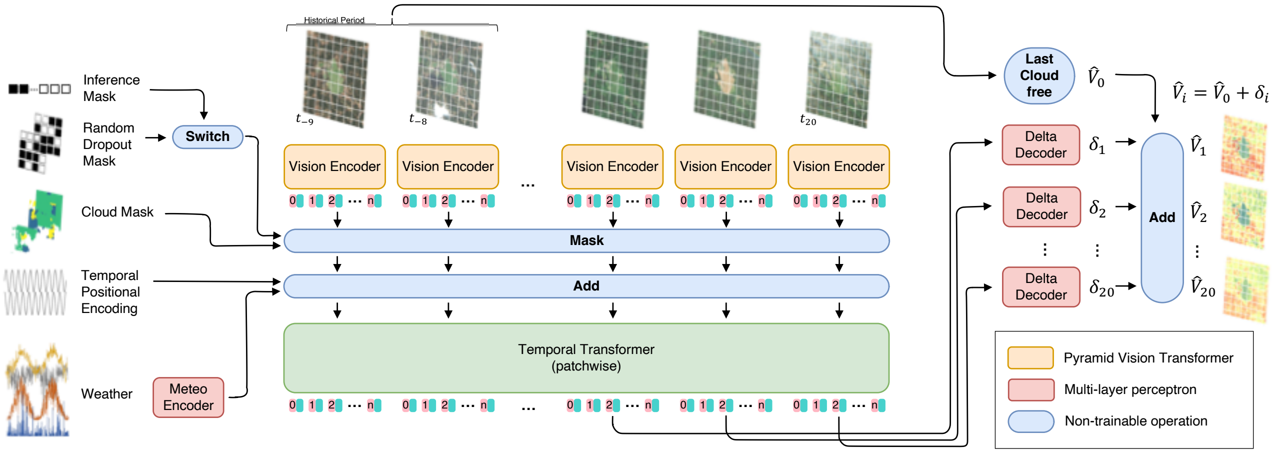

Figure 2. The architecture of our proposed Contextformer.

Transformers [76] offer a compelling approach to handle multi-modal data [29]. Their efficacy in remote sensing has been shown multiple times [12, 22, 44, 75]. In particular, geospatial foundation models [12, 47, 58, 68, 75, 89] make use of this approach, often through masked token modeling [25] with Vision Transformers (ViT, [19]). Our Contextformer differs from geospatial foundation models by focusing specifically on vegetation forecasting and only utilizing a pre-trained vision transformer as vision backbone.

3. Methods

3.1. Task

We predict the future NDVI, a remote sensing proxy of vegetation state ($V^t \in \mathbb{R}^{H\times W}, t \in [T+1, T+K]$) conditioned on past satellite imagery ($X^t \in \mathbb{R}^{H\times W}, t \in [1, T]$), past and future weather ($C^t \in \mathbb{R}, t \in [1, T+K]$) and static elevation maps ($E \in \mathbb{R}^{H\times W}$). Hence, denoting a model $f(.;\theta)$ with parameters θ, we obtain vegetation forecasts as:

$$\hat{V}^{T+1:T+K} = f(X^{1:T}, C^{1:T+K}, E; \theta) \qquad (1)$$

In this paper most models are deep neural networks, trained with stochastic gradient descent to maximize a Gaussian Likelihood. More specifically, the optimal parameters θ^* are obtained by minimizing the mean squared error over valid pixels $V_*^t = V^t \odot M_Q^t \odot M_L$, where $M_Q \in \{0,1\}^{H\times W}$ masks pixels that are cloudy, cloud shadow or snow, $M_L \in \{0,1\}^{H\times W}$ masks pixels that are not cropland, forest, grassland or shrubland and $\odot$ denotes elementwise multiplication. Hence the training objective (leaving out dimensions for simplicity) is

$$\theta^* = \underset{\theta}{argmin} \frac{\sum (V - \hat{V})^2 \odot M_Q \odot M_L}{\sum M_Q \odot M_L} \qquad (2)$$

In this work $H = W = 128$px, $T = 10$ and $K = 20$.

3.2. Our proposed Contextformer model

To tackle the vegetation forecasting task, we develop the Contextformer, a multi-modal transformer model operating on local spatial context patches (hence the name). Next to historical satellite imagery, it leverages an elevation map and meteorological data to predict vegetation status.

Overview Our proposed Contextformer adopts an *encode-process-decode* [9] setup. Encoders and decoders work spatially without temporal fusion, while the processor temporally translates latent features in local context patches. It includes meteo and vision encoders, a temporal transformer processor, and a decoder predicting the delta from the last cloud-free NDVI observation (see Fig. 2).

Encoders The meteo encoder (for weather) and the delta decoders are parameterized as multi-layer perceptrons (MLPs) (Fig. 2 red boxes). For the vision encoder (Fig. 2 yellow boxes), we follow the MMST-ViT model for crop yield prediction [40] and use a Pyramid Vision Transformer (PVT) v2 B0 [78, 79], which is particularly suitable for dense prediction tasks. It divides the images of each time step into patches of 4 × 4 px and then creates patch-wise embeddings with a global receptive field. We merge multi-scale features from all stages of an ImageNet pre-trained PVT v2 B0, upscale them, concatenate and project. The features for each image stack (satellite & elevation) contain uni-temporal multi-scale and spatial context information.

Masked Token Modeling During training, we implement masked token modeling [25], randomly switching ($p = 0.5$) between inference mode (dropping all patches for time steps 10 to 30) and random dropout mode (masking 70% of patches for time steps 3 to 30). At test time, only inference mode is used to prevent exposure to future images. Cloudy patches are removed using the cloud mask. After replacing the masked values with a learned masking token, each patch receives sinusoidal temporal positional encoding [75] and weather embeddings from the meteo encoder.

Processor The temporal transformer (Fig. 2 green box) processes patches in parallel across 30 time steps, but spatially only within each 4×4 px patch. The idea here is that for ecosystem processes, spatial context is crucial but does not change dynamically. To address Sentinel 2's sub-pixel inaccuracies causing slight shifts over time, we maintain a small local context (4×4 px) within the temporal encoder. This approach substantially reduces memory usage during training (by 16×), allowing for larger batch sizes. Our implementation of the temporal transformer follows Presto's transformer encoder [75], based on the standard ViT [19].

Output Our Contextformer exploits vegetation dynamics persistence by predicting deviations from an initial state. More specifically, we use the last cloud-free NDVI observation from the historical period (10 time steps) as the initial prediction $\hat{V}^0$. Then, the delta decoder predicts a deviation δ^i for each future period token embedding (Fig. 2 right side). The final NDVI prediction is computed as $\hat{V}^i = \hat{V}^0 + \delta^i$. SGED-ConvLSTM [33] employed a similar delta framework, but predicting deviations iteratively, which in our multi-step prediction setting would result in cumulative outputs with undesirable training gradients.

3.3. GreenEarthNet Dataset

We present GreenEarthNet, a tailored dataset for high-resolution geospatial vegetation forecasting. It contains spatio-temporal minicubes [41], that are a collection of 30 5-daily satellite images (10 historical, 20 future), 150 daily meteorological observations and an elevation map. Spatial dimensions are 128×128px (2.56×2.56km). To enable cross-dataset model comparisons, we re-use the training locations and predictor dimensions from the EarthNet2021 [60] dataset for Earth surface forecasting.

Satellite and Meteo Layers GreenEarthNet includes Sentinel 2 [43] satellite bands blue, green, red, and near-infrared at 20m (consistent with EarthNet2021) and E-OBS [13] interpolated meteorological station data, which represents high quality meteorology over Europe [7]. More specifically, the meteorological drivers wind speed, relative humidity, and shortwave downwelling radiation, alongside the rainfall, sea-level pressure, and temperature (daily mean, min & max) are included. To enable reproducible research and minicube generation anywhere on Earth, we open source a Python package called *EarthNet Minicuber*[1], which generates multi-modal minicubes in a cloud native manner: only downloading the data chunks actually needed, instead of a full Sentinel 2 tile.

Cloud mask Vegetation proxies derived from optical satellite imagery are only meaningful if observations with clouds, shadows and snow are excluded such that anomalies due to clouds can be distinguished from vegetation anomalies. We train a UNet with Mobilenetv2 encoder [64] on the CloudSEN12 dataset [4] to detect clouds and cloud shadows from RGB and Nir bands. Tab. 1 compares precision, recall and F1 scores for detecting faulty pixels. Our approach outperforms Sen2Cor [43] (used in EarthNet2021), FMask [55] and KappaMask [18] baselines by a large margin. If using Sen2Cor in addition, to allow for snow masking, precision drops, but recall increases: i.e. the cloud mask gets more conservative. Using all 13 Sentinel 2 L2A bands is better than just using 4 bands, however such a model is not directly applicable to GreenEarthNet.

Algorithm	Works /w GreenEarthNet	Prec	Rec	F1
Sen2Cor	Yes	0.83	0.60	0.70
FMask	No	0.85	0.85	0.85
KappaMask	No	0.74	0.88	0.81
UNet RGBNir	Yes	0.91	0.90	0.90
UNet+Sen2CorSnow	Yes	0.83	**0.93**	0.88
UNet 13Bands	No	**0.94**	0.92	**0.93**

Table 1. Precision, recall and F1-score of different Sentinel 2 cloud masking algorithms.

Test sets Due to meso-scale circulation patterns, weather has high spatial correlation lengths. For GreenEarthNet, we design test sets ensuring independence not only in the high-resolution satellite data but also in the coarse-scale meteorology between training and test minicubes. More specifically, we introduce the subsets

- *Train*, 23816 minicubes in years 2017-2020
- *Val* 245 minicubes close to training locations, year 2020
- *OOD-t test* same locations as Val, years 2021-2022
- *OOD-s test*, 800 minicubes stratified over 1° × 1° lat-lon grid cells outside training regions, years 2017-2019
- *OOD-st test* same as OOD-s, but years 2021-2022

OOD-t serves as the primary test set, assessing the models' ability to extrapolate in time. Similarly, *val* enables early stopping based on temporal extrapolation skill. *OOD-s* and *OOD-st* assess spatial and spatio-temporal extrapolation respectively. These test sets align with EarthNet2021: *Val*/*OOD-t* locations overlap with EarthNet2021 IID tests, while *OOD-s*/*OOD-st* locations are distant from EarthNet2021 training data. Minicubes are created for four periods during the European growing season [63] each year: March-May (MAM), May-July (MJJ), July-September (JAS), and September-November (SON).

Additional Layers We add the ESA Worldcover Landcover map [90] for selecting only vegetated pixels during evaluation, the Geomorpho90m Geomorphons map [2] for further evaluation and the ALOS [72], Copernicus [20] and NASA [14] DEMs, to provide uncertainty in the elevation maps. Finally, we provide georeferencing for each minicube, enabling their extension with further data.

[1] https://pypi.org/project/earthnet-minicuber/

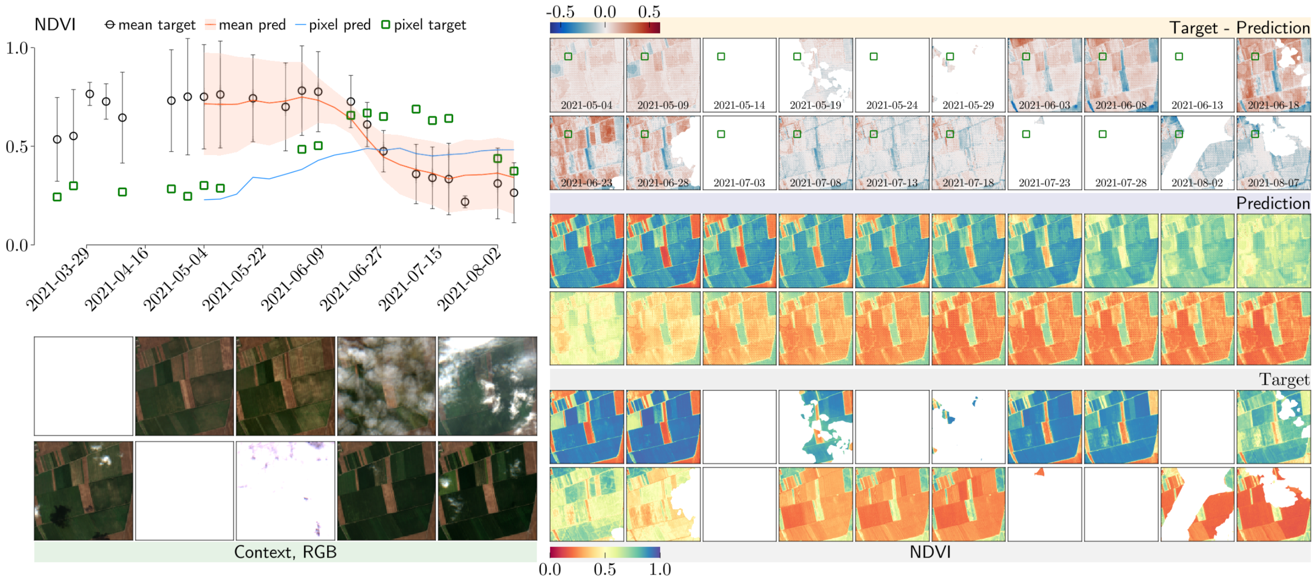

Figure 3. Qualitative Results of Contextformer for one OOD-t minicube located near Oradea, Romania. The top-left shows timeseries for all pixels (mean and std. dev.) and for a single pixel (green square on top right). The right side shows image timeseries of cloud-masked target and predicted NDVI alongside their difference.

3.4. Evaluation

We resort to traditional metrics in environmental modeling:

- R^2, the squared pearson correlation coefficient
- RMSE, the root mean squared error
- NSE $= 1 - \frac{MSE(V, \hat{V})}{Var[V]}$, the nash-sutcliffe efficiency [51], a measure of relative variability
- $|\text{bias}| = |\overline{V} - \overline{\hat{V}}|$, the absolute bias

In addition, we propose to measure if a model is better than the NDVI climatology, by computing the *Outperformance score*: the percentage of minicubes, for which the model is better in at least 3 out of the 4 metrics. Here, better means their score difference (ordering s.t. higher=better) exceeds 0.01 for RMSE and |bias| and 0.05 for NSE and R^2. We also report the RMSE over only the first 25 days (5 time steps) of the target period.

We compute all metrics per pixel over clear-sky timesteps. We then consider only pixels with vegetated landcover (cropland, grassland, forest, shrubland), no seasonal flooding (minimum NDVI > 0), enough observations (≥ 10 during target period, ≥ 3 during context period) and considerable variation (NDVI std. dev > 0.1). All these pixelwise scores are grouped by minicube and landcover, and then aggregated to account for class imbalance. Finally, the macro-average of the scores per landcover class is computed. In this way, the scores represent a conservative estimate of the expected performance of dynamic vegetation modeling during a new year or at a new location.

3.5. Baselines

We evaluate Contextformer against diverse baselines, spanning non-ML methods, time series forecasting, top EarthNet2021 performers [60], classical models, and two cutting-edge video prediction models. This selection aims to address uncertainty in optimal models for vegetation forecasting, considering factors like spatial context relevance. While existing spatio-temporal earth surface forecasting models are strong baselines, recent advances in video prediction, treating satellite image time series as video data, may offer competitive advantages.

Non-ML baselines We evaluate three non-ML baselines related to ecological memory: persistence [60] (last cloud free NDVI pixel), previous year [61] (linearly interpolated) and climatology (mean NDVI seasonal cycle).

Local time series models We compare against three common time series models: Kalman filter, LightGBM [32] and Prophet [74] from the Python library darts [27]. These are trained on timeseries from a single pixel and applied to forecast this pixel, given future weather as covariates. They are expensive to run: a single minicube takes $\sim$ 3h on an 8-CPU machine, $\mathcal{O}(10^4)$ slower than deep learning. We also evaluate a global timeseries model: the LSTM (ConvLSTM with 1x1 kernel). The time series models should be strong if spatial interactions are less predictive for vegetation.

EarthNet2021 models We also evaluate the Top-3 models from the EarthNet2021 challenge leaderboard[2] using their trained weights: a regular ConvLSTM [17], an encode-

[2] https://web.archive.org/web/20230228215255/https://www.earthnet.tech/en21/ch-leaderboard/

	Model	R^2 ↑	RMSE ↓	NSE ↑	|bias| ↓	Outperform ↑ Climatology	RMSE ↓ 25 days	#Params
NON-ML	Persistence	0.00	0.23	-1.28	0.17	21.8%	0.09	0
	Previous year	0.56	0.20	-0.40	0.14	19.3%	0.18	0
	Climatology	0.58	0.18	-0.34	0.13	n.a.	0.16	0
LOCAL TS	Kalman filter	0.41	0.19	-0.57	0.13	27.0%	0.16	$\mathcal{O}(10)$
	LightGBM	0.51	0.17	-0.22	0.12	42.2%	0.11	n.a.
	Prophet	0.57	0.16	-0.05	0.11	60.6%	0.13	$\mathcal{O}(10)$
EN21	ConvLSTM [17]	0.51	0.18	-0.37	0.12	43.9%	0.12	0.2M
	SG-ConvLSTM [33]	0.53	0.19	-0.33	0.14	45.8%	0.11	0.7M
	Earthformer [22]	0.49	0.17	-0.27	0.12	47.2%	0.11	60.6M
THIS STUDY	ConvLSTM [61]	0.58 ±0.01	0.16 ±0.00	-0.13 ±0.02	0.11 ±0.00	53.1% ±1.2%	0.11 ±0.00	1.0M
	Earthformer [22]	0.52	0.16	-0.13	0.10	56.5%	0.09	60.6M
	PredRNN [81]	**0.62** ±0.00	0.15 ±0.00	0.03 ±0.00	0.10 ±0.00	64.7% ±1.2%	0.10 ±0.00	1.4M
	SimVP [73]	0.60 ±0.00	0.15 ±0.00	0.03 ±0.01	**0.09** ±0.00	64.1% ±1.0%	0.10 ±0.00	6.6M
	Contextformer (Ours)	**0.62** ±0.00	**0.14** ±0.00	**0.09** ±0.01	**0.09** ±0.00	**66.8%** ±0.3%	**0.08** ±0.00	6.1M

Table 2. Quantitative Results. Mean (±std. dev.) are computed from three different random seeds.

process-decode ConvLSTM called SGED-ConvLSTM [33] and the Earthformer [22], a transformer model using cuboid-attention.

Additionally, we train and fine-tune both the ConvLSTM and Earthformer on GreenEarthNet. For the ConvLSTM, we follow the original Shi et al. [66] encoding-forecasting setup, which is different from ConvLSTM flavors studied on EarthNet2021 [17, 33] but has demonstrated improved performance on a similar problem in Africa [61]. We condition the Earthformer [22] through early fusion during historical steps and latent fusion during future steps.

Video prediction models We adapt PredRNN, SimVP, and two UNet-based models. The next-frame UNet [56] predicts one step ahead, while the next-cuboid UNet [60] predicts all steps at once. PredRNN [80, 81] has improved autoregressive information flow, and SimVP [73] performs direct multi-step prediction. We enhance both with weather conditioning using feature-wise linear modulation [54].

3.6. Implementation details

We build all of our ConvNets with a PatchMerge-style architecture similar to Earthformer [22]. For SimVP and PredRNN, such encoders and decoders are more powerful, but have slightly more parameters than in the original papers. We use GroupNorm [86] and LeakyReLU activation [87] for ConvNets and ConvLSTMs. For Contextformer, we use LayerNorm [5] and GELU activation [26]. For ConvNets, skip connections preserve high-fidelity. Our framework is implemented in PyTorch, and models are trained on Nvidia A40 and A100 GPUs. We use the AdamW [42] optimizer and tune the learning rate and a few hyperparameters per model. More implementation details can be found in the supplementary materials.

4. Experiments

4.1. Baseline comparison

We conduct experiments for predicting vegetation state across Europe in 2021 and 2022 at $20m$ resolution and compare the Contextformer against a wide range of baselines. The quantitative results are shown in table 2. For Contextformer, ConvLSTM, PredRNN and SimVP, we report the mean (±std. dev.) from three different random seeds. Earthformer has an order of magnitude more parameters, making training more expensive, which is why we only report one random seed. We find the Contextformer outperforms (or performs on par with) every baseline on all metrics. It achieves $R^2 = 0.62$ and 0.14 RMSE on the full 100 days lead time, which is further improved to 0.08 RMSE during the first 25 days lead time. The closest competitors are PredRNN and SimVP, with PredRNN having on par $R^2 = 0.62$ and SimVP on par $|\text{bias}| = 0.09$.

The Contextformer and the other video prediction baselines trained in this study are the first models to outperform the Climatology baseline: the ConvLSTM reaches 53.1% outperformance score, while the Contextformer achieves 66.8% (with consistent ranking across thresholds, see supplementary material). For the top-3 models (PredRNN, SimVP and Contextformer) and all metrics, differences to the climatology are highly significant when tested for all pixels (with Wilcoxon signed-rank test, $\alpha = 0.001$), but also for each land cover or for smaller subsets of 100 minicubes. ConvLSTM and Earthformer have overall lower skill. They mostly excels at RMSE and |bias|, where they can perform similar to other methods, yet have way lower performance for NSE and R^2.

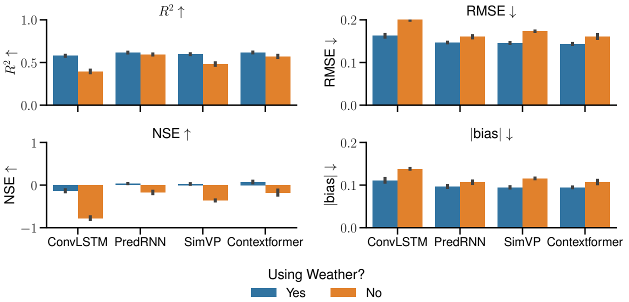

Figure 4. Model performance comparing meteo-guided models (blue) with the ablation not using weather (black bar is std. dev. from three random seeds).

The models trained on EarthNet2021 (ConvLSTM [17], SGED-ConvLSTM [33] and Earthformer [22]) perform poorly. None of the approaches consistently beats the Climatology, particularly the R^2 is much lower (from 0.49 for Earthformer up to 0.53 for SGED-ConvLSTM). Likely, this is a result of the focus on perceptual quality that was reflected in the EarthNet2021 metrics, as well as the overall lower data quality due to a faulty cloud mask.

Finally, other local time series baselines and non-ML baselines also underperform the Contextformer. The strongest pixelwise model is Prophet [74], with an outperformance score of 60.6%, followed by the climatology. Note, all of these baselines have access to a lot more information than the deep learning-based models (6 years vs. 50 days). Hence, this model comparison gives an indication, that spatial context is useful for vegetation forecasting, but leveraging them is challenging, as temporal dynamics are more dominant. In addition, the local time series models are all very slow, compared to the deep learning solutions presented in this work, which perform predictions within seconds (see supplementary material).

Qualitative results of the Contextformer model for one minicube are reported in fig. 3. The model clearly learns the complex dynamics of vegetation, with a strong seasonal evolution of the crop fields. It interpolates faithfully those pixels, which are masked in the target, and contains strong temporal consistency. However, as the lead time increases, predictions become less explicit, with a tendency towards oversmoothing.

4.2. Weather guidance

Our meteo-guided models benefit from the weather conditioning. Fig. 4 compares ConvLSTM, PredRNN, SimVP and Contextformer (blue) against a variant without weather conditioning (orange). For all metrics, using weather outperforms not using it. The ConvLSTM has the largest performance gain due to meteo-guidance, yet it is also the weakest model. This could possibly be due to the ConvLSTMs smaller receptive field and hence lower capacity at leveraging spatial context, which may to some degree compensate predictive capacity from weather.

Model	Original $R^2 \uparrow$	Shuffled $R^2 \uparrow$	Diff ↑
Climatology	0.58	-	-
1x1 LSTM	0.53	0.53	0.00
Next-frame UNet	0.51	0.48	-0.03
Next-cuboid UNet	0.56	0.43	-0.13
ConvLSTM	0.58	0.46	-0.12
PredRNN	0.62	0.45	-0.17
SimVP	0.60	0.49	-0.11
Contextformer	0.62	0.55	-0.07

Table 3. Model skill when spatial interactions are broken through shuffling.

For PredRNN and SimVP, we conduct an extended ablation study on weather guidance (see supplementary material). Weather conditioning methods (concatenation, FiLM [54], and cross-attention [62]) have a minor impact on performance when applied appropriately: cross-attention is most useful with latent fusion, FiLM outperforms concatenation, and is suitable for early fusion.

4.3. The role of spatial interactions

Unlike video prediction, satellite images show minimal spatial movement. Field and forest boundaries remain mostly fixed, with the largest variations occurring within these edges over time. Hence, it is unclear whether spatio-temporal models, accounting for interactions, are suitable for modeling vegetation dynamics. However, at $20m$ resolution, lateral processes may occur, not captured by predictors. For example, grasslands near a river or on a north-facing slope may react differently to meteorological drought. Additionally, weather affects trees at the forest edge differently from those in the center.

We compare model performance with spatially shuffled input, i.e. explicitly breaking spatial interactions [59]. We shuffle across batch and space, to also destroy image statistics. We evaluate Contextformer, ConvLSTM, PredRNN, and SimVP, skipping Earthformer due to high training cost. In addition we also study three baselines: a pixelwise (1x1) LSTM, the next-frame UNet and the next-cuboid UNet (see sec. 3.5). The pixelwise LSTM is a global timeseries model unable to capture spatial interactions. The next-frame UNet models spatial interactions, but does not consider temporal memory. All other models can leverage spatio-temporal dependencies, though the ConvLSTM only has a small local receptive field ($\sim$ 100m around each pixel). The results are reported in tab. 3. As can be expected, the pixelwise LSTM can be trained with spatial shuffled pixels without performance loss. All other models, though, exhibit a drop in per-

Ablation	$R^2 \uparrow$	RMSE $\downarrow$	Outperf $\uparrow$ Climatology
MLP vision encoder	0.58	0.15	58.3%
PVT encoder (frozen)	0.57	0.17	46.1%
PVT encoder	0.62	0.15	62.3%
/w cloud mask token	0.61	0.16	61.8%
/w learned $\hat{V}^0$	0.62	0.16	60.6%
/w last pixel $\hat{V}^0$	0.62	0.15	65.1%
Contextformer-6M	0.62	0.14	66.8%
Contextformer-16M	0.61	0.14	67.3%

Table 4. Model ablations. The Contextformer uses a PVT encoder, a cloud mask token and the last cloud free pixel as $\hat{V}_0$.

formance under pixel shuffling. For Contextformer, ConvLSTM, PredRNN and SimVP, R^2 drops by at least 0.07 and $RMSE$ increases by at least 0.04.

4.4. Ablation Study of Contextformer components

We conduct experiments to show how each key component in our Contextformer affects predictive skill. Tab. 4 lists the results of our ablation studies. First, we find that continued training of a pre-trained PVT vision encoder (outperformance score 62.3%) outperforms both a MLP vision encoder and a frozen pre-trained PVT. Second, adding the delta-prediction scheme with an initial vegetation state estimate $\hat{V}^0$ constructed by the last historical cloud free NDVI pixel further improves the outperformance to 65.1% – the version directly predicting NDVI is *PVT encoder*. Instead using a learned MLP decoder to estimate $\hat{V}^0$ is inferior. Third, using the cloud mask to drop out faulty tokens from the PVT encoder decreases model skill, if used alone, but if used on top of the delta-prediction scheme in the final model Contextformer-6M, it gives another boost to 66.8% outperformance. Finally, scaling the model size of the Contextformer to 16M parameters is not helpful when trained on GreenEarthNet, indicating the need for an even larger dataset for further performance gains.

4.5. Contextformer Strengths and Limitations

We assess the performance at spatio-(temporal) extrapolation of the Contextformer on the OOD-s and OOD-st test sets (tab. 5). The Contextformer can extrapolate in space and time. However, the margin to the climatology does shrink. Additional training data could be beneficial. While spatial extrapolation is theoretically not required for modeling vegetation dynamics (only temporal extrapolation is), it can enhance inference speed and expand applicability over larger areas. However, it is worth noting that certain locations may be harder to predict due to factors unrelated to weather, e.g. anthropogenic influences (harvesting, fires).

	OOD-s		OOD-st	
Model	$R^2 \uparrow$	RMSE $\downarrow$	$R^2 \uparrow$	RMSE $\downarrow$
Climatology	0.50	0.15	0.56	0.19
Contextformer	0.54	0.15	0.58	0.14

Table 5. Model skill at spatial (OOD-s) and spatio-temporal (OOD-st) extrapolation.

For practical applications another aspect needs to be studied in future work: at inference time weather comes from uncertain weather forecasts. Here, we first wanted to learn the impact of weather on vegetation and thus took the historical meteo data which has the least error. We expect the weather forecast uncertainty (represented by realizations / scenarios) to mostly propagate, but not present a covariate shift larger than the inter-annual variability, which our models are robust to (OOD-t evaluation).

5. Conclusion

We proposed Contextformer, a multi-modal transformer model designed for fine-resolution vegetation greenness forecasting. It leverages spatial context through a Pyramid Vision Transformer backbone while maintaining parameter efficiency. The temporal component is a transformer that independently models the dynamics of local context patches over time, incorporating meteorological data. We additionally introduce the novel GreenEarthNet dataset tailored for self-supervised vegetation forecasting and compare Contextformer against an extensive set of baselines.

Contextformer outperforms the previous state-of-the-art, especially on nash-sutcliffe efficiency and surpasses even strong freshly trained video prediction baselines. To our knowledge, we are the first to consider a climatology baseline and outperforming it with models. Given the pronounced seasonality of vegetation dynamics, this suggests real-world applicability for our models, particularly the Contextformer, in crucial scenarios like humanitarian anticipatory action or carbon monitoring.

Author contributions. VB experiments, figures, writing. CR experiments, writing. CRM supervision, figures, writing. ZG experiments. LA figures. NC, JC, NL, MW writing. MR funding, supervision, writing. All authors contributed to discussing the results.

Acknowledgments. We thank Reda ElGhawi, Christian Reimers, Annu Panwar and Xingjian Shi. MW thanks the European Space Agency for funding the project DeepExtremes (AI4Science ITT). CRM and LA (DeepCube, grant agreement No 101004188) and NL and JC (XAIDA, grant agreement No 101003469) thank the European Union's Horizon 2020 research and innovation programme.

References

[1] Rehaan Ahmad, Brian Yang, Guillermo Ettlin, Andrés Berger, and Pablo Rodríguez-Bocca. A machine-learning based ConvLSTM architecture for NDVI forecasting. *International Transactions in Operational Research*, 30(4):2025–2048, 2023. 2

[2] Giuseppe Amatulli, Daniel McInerney, Tushar Sethi, Peter Strobl, and Sami Domisch. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. *Scientific Data*, 7(1):162, 2020. 4

[3] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS journal of photogrammetry and remote sensing*, 140:20–32, 2018. 2

[4] Cesar Aybar, Luis Ysuhuaylas, Jhomira Loja, Karen Gonzales, Fernando Herrera, Lesly Bautista, Roy Yali, Angie Flores, Lissette Diaz, Nicole Cuenca, Wendy Espinoza, Fernando Prudencio, Valeria Llactayo, David Montero, Martin Sudmanns, Dirk Tiede, Gonzalo Mateo-García, and Luis Gómez-Chova. CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2. *Scientific Data*, 9(1):782, 2022. 4

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv*, 1607.06450, 2016. 6

[6] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. FitVid: Overfitting in Pixel-Level Video Prediction. *arxiv*, 2106.13195, 2021. 2

[7] Moritz Bandhauer, Francesco Isotta, Mónika Lakatos, Cristian Lussana, Line Båserud, Beatrix Izsák, Olivér Szentes, Ole Einar Tveito, and Christoph Frei. Evaluation of daily precipitation analyses in e-obs (v19. 0e) and era5 by comparison to regional high-resolution datasets in european regions. *International Journal of Climatology*, 42(2):727–747, 2022. 4

[8] Adam B. Barrett, Steven Duivenvoorden, Edward E. Salakpi, James M. Muthoka, John Mwangi, Seb Oliver, and Pedram Rowhani. Forecasting vegetation condition for drought early warning systems in pastoral communities in Kenya. *Remote Sensing of Environment*, 248:111886, 2020. 2

[9] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arxiv*, 1806.01261, 2018. 3

[10] Vitus Benson. Code and pre-trained model weights for benson et. al., CVPR (2024) - multi-modal learning for geospatial vegetation forecasting. *Zenodo*, 10.5281/zenodo.10793870, 2024. 1

[11] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970): 533–538, 2023. 2

[12] 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. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 3

[13] Richard C. Cornes, Gerard van der Schrier, Else J. M. van den Besselaar, and Philip D. Jones. An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409, 2018. 4

[14] R. Crippen, S. Buckley, P. Agram, E. Belz, E. Gurrola, S. Hensley, M. Kobrick, M. Lavalle, J. Martin, M. Neumann, Q. Nguyen, P. Rosen, J. Shimada, M. Simard, and W. Tung. NASADEM global elevation model: Methods and progress. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 125–128. Copernicus GmbH, 2016. 4

[15] Changlu Cui, Wen Zhang, ZhiMing Hong, and LingKui Meng. Forecasting ndvi in multiple complex areas using neural network techniques combined feature engineering. *International Journal of Digital Earth*, 13(12):1733–1749, 2020. 2

[16] Mauro Dalla Mura, Saurabh Prasad, Fabio Pacifici, Paulo Gamba, Jocelyn Chanussot, and Jón Atli Benediktsson. Challenges and opportunities of multimodality and data fusion in remote sensing. *Proceedings of the IEEE*, 103(9): 1585–1601, 2015. 2

[17] 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. 2, 5, 6, 7

[18] Marharyta Domnich, Indrek Sünter, Heido Trofimov, Olga Wold, Fariha Harun, Anton Kostiukhin, Mihkel Järveoja, Mihkel Veske, Tanel Tamm, Kaupo Voormansik, Aire Olesk, Valentina Boccia, Nicolas Longepe, and Enrico Giuseppe Cadau. KappaMask: AI-Based Cloudmask Processor for Sentinel-2. *Remote Sensing*, 13(20):4100, 2021. 4

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2020. 3, 4

[20] ESA. Copernicus DEM - Global and European Digital Elevation Model (COP-DEM). 2021. 4

[21] Aya Ferchichi, Ali Ben Abbes, Vincent Barra, and Imed Riadh Farah. Forecasting vegetation indices from spatio-temporal remotely sensed data using deep learning-based approaches: A systematic literature review. *Ecological Informatics*, page 101552, 2022. 2

[22] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring Space-

Time Transformers for Earth System Forecasting. In *Advances in Neural Information Processing Systems*, 2022. 2, 3, 6, 7

[23] 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. 2

[24] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[26] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv*, 1606.08415, 2023. 6

[27] Julien Herzen, Francesco Lässig, Samuele Giuliano Piazzetta, Thomas Neuer, Léo Tafti, Guillaume Raille, Tomas Van Pottelbergh, Marek Pasieka, Andrzej Skrodzki, Nicolas Huguenin, Maxime Dumonal, Jan Kościsz, Dennis Bader, Frédérick Gusset, Mounir Benheddi, Camila Williamson, Michal Kosinski, Matej Petrik, and Gaël Grosch. Darts: User-Friendly Modern Machine Learning for Time Series. *Journal of Machine Learning Research*, 23 (124):1–6, 2022. 5

[28] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[29] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 3

[30] Lei Ji and A.J. Peters. Forecasting vegetation greenness with satellite and climate data. *IEEE Geoscience and Remote Sensing Letters*, 1(1):3–6, 2004. 2

[31] Lingjun Kang, Liping Di, Meixia Deng, Eugene Yu, and Yang Xu. Forecasting vegetation index based on vegetation-meteorological factor interactions with artificial neural network. In *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, pages 1–6. IEEE, 2016. 2

[32] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 5

[33] 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. 2, 4, 6, 7

[34] Spyros Kondylatos, Ioannis Prapas, Michele Ronco, Ioannis Papoutsis, Gustau Camps-Valls, María Piles, Miguel-Ángel Fernández-Torres, and Nuno Carvalhais. Wildfire Danger Prediction and Understanding With Deep Learning. *Geophysical Research Letters*, 49(17):e2022GL099368, 2022. 2

[35] Basil Kraft, Martin Jung, Marco Körner, Christian Requena Mesa, José Cortés, and Markus Reichstein. Identifying Dynamic Memory Effects on Vegetation State Using Recurrent Neural Networks. *Frontiers in Big Data*, 2, 2019. 1, 2

[36] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, page eadi2336, 2023. 2

[37] Thomas Lees, Gabriel Tseng, Clement Atzberger, Steven Reece, and Simon Dadson. Deep Learning for Vegetation Health Forecasting: A Case Study in Kenya. *Remote Sensing*, 14(3):698, 2022. 2

[38] Thomas Lees, Gabriel Tseng, Clement Atzberger, Steven Reece, and Simon Dadson. Deep learning for vegetation health forecasting: a case study in kenya. *Remote Sensing*, 14(3):698, 2022. 2

[39] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, 2022. 2

[40] Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, Li Chen, Shelby Williams, Robert Minvielle, Xiangming Xiao, Drew Gholson, Nicolas Ashwell, Tri Setiyono, Brenda Tubana, Lu Peng, Magdy Bayoumi, and Nian-Feng Tzeng. MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5774–5784, 2023. 3

[41] David Montero Loaiza, Guido Kraemer, Anca Anghelea, Cesar Luis Aybar Camacho, Gunnar Brandt, Gustau Camps-Valls, Felix Cremer, Ida Flik, Fabian Gans, Sarah Habershon, Chaonan Ji, Teja Kattenborn, Laura Martínez-Ferrer, Francesco Martinuzzi, Martin Reinhardt, Maximilian Söchting, Khalil Teber, and Miguel Mahecha. Data Cubes for Earth System Research: Challenges Ahead. *EarthArXiv*, 5649, 2023. 4

[42] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2022. 6

[43] Jérôme Louis, Vincent Debaecker, Bringfried Pflug, Magdalena Main-Knorn, Jakub Bieniarz, Uwe Mueller-Wilm, Enrico Cadau, and Ferran Gascon. SENTINEL-2 SEN2COR: L2A Processor for Users. In *Proceedings Living Planet Symposium 2016*, pages 1–8, Prague, Czech Republic, 2016. Spacebooks Online. 4

[44] Xianping Ma, Xiaokang Zhang, and Man-On Pun. A cross-modal multiscale fusion network for semantic segmentation of remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3463–3474, 2022. 3

[45] Yue Ma, Yingjie Hu, Glenn R. Moncrieff, Jasper A. Slingsby, Adam M. Wilson, Brian Maitner, and Ryan Zhenqi Zhou.

Forecasting vegetation dynamics in an open ecosystem by integrating deep learning and environmental variables. *International Journal of Applied Earth Observation and Geoinformation*, 114:103060, 2022. 2

[46] Francesco Martinuzzi, Miguel D Mahecha, Gustau Camps-Valls, David Montero, Tristan Williams, and Karin Mora. Learning extreme vegetation response to climate forcing: A comparison of recurrent neural network architectures. *EGUsphere*, 2023:1–32, 2023. 2

[47] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards Geospatial Foundation Models via Continual Pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023. 3

[48] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020. 2

[49] Derege Tsegaye Meshesha, Muhyadin Mohammed Ahmed, Dahir Yosuf Abdi, and Nigussie Haregeweyn. Prediction of grass biomass from satellite imagery in Somali regional state, eastern Ethiopia. *Heliyon*, 6(10), 2020. 1

[50] Charlie Nash, João Carreira, Jacob Walker, Iain Barr, Andrew Jaegle, Mateusz Malinowski, and Peter W. Battaglia. Transframer: Arbitrary frame prediction with generative models. *Trans. Mach. Learn. Res.*, 2023, 2022. 2

[51] J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970. 5

[52] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *1st Workshop on the Synergy of Scientific and Machine Learning Modeling @ ICML2023*, 2023. 2

[53] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *AI for Earth and Space Science, workshop at ICLR 2022*, 2202.11214, 2022. 2

[54] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. 6, 7

[55] Shi Qiu, Zhe Zhu, and Binbin He. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sensing of Environment*, 231: 111205, 2019. 4

[56] Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020. 6

[57] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021. 2

[58] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 3

[59] Christian Requena-Mesa, Markus Reichstein, Miguel Mahecha, Basil Kraft, and Joachim Denzler. Predicting Landscapes from Environmental Conditions Using Generative Networks. In *Pattern Recognition*, pages 203–217, Cham, 2019. Springer International Publishing. 7

[60] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1142, 2021. 2, 4, 5, 6

[61] Claire Robin, Christian Requena-Mesa, Vitus Benson, Lazaro Alonso, Jeran Poehls, Nuno Carvalhais, and Markus Reichstein. Learning to forecast vegetation greenness at fine resolution over Africa with ConvLSTMs. *Climate Change AI workshop at NeurIPS 2022*, 2022. 2, 5, 6

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 7

[63] Thomas Rötzer and Frank-M. Chmielewski. Phenological maps of Europe. *Climate Research*, 18(3):249–257, 2001. 4

[64] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 4

[65] Mohamad Hakam Shams Eddin and Juergen Gall. Focal-TSMP: Deep learning for vegetation health prediction and agricultural drought assessment from a regional climate simulation. *EGUsphere*, pages 1–50, 2023. 2

[66] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 2, 6

[67] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[68] Michael J. Smith, Luke Fleming, and James E. Geach. EarthPT: A foundation model for Earth Observation. *arxiv*, 2309.07207, 2023. 2, 3

[69] Max J Steinhausen, Paul D Wagner, Balaji Narasimhan, and Björn Waske. Combining sentinel-1 and sentinel-2 data for improved land use and land cover mapping of monsoon regions. *International journal of applied earth observation and geoinformation*, 73:595–604, 2018. 2

[70] Corinne Stucker, Vivien Sainte Fare Garnot, and Konrad Schindler. U-TILISE: A Sequence-to-Sequence Model for Cloud Removal in Optical Satellite Time Series. 61:1–16. 2

[71] Joan Sturm, Maria J. Santos, Bernhard Schmid, and Alexander Damm. Satellite data reveal differential responses of Swiss forests to unprecedented 2018 drought. *Global Change Biology*, 28(9):2956–2978, 2022. 1

[72] T. Tadono, H. Nagai, H. Ishida, F. Oda, S. Naito, K. Minakawa, and H. Iwamoto. Generation of the 30 m mesh global digital surface model by ALOS Prism. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B4:157–162, 2016. 4

[73] Cheng Tan, Zhangyang Gao, and Stan Z. Li. SimVP: Towards Simple yet Powerful Spatiotemporal Predictive Learning. *arxiv*, 2211.12509, 2023. 2, 6

[74] Sean J. Taylor and Benjamin Letham. Forecasting at Scale. *The American Statistician*, 72(1):37–45, 2018. 2, 5, 7

[75] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. *arxiv*, 2304.14065, 2023. 2, 3, 4

[76] 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*. Curran Associates, Inc., 2017. 3

[77] Lei Wang, Xin Xu, Yue Yu, Rui Yang, Rong Gui, Zhaozhuo Xu, and Fangling Pu. SAR-to-Optical Image Translation Using Supervised Cycle-Consistent Adversarial Networks. *IEEE Access*, 7:129136–129149, 2019. 2

[78] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2, 3

[79] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with Pyramid Vision Transformer. *Computational Visual Media*, 8(3):415–424, 2022. 2, 3

[80] 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*. Curran Associates, Inc., 2017. 2, 6

[81] 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, 2023. 2, 6

[82] Andrew Whyte, Konstantinos P Ferentinos, and George P Petropoulos. A new synergistic approach for monitoring wetlands using sentinels-1 and 2 data with object-based machine learning algorithms. *Environmental Modelling & Software*, 104:40–54, 2018. 2

[83] Aleksandra Wolanin, Gustau Camps-Valls, Luis Gómez-Chova, Gonzalo Mateo-García, Christiaan van der Tol, Yongguang Zhang, and Luis Guanter. Estimating crop primary productivity with Sentinel-2 and Landsat 8 using machine learning methods trained with radiative transfer simulations. *Remote Sensing of Environment*, 225:441–457, 2019. 1

[84] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, pages 22419–22430. Curran Associates, Inc., 2021. 2

[85] Haixu Wu, Zhiyu Yao, Jianmin Wang, and Mingsheng Long. MotionRNN: A Flexible Model for Video Prediction With Spacetime-Varying Motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15435–15444, 2021. 2

[86] Yuxin Wu and Kaiming He. Group Normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 6

[87] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical Evaluation of Rectified Activations in Convolutional Network. *arxiv*, 1505.00853, 2015. 6

[88] Xian Yang, Yifan Zhao, and Ranga Raju Vatsavai. Deep Residual Network with Multi-Image Attention for Imputing Under Clouds in Satellite Imagery. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 643–649, 2022. 2

[89] Fanglong Yao, Wanxuan Lu, Heming Yang, Liangyu Xu, Chenglong Liu, Leiyi Hu, Hongfeng Yu, Nayu Liu, Chubo Deng, Deke Tang, Changshuo Chen, Jiaqi Yu, Xian Sun, and Kun Fu. RingMo-Sense: Remote Sensing Foundation Model for Spatiotemporal Prediction via Spatiotemporal Evolution Disentangling. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2023. 3

[90] Daniele Zanaga, Ruben Van De Kerchove, Wanda De Keersmaecker, Niels Souverijns, Carsten Brockmann, Ralf Quast, Jan Wevers, Alex Grosu, Audrey Paccini, Sylvain Vergnaud, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Ivelina Georgieva, Myroslava Lesiv, Sarah Carter, Martin Herold, Linlin Li, Tsendbazar, Nandin-Erdene, Fabrizio Ramoino, and Olivier Arino. ESA WorldCover 10 m 2020 v100. 2021. 4

[91] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, 2023. 2

[92] Yelu Zeng, Dalei Hao, Alfredo Huete, Benjamin Dechant, Joe Berry, Jing M. Chen, Joanna Joiner, Christian Frankenberg, Ben Bond-Lamberty, Youngryel Ryu, Jingfeng Xiao, Ghassem R. Asrar, and Min Chen. Optical vegetation indices for monitoring terrestrial ecosystems globally. *Nature Reviews Earth & Environment*, pages 1–17, 2022. 2