EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task.

Christian Requena-Mesa[1,2,3,*], Vitus Benson[1,*], Markus Reichstein[1,4], Jakob Runge[2,5], Joachim Denzler[2,3,4]

1) Biogeochemical Integration, Max-Planck-Institute for Biogeochemistry, Jena, Germany
2) Institute of Data Science, German Aerospace Center (DLR), Jena, Germany
3) Computer Vision Group, University of Jena, Jena, Germany
4) Michael-Stifel-Center Jena for Data-driven and Simulation Science, Jena, Germany
5) Technische Universität Berlin, Berlin, Germany

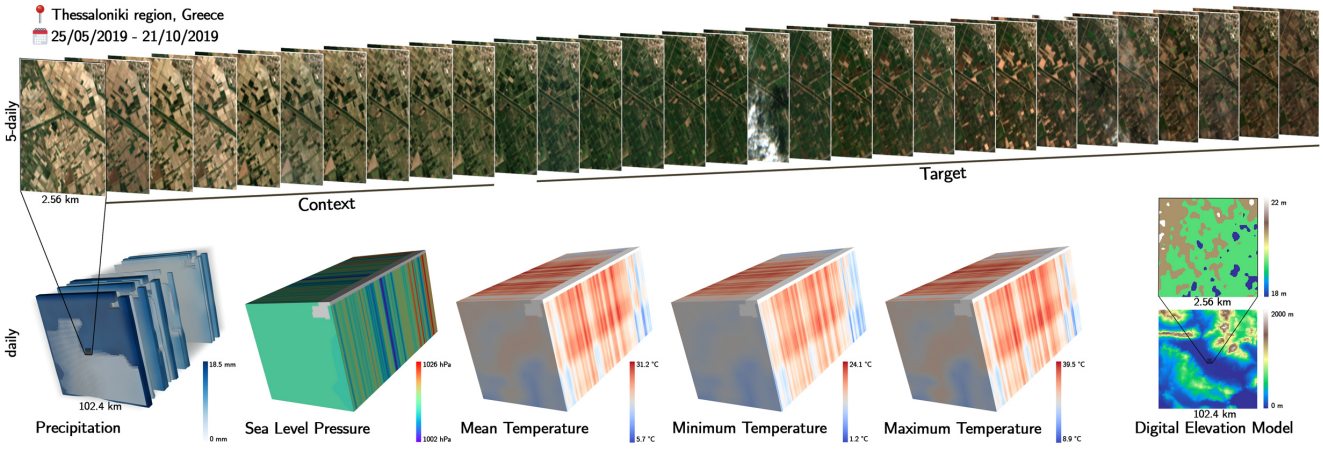

Figure 1: Overview visualization of one of the over 32000 samples in EarthNet2021.

Abstract

Satellite images are snapshots of the Earth surface. We propose to forecast them. We frame Earth surface forecasting as the task of predicting satellite imagery conditioned on future weather. EarthNet2021 is a large dataset suitable for training deep neural networks on the task. It contains Sentinel 2 satellite imagery at 20 *m resolution, matching topography and mesoscale* (1.28 *km) meteorological variables packaged into* 32000 *samples. Additionally we frame EarthNet2021 as a challenge allowing for model intercomparison. Resulting forecasts will greatly improve* ($> \times 50$) *over the spatial resolution found in numerical models. This allows localized impacts from extreme weather to be predicted, thus supporting downstream applications such as crop yield prediction, forest health assessments or biodiversity monitoring. Find data, code, and how to participate at* www.earthnet.tech.

1. Introduction

Seasonal weather forecasts are potentially very valuable in support of sustainable development goals such as zero hunger or life on land. Spatio-temporal deep learning is expected to improve the predictive ability of seasonal weather forecasting [52]. Yet it is unclear, how exactly this expectation will materialize. One possible way can be found by carefully thinking about the target variable. The above mentioned goals illustrate that ultimately it will not directly be the seasonal meteorological forecasts but rather derived impacts (e.g. agricultural output and ecosystem health) that are of most use to humanity. Such impacts, especially those affecting vegetation, materialize on the land surface. Meaning, they can be observed on satellite imagery. Thus, high-resolution impact forecasting can be phrased as the prediction of satellite imagery [15, 24, 29, 38, 73]. Prediction of future frames is also the metier of video prediction [2, 22, 37, 43, 46]. Yet, satellite image forecasting can also leverage additional future drivers, such as the output of numerical weather simulations with earth system models. The

*Joint first authors. {crequ,vbenson}@bgc-jena.mpg.de

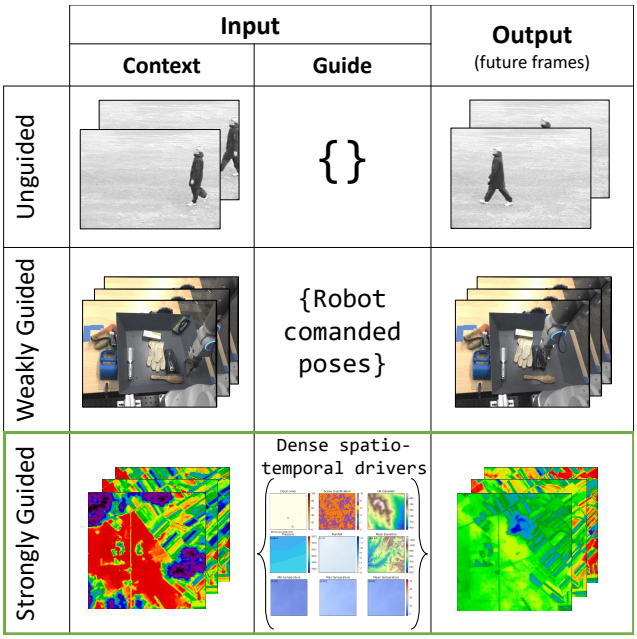

Figure 2: Video prediction could be unguided, weakly guided or strongly guided. EarthNet2021 is the first dataset specifically designed for the development of spatio-temporal strongly guided video prediction methods.

general setting of video prediction with additional drivers is called guided video prediction. We define *Earth surface forecasting* as the prediction of satellite imagery conditioned on future weather.

Our main **contributions** are summarized as follows:

- We motivate the novel task of Earth surface forecasting as guided video prediction and define an evaluation pipeline based the EarthNetScore ranking criterion.
- We introduce EarthNet2021, a carefully curated large-scale dataset for Earth surface forecasting conditional on meteorological projections.
- We start model intercomparison in Earth surface forecasting with a pilot study encouraging further research in deep learning video prediction models.

2. Related work

Earth surface forecasting lays in the intersection of video prediction and data-driven Earth system modeling.

Video prediction. In Fig. 2 we classify video prediction tasks into three types depending on the input data used. Traditionally video prediction models are conditioned on past context frames and predict future target frames (i.e. unguided, [17, 18, 22, 25, 31, 36, 40, 43, 60]). The used models inherit many characteristics known the be useful in modeling Earth surface phenomena: short-long term memory effects [35, 55], short-long range spatial relationships [53], as well as, the ability to generate stochastic predictions [2, 23, 37], ideal to generate ensemble forecasts of Earth surface for effective uncertainty management [72]. Guided video prediction is the setting where on top of past frames models have access to future information. We further differ between weak and strong guiding (Fig. 2). Weakly guided models are provided with sparse information of the future, for example robot commands [22]. In contrast strongly guided models leverage dense spatio-temporal information of the future. This is the setting of EarthNet2021. Some past works resemble the strongly guided setting, however either the future information are derived from the frames themselves, making the approaches not suitable for prediction [67] or they use the dense spatial information but discard the temporal component [42, 53].

Modelling weather impact with machine learning. Both impact modeling and weather forecasting have been tackled with machine learning methods of different complexity. One string of literature has focused on forecasting imagery from weather satellites [29, 38, 62, 70] while another one has focused on predicting reanalysis data or emulating general circulation models [50, 51, 57, 68]. For localized weather, statistical downscaling has been leveraged [5, 44, 64, 65] (Fig. 3A). Direct impacts of extreme weather have been predicted one at a time (Fig. 3B), examples being crop yield [1, 9, 32, 48, 58], vegetation index [15, 26, 49, 69], drought index [47] and soil moisture [19].

3. Motivation

While satellite imagery prediction is an interesting task for video prediction modelers, it is similarly important for domain experts, i.e., climate and land surface scientists. We focus on predicting localized impacts of extreme weather. This is highly relevant since extreme weather impacts very heterogeneously at the local scale [34]. Very local factors, such as vegetation, soil type, terrain elevation or slope, determine whether a plot is resilient to a heatwave or not. For example, ecosystems next to a river might survive droughts more easily than those on south-facing slopes. However, the list of all possible spatio-temporal interactions is far from being mechanistically known; hence, it is a source of considerable amount of uncertainty and an opportunity for powerful data-driven methods.

Predicting localized weather impacts can be tackled in three main ways (Fig. 3). All approaches make use of seasonal weather forecasts [8, 10](2 – 6 months ahead). The classical approach (Fig. 3A), attempts the hyper-resolution of the weather forecast for particular geolocations using statistical [7, 66] or dynamical [39] downscaling, that is, correlating the past observed weather with past mesoscale model outputs and using the estimated relationship. The downscaled weather can then be used in mechanistic models (e.g.

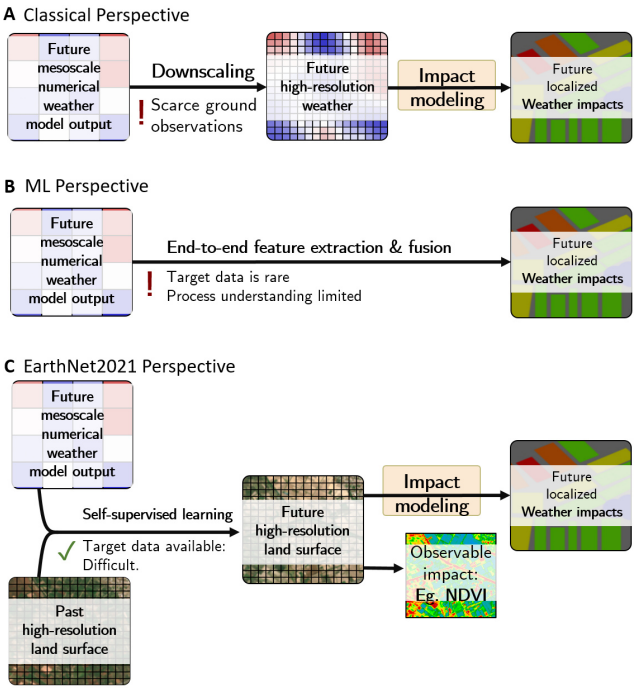

Figure 3: Three ways of extreme weather impact prediction are A) downscaling meteorological forecasts and subsequent impact modeling (*e.g.* runoff models), B) acquiring target data at high-resolution and using supervised learning or C) leveraging Earth surface forecasting: This gives directly obtainable impacts (*e.g.* NDVI) while still allowing for impact modeling. Compared to A) and B), large amounts of satellite imagery are available, thus together with self-supervised (i.e. no labeling required) deep learning large-scale impact prediction becomes feasible.

of river discharge) for impact extraction. However, weather downscaling is a difficult task because it requires ground observations from weather stations, which are sparse. A more direct way (Fig. 3B) is to correlate a desired future impact variable, such as crop yields or flood risk, with past data (e.g., weather data and vegetation status, [48]). Yet again, this approach requires ground truth data of target variables, which is scarce, thus limiting the global applicability of the approach.

Instead, by defining the task of Earth surface forecasting we propose to use satellite imagery as an intermediate step (Fig. 3C). From satellite imagery, multiple indices describing the vegetation state such as the normalized differenced vegetation index (NDVI) or the enhanced vegetation index (EVI) can be directly observed. These give insights on vegetation anomalies, which in turn describe the ecosystem impact at a very local scale. Because of satellite imagery's vast availability, there is no data scarcity. While technically difficult, forecasting weather impacts via satellite imagery prediction is feasible. Additionally, satellite imagery is also used to extract further processed weather impact data products, such as biodiversity state [21], crop yield [58], soil moisture [19] or ground biomass [49]. In short, Earth surface prediction is promising for forecasting highly localized climatic impacts.

4. EarthNet2021 Dataset

4.1. Overview

Data sources. With EarthNet2021 we aim at creating the first dataset for the novel task of Earth surface forecasting. The task requires satellite imagery time series at high temporal and spatial resolution and additional climatic predictors. The two primary public satellite missions for high-resolution optical imagery are Landsat and Sentinel 2. While the former only revisits each location on Earth every 16 days, the latter does so every five days. Thus we choose Sentinel 2 imagery [41] for EarthNet2021. The additional climatic predictors should ideally come from a seasonal weather model. Obtaining predictions from a global seasonal weather model starting at multiple past time steps is computationally very demanding. Instead, we approximate the forecasts using the E-OBS [14] observational dataset, which essentially contains interpolated ground truth observed weather from a number of stations over Europe. This also makes the task easier as uncertain weather forecasts are replaced with certain observations. Since E-OBS limits the spatial extent to Europe, we use the appropriate high-resolution topography: EU-DEM [3].

Individual samples. After data processing, EarthNet2021 contains over 32000 samples, which we call *minicubes*. A single minicube is visualized in Fig. 1. It contains 30 5-daily frames (128×128 pixel or 2.56×2.56 km) of four channels (blue, green, red, near-infrared) of satellite imagery with binary quality masks at high-resolution (20 m), 150 daily frames (80×80 pixel or 102.4×102.4 km) of five meteorological variables (precipitation, sea level pressure, mean, minimum and maximum temperature) at mesoscale resolution (1.28 km) and the static digital elevation model at both high- and mesoscale resolution. The minicubes reveal a strong difference between EarthNet2021 and classic video prediction datasets. In the latter, the objects move in a 3d space, but images are just a 2d projection of this space. For satellite imagery, this effect almost vanishes as the Earth surface locally is very similar to a 2d space.

4.2. Generation

Challenges. In general, geospatial datasets are not analysis-ready for standard computer vision. While the former often contain large files together with information about the projection of the data, the latter requires many small data sam-

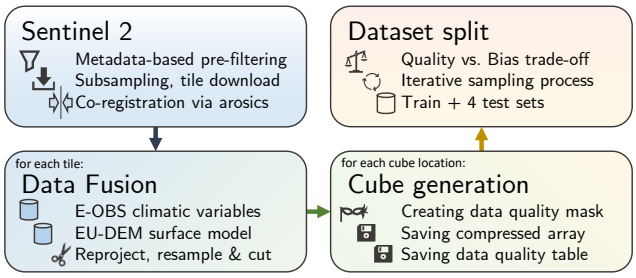

Figure 4: The dataset generation scheme of EarthNet2021.

ples on an Euclidean grid. With EarthNet2021 we aim to bridge the gaps and transform geospatial data into analysis-ready samples for deep learning. To this end, we had to gather the satellite imagery, combine it with additional predictors, generate individual data samples and split these into training and test sets – challenges which are described in the following paragraphs and lead to our dataset generation scheme, see Fig. 4.

Obtaining Sentinel 2 satellite imagery. Downloading the full archive of Sentinel 2 imagery over Europe would require downloading Petabytes, rendering the approach unfeasible. Luckily pre-filtering is possible as the data is split by the military grid reference system into so-called tiles and for each tile metadata can be obtained from the AWS Open Data Registry[1] before downloading. We pre-filter and only download a random subset of 110 tiles with at least 80% land visible on the least-cloudy day and minimum 90% data coverage. For each tile we download blue, green, red, near-infrared and scene-classification bands over the time series corresponding to the 5-day interval with the lowest off-nadir angle. We use the sentinel-hub library[2] to obtain an archive of over 30 TB raw imagery from November 2016 to May 2020. In it we notice spatial jittering between consecutive Sentinel 2 intakes, possibly due to the tandem satellites not being perfectly co-registered. We try to compensate this artifact by co-registering the time series of each tile. We use the global co-registration from the arosics library[3] [56] inside a custom loop.

Data fusion with E-OBS and EU-DEM. For each of the 110 tiles we fuse their time series with additional data. More particularly we gathered E-OBS[4] weather variables (daily mean temperature (TG); daily minimum temperature (TN); daily maximum temperature (TX); daily precipitation sum (RR); and daily averaged sea level pressure (PP)) at 11.1 km resolution and the EU-DEM[5] digital surface model at 25 m resolution. We re-project, resample and

[1] https://registry.opendata.aws/sentinel-2/
[2] https://sentinelhub-py.readthedocs.io/
[3] https://pypi.org/project/arosics/
[4] https://surfobs.climate.copernicus.eu/
[5] eea.europa.eu/data-and-maps/data/eu-dem/

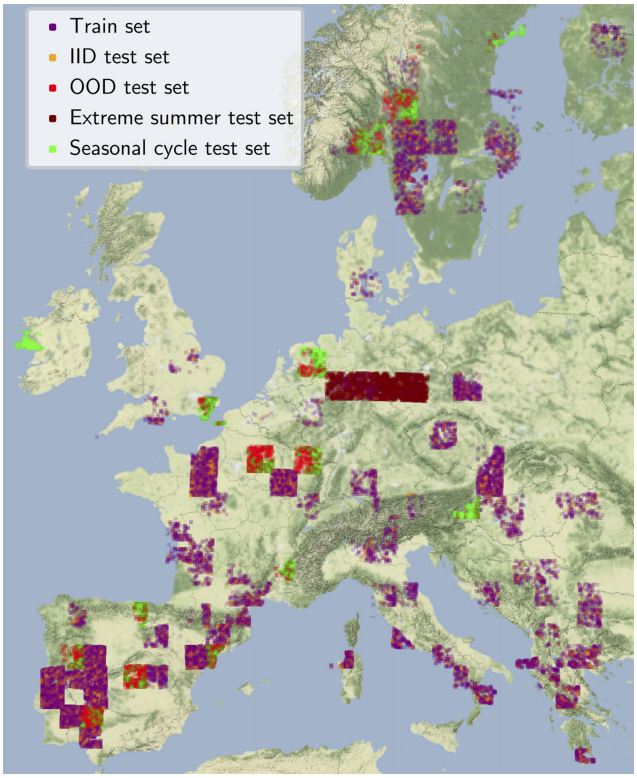

Figure 5: Spatial distribution of the samples in EarthNet2021.

cut them to two data windows. The high-resolution window has 20 m ground resolution, matching the Sentinel 2 imagery, and the mesoscale window has 1.28 km ground resolution.

Generation of minicubes. Given the fused data, we create a target minicube grid, cutting each tile into a regular spatial grid and a random temporal grid. Spatially, high-resolution windows of minicube do not overlap, while mesoscale windows do. Temporally, minicubes at the same location never overlap. For each location in the minicube grid, we extract the data from our fused archive, generate a data quality (i.e. cloud) mask based on heuristic rules (similar to [45]) and save the minicube in a compressed numpy array [28]. We also generate data quality indicators; these will be useful for selecting cubes during dataset splitting.

Creating the dataset split. In the raw EarthNet2021 corpus are over 1.3 million minicubes. Unfortunately, most of them are of very low quality, mainly because of clouds. Now just taking minicubes above a certain quality threshold creates another problem: selection bias. To give an intuitive example: most frequently, high-quality (cloud-free) samples are found during summer on the Iberian Peninsula, whereas there barely are 4 consecutive weeks without clouds on the

British Islands. We address this trade-off by introducing an iterative filtering process. Until 32000 minicubes are collected, we iteratively loosen quality restrictions for selecting high quality cubes and filling them up to obtain balance among starting months and between northern and southern geolocations. In a similar random-iterative process we separate 15 tiles from all downloaded tiles to create a spatial out-of-domain (OOD) test set (totalling 4214 minicubes) and randomly split the remainder tiles into 23904 minicubes for training and 4219 for in-domain (IID) testing.

4.3. Description

Statistics. The EarthNet2021 dataset spans across wider Central and Western Europe. Its training set contains 23904 samples from 85 regions (Sentinel 2 tiles) in the spatial extent. Fig. 5 visualizes the spatial distribution of samples. 71% of the minicubes in the training set lay in the southern half, which also contains more landmass. In the northern half we observe a strong clustering in the vicinity of the Oslofjord, which is possibly random. Temporally most minicubes cover the period of May to October (Fig. 6a). While this certainly biases the dataset, it might actually be desirable because some of the most devastating climate impacts (e.g., heatwaves, droughts, wildfires) occur during summer. Fig. 6b shows the bias-quality trade-off, observe that most high quality minicubes are from summer in the Mediterranean. Also, it shows that EarthNet2021 does not contain samples covering winter in the northern latitudes. This is possibly an effect of our very restrictive quality masking wrongly classifying snow as clouds.

Comparison to other datasets. Earth surface forecasting is a novel task, thus there are no such datasets prior to EarthNet2021. Still, since it also belongs to the broader set of analysis-ready datasets for deep learning, we can assert that it is large enough for training deep neural networks. In supplementary table 3 we compare a range of datasets using either satellite imagery or being targeted to video prediction models. By pure sample size, EarthNet2021 ranks solid, yet, the number is misleading since individual samples are different. By additionally comparing the size in gigabytes, we assert that EarthNet2021 is indeed a large dataset.

Limitations. Clearly, EarthNet2021 limits models to work solely on Earth surface forecasting in Europe. Additionally, the dataset is subject to a selection bias; therefore, there are areas in Europe for which model generalizability could be problematic. Furthermore, EarthNet2021 leverages observational products instead of actual forecasts. Thus, while this certainly is practical for a number of reasons, Earth surface models trained on EarthNet2021 should be viewed as experimental and might not be plug-and-play into production with seasonal weather model forecasts.

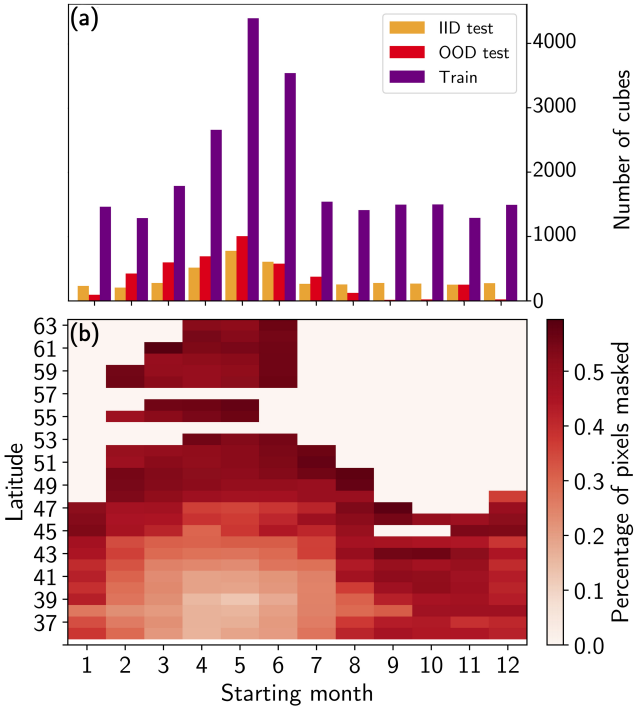

Figure 6: Monthly bias of samples. (a) Shows the monthly number of minicubes and (b) shows the data quality measured by the percentage of masked (mainly cloudy) pixels over both, months and latitude.

5. EarthNet2021 Challenge

5.1. Overview

Model intercomparison. Modeling efforts are most useful when different models can easily be compared. Then, strengths and weaknesses of different approaches can be identified and state-of-the-art methods selected. We propose the EarthNet2021 challenge as a model intercomparison exercise in Earth surface forecasting built on top of the EarthNet2021 dataset. This motivation is reflected in the design of the challenge. We define an evaluation protocol by which approaches can be compared as well as provide a framework, such that knowledge between modelers is easily exchanged. There is no reward other than scientific contribution and a publicly visible leaderboard. Evaluating Earth surface forecasts is not trivial. Since it is a new task, there is not yet a commonly used criterion. We design the EarthNetScore as a ranking criterion balancing multiple goals and center the evaluation pipeline around it (see Fig. 7). Moreover, we motivate four challenge tracks. These allow comparison of models' validity and robustness and applicability to extreme events and the vegetation cycle.

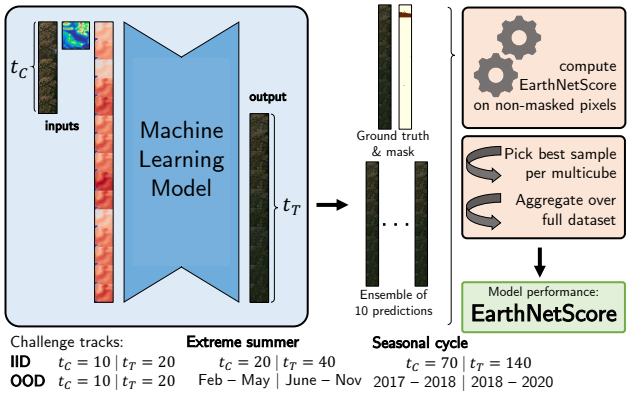

Figure 7: Evaluation pipeline for models on EarthNet2021. For predicting the t_T target frames, a model can use satellite images from the t_C context frames, the static DEM and mesoscale climatic variables including those from the target time steps.

EarthNet2021 framework. To facilitate research we provide a flexible framework which kick-starts modelers and removes double work between groups. The evaluation pipeline is packaged in the EarthNet2021 toolkit and leverages multiprocessing for fast inference. Additionally, challenge participants are encouraged to use the model intercomparison suite. It shall give one entry point for running a wide range of models. Currently it features model templates in PyTorch and TensorFlow and additional graphical output useful for debugging and visual comparison.

5.2. EarthNetScore

Components. Evaluating Earth surface predictions is nontrivial. Firstly, because the topological space of spectral images is not a metric space, secondly, because clouds and other data degradation hinder evaluation, and, thirdly, because ranking simply by root mean squared error might lead to ordinal rankings of imperfect predictions that experts would not agree to. Instead of using a single score, we define the *EarthNetScore* ENS by combining multiple components. As the first component, we use the median absolute deviation MAD. It is a robust distance in pixel-space which is justified by the goal that predicted and target values should be close. Secondly, OLS, the difference of ordinary least squares linear regression slopes of pixelwise Normalized Difference Vegetation Index (NDVI) timeseries, gives an indicator as to whether the predictions are able to reproduce the trend in vegetation change. This together with the Earth mover distance EMD between pixelwise NDVI time series over the short span of 20 time steps is a good proxy of the fit (in the distribution and direction) of the vegetation time series. The time series based metrics OLS and EMD are also largely robust to missing data points, which poses a consistency constraint on model predictions at output pixels for which no target data is available. Finally, the structural similarity index $SSIM$ is a perceptual metric imposing predicted frames to have similar spatial structure to the target satellite images. All component scores are modified to work properly in the presence of a data quality mask, rescaled to match difficulty and transformed to lay between 0 (worst) and 1 (best).

Computation. Combining these components is another challenge. We would like to define the EarthNetScore as:

$$ENS = \frac{4}{(\frac{1}{MAD} + \frac{1}{OLS} + \frac{1}{EMD} + \frac{1}{SSIM})}. \quad (1)$$

This is the harmonic mean of the four components, so it lays strongly to the worst performing component. Yet, computing ENS over a full test set requires further clarification. Earth surface forecasting is a stochastic prediction task, models are allowed to output an ensemble of predictions. Thus, for each minicube in the test set there might multiple predictions (up to 10). In line with what is commonly done in video prediction, we compute the subscores for each one of them but, only take the prediction for which equation 1 is highest for model intercomparison. In other words, the evaluation pipeline only accounts for the best prediction per minicube. This is superior to average predictions as it allows for an ensemble of discrete, sharp, plausible outcomes, something desired for Earth surface forecasting given its highly multimodal nature. Still, this evaluation scheme, suffers severely from not being able to rank models according to their modeled distribution. Once the components for the best predictions for all minicubes in the dataset are collected, we average each component and then calculate the ENS by feeding the averages to equation 1. Then, the ENS ranges from 0 to 1, where 1 is a perfect prediction.

5.3. Tracks

Main (IID) track. The EarthNet2021 main track checks model validity. It uses the IID test set, which has minicubes that are very similar (yet randomly split) as those seen during training. Models get 10 context frames of high resolution 5-daily multispectral satellite imagery (time [t-45, t]), static topography at both mesoscale and high resolution, and mesoscale dynamic climate conditions for 150 past and future days (time [t-50, t+100]). Models shall output 20 frames of high-resolution sentinel 2 bands red, green, blue and near-infrared for the next 100 days (time [t+5,t+100]). These predictions are then evaluated with the EarthNetScore on cloud-free pixels from the ground truth. This track follows the assumption that, in production, any Earth surface forecasting model would have access to all

	IID					OOD				
	ENS	**MAD**	**OLS**	**EMD**	**SSIM**	**ENS**	**MAD**	**OLS**	**EMD**	**SSIM**
Persistence	0.2625	0.2315	0.3239	0.2099	0.3265	0.2587	0.2248	0.3236	0.2123	0.3112
Channel-U-Net	0.2902	0.2482	0.3381	0.2336	0.3973	0.2854	0.2402	0.3390	0.2371	0.3721
Arcon	0.2803	0.2414	0.3216	0.2258	0.3863	0.2655	0.2314	0.3088	0.2177	0.3432

Table 1: Models performance on EarthNet2021. See supplementary material for Extreme and Seasonal test sets.

prior Earth observation data, thus the test set has the same underlying distribution as the training set.

Robustness (OOD) track. In addition to the main track, we offer a robustness track. Even on the same satellite data, deep learning models might generalize poorly across geolocations [6], thus it is important to check model performance on an out-of-domain (OOD) test set. This track has a weak OOD setting; in which minicubes solely are from different Sentinel 2 tiles than those seen during training, which is possibly only a light domain shift. Still, it is useful as a first benchmark to check applicability of models outside the training domain.

Extreme summer track. Furthermore, EarthNet2021 contains two tracks particularly focused on Earth system science hot topics, which should both be understood as more experimental. The extreme summer track contains cubes from the extreme summer 2018 in northern Germany [4], with 4 months of context (20 frames) starting from February and 6 months (40 frames) starting from June to evaluate predictions. For these locations, only cubes before 2018 are in the training set. Being able to accurately downscale the prediction of an extreme heat event and to predict the vegetation response at a local scale would greatly benefit research on resilience strategies. In addition, the extreme summer track can in some sense be understood as a temporal OOD setting.

Seasonal cycle track. While not the focus of EarthNet2021, models are likely able to generate predictions for longer horizons. Thus, we include the seasonal cycle track covering multiple years of observations; hence, including vegetation full seasonal cycle. This track is also in line with the recently rising interest in seasonal forecasts within physical climate models. It contains minicubes from the spatial OOD setting also used for the robustness tracks, but this time each minicube comes with 1 year (70 frames) of context frames and 2 years (140 frames) to evaluate predictions. For this longer prediction length, we change the EarthNetScore OLS component to be calculated over disjoint windows of 20 frames each.

6. Models

As first baselines in the EarthNet2021 model intercomparison, we provide three models. One is a naive averaging persistence baseline while the other two are deep learning models slightly modified for guided video prediction. Performance is reported in Tab. 1.

Persistence baseline The EarthNet2021 unified toolkit comes with a pre-implemented baseline in NumPy. It simply averages cloud-free pixels over the context frames and uses that value as a constant prediction. Performance is shown in table 1.

Autorregressive Conditional video prediction baseline The Autorregressive Conditional video prediction baseline (Arcon) is based on Stochastic adversarial video prediction (SAVP, [37]) that was originally was used as an unguided or weakly guided video prediction model. We extend SAVP for EarthNet2021 by stacking the guiding variables as extra video channels. To this end, climatic variables had to be resampled to match imagery resolution. In addition, SAVP cannot make use of the different temporal resolution of predictors and targets (daily vs. 5 daily) so predictors were reduced by taking the 5-daily mean, these steps resulted in guiding information loss. Since there is no moving objects in satellite imagery, but just a widely variable background, all SAVP components specifically designed for motion prediction were disabled. Image generation from scratch and reuse of context frames as background was enabled. Different to traditional video input data, EarthNet2021 input satellite imagery is defective, as a model shall not forecast clouds and other artifacts. Thus, different to the original implementation, we train Arcon just with mean absolute error over non-masked pixels; in particular, this means no adversarial loss was used.

Arcon outperforms the persistence baseline in every test set except the full seasonal cycle test (see table 1 for IID and OOD results), where, possibly the model breaks down when fed a context length higher than 10. The model shows degrading forecasting performance at the longer temporal horizon (see Fig. 8). These results give us two hints. First, it is necessary to overhaul and adapt current video prediction models to make them capable of tackling the strongly guided setting. Second, since the slightly adapted SAVP

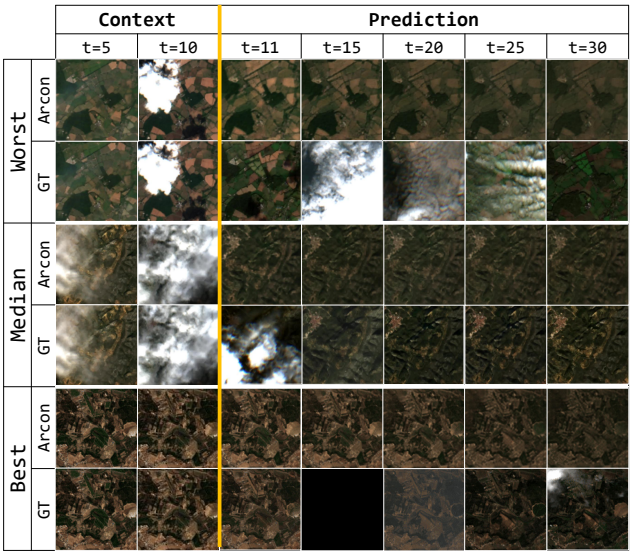

Figure 8: Worst, median and best samples predicted by Arcon according to EarthNetScore over the full IID test set. Yellow line marks the beginning of the predicted frames.

shows skill over the persistence baseline, we can anticipate current video prediction models to be a useful starting point.

Channel-U-Net baseline

This architecture is inspired by the winning solution to the 2020 Traffic4cast challenge [12]. Traffic map forecasting is to a certain degree similar to the proposed task of Earth surface forecasting. The solution used a U-Net architecture [54] with dense connections between layers. All available context time step inputs were stacked along the channel dimension and fed into the network. Subsequently the model outputs all future time steps stacked along the channel dimension, which are then reshaped for proper evaluation. We call such an approach a Channel-U-Net. Here we present a Channel-U-Net with an ImageNet [16] pre-trained DenseNet161 [30] encoder from the Segmentation Models PyTorch library[6]. As inputs we feed the center 2x2 meteorological predictors upsampled to full spatial resolution, the high-resolution DEM and all the satellite channels from the 10 context time steps, resulting in 191 input channels. The model outputs 80 channels activated with a sigmoid, corresponding to the four color channels for each of the 20 target time steps. We trained the model for 100 Epochs on a quality masked L1 loss with Adam [33], an initial learning rate of 0.002, decreased by a factor 10 after 40, 70 and 90 epochs. We use a batch size of 64 and 4 x V100 16GB GPUs. For the extreme and the seasonal tracks we slide the model through the time series by feeding back its previous outputs as new inputs after the initial prediction, which uses the last 10 frames of context available.

[6] https://smp.readthedocs.io/

Channel-U-Net is the overall best performing model, even though it does not model temporal dependencies explicitly. This model also underperforms the persistence baseline on the seasonal test set, possibly due to the sliding window approach used for the much longer prediction length.

7. Outlook

If solved, Earth surface forecasting will greatly benefit society by providing seasonal predictions of climate impacts at a local scale. These are extremely informative for implementing preventive mitigation strategies. EarthNet2021 is a stepping-stone towards the collaboration that is necessary between modelers from Computer Vision and domain experts from the Earth System Sciences. The dataset requires developing guided video prediction models, a unique opportunity for video prediction researchers to extend their approaches. Since the guided setting allows modeling in a more controlled environment, there is the possibility, that gained knowledge can also be transferred back to general (unguided) video prediction. Eventually, numerical Earth System Models [20] could benefit from the data-driven high-resolution modelling by Earth surface forecasting models. In so-called hybrid models [52], both components could be combined.

EarthNet2021 is the first dataset for spatio-temporal Earth surface forecasting. As such, it comes with a number of limitations, including some that will be discovered during model development. Through the EarthNet2021 framework, especially the model intercomparison suite, we hope to create a space for communication between different stakeholders. Then, we could remove pressing issues iteratively. We hope EarthNet2021 will serve as a starting point for community building around high-resolution Earth surface forecasting.

Acknowledgements We thank three anonymous reviewers and the area chair for their constructive reviews. We are grateful to the Centre for Information Services and High Performance Computing [Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)] TU Dresden for providing its facilities for high throughput calculations.

Authors contribution C.R. Experimental design, EarthNet model intercomparison suite, baseline persistence and Arcon models, website, documentation, manuscript. **V.B.** Experimental design, dataset creation pipeline, EarthNet toolkit package, baseline channel-net model, website, documentation, manuscript. **M.R.** Experimental design, metric selection, manuscript revision. **J.R.** Experimental design, manuscript revision. **J.D.** Experimental design, metric selection, manuscript revision.

Code and data availability All information is provided on our website www.earthnet.tech.

References

[1] Khalid A Al-Gaadi, Abdalhaleem A Hassaballa, ElKamil Tola, Ahmed G Kayad, Rangaswamy Madugundu, Bander Alblewi, and Fahad Assiri. Prediction of potato crop yield using precision agriculture techniques. *PloS one*, 11(9):e0162219, 2016.

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

[3] A Bashfield and A Keim. Continent-wide dem creation for the european union. In *34th International Symposium on Remote Sensing of Environment. The GEOSS Era: Towards Operational Environmental Monitoring. Sydney, Australia*, pages 10–15, 2011.

[4] Ana Bastos, P Ciais, P Friedlingstein, S Sitch, Julia Pongratz, L Fan, JP Wigneron, Ulrich Weber, Markus Reichstein, Z Fu, et al. Direct and seasonal legacy effects of the 2018 heat wave and drought on european ecosystem productivity. *Science advances*, 6(24):eaba2724, 2020.

[5] Joaquín Bedia, Jorge Baño-Medina, Mikel N Legasa, Maialen Iturbide, Rodrigo Manzanas, Sixto Herrera, Ana Casanueva, Daniel San-Martín, Antonio S Cofiño, and José Manuel Gutiérrez. Statistical downscaling with the downscaler package (v3. 1.0): contribution to the value intercomparison experiment. *Geoscientific Model Development*, 13(3), 2020.

[6] Vitus Benson and Alexander Ecker. Assessing out-of-domain generalization for robust building damage detection. *AI for Humanitarian Assistance and Disaster Response workshop (NeurIPS 2020)*, 2020.

[7] J Boé, L Terray, F Habets, and E Martin. A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. *Journal of Geophysical Research: Atmospheres*, 111(D23), 2006.

[8] Gilbert Brunet, Melvyn Shapiro, Brian Hoskins, Mitch Moncrieff, Randall Dole, George N Kiladis, Ben Kirtman, Andrew Lorenc, Brian Mills, Rebecca Morss, et al. Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bulletin of the American Meteorological Society*, 91(10):1397–1406, 2010.

[9] Yaping Cai, Kaiyu Guan, David Lobell, Andries B Potgieter, Shaowen Wang, Jian Peng, Tianfang Xu, Senthold Asseng, Yongguang Zhang, Liangzhi You, et al. Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches. *Agricultural and forest meteorology*, 274:144–159, 2019.

[10] Pierre Cantelaube and Jean-Michel Terres. Seasonal weather forecasts for crop yield modelling in europe. *Tellus A: Dynamic Meteorology and Oceanography*, 57(3):476–487, 2005.

[11] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838, 2020.

[12] Sungbin Choi. Utilizing unet for the future traffic map prediction task traffic4cast challenge 2020. *arXiv preprint arXiv:2012.00125*, 2020.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[14] Richard C Cornes, Gerard van der Schrier, Else JM 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.

[15] Monidipa Das and Soumya K Ghosh. Deep-step: A deep learning approach for spatiotemporal prediction of remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 13(12):1984–1988, 2016.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423, 2017.

[18] Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections, 2017.

[19] Natalia Efremova, Dmitry Zausaev, and Gleb Antipov. Prediction of soil moisture content based on satellite data and sequence-to-sequence networks. *NeurIPS 2018 Women in Machine Learning workshop*, 2019.

[20] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[21] Mathieu Fauvel, Mailys Lopes, Titouan Dubo, Justine Rivers-Moore, Pierre-Louis Frison, Nicolas Gross, and Annie Ouin. Prediction of plant diversity in grasslands using sentinel-1 and-2 satellite image time series. *Remote Sensing of Environment*, 237:111536, 2020.

[22] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *arXiv preprint arXiv:1605.07157*, 2016.

[23] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pages 3233–3246. PMLR, 2020.

[24] Feng Gao, Jeff Masek, Matt Schwaller, and Forrest Hall. On the blending of the landsat and modis surface reflectance: Predicting daily landsat surface reflectance. *IEEE Transactions on Geoscience and Remote sensing*, 44(8):2207–2218, 2006.

[25] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell. Disentangling propagation and generation for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[26] Andrea Gobbi, Marco Cristoforetti, Giuseppe Jurman, and Cesare Furlanello. High resolution forecasting of heat waves impacts on leaf area index by multiscale multitemporal deep learning. *arXiv preprint arXiv:1909.07786*, 2019.

[27] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–17, 2019.

[28] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[29] Seungkyun Hong, Seongchan Kim, Minsu Joh, and Sa-Kwang Song. Psique: Next sequence prediction of satellite images using a convolutional sequence-to-sequence network. *Workshop on Deep Learning for Physical Sciences (NeurIPS 2017)*, 2017.

[30] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[31] Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *International Conference on Machine Learning*, pages 1771–1779. PMLR, 2017.

[32] Elisa Kamir, François Waldner, and Zvi Hochman. Estimating wheat yields in australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160:124–135, 2020.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Felix N Kogan. Remote sensing of weather impacts on vegetation in non-homogeneous areas. *International Journal of remote sensing*, 11(8):1405–1419, 1990.

[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.

[36] David P Kreil, Michael K Kopp, David Jonietz, Moritz Neun, Aleksandra Gruca, Pedro Herruzo, Henry Martin, Ali Soleymani, and Sepp Hochreiter. The surprising efficiency of framing geo-spatial time series forecasting as a video prediction task–insights from the iarai traffic4cast competition at neurips 2019. In *NeurIPS 2019 Competition and Demonstration Track*, pages 232–241. PMLR, 2020.

[37] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

[38] Jae-Hyeok Lee, Sangmin S Lee, Hak Gu Kim, Sa-Kwang Song, Seongchan Kim, and Yong Man Ro. Mcsip net: Multichannel satellite image prediction via deep neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):2212–2224, 2019.

[39] Jeff Chun-Fung Lo, Zong-Liang Yang, and Roger A Pielke Sr. Assessment of three dynamical climate downscaling methods using the weather research and forecasting (wrf) model. *Journal of Geophysical Research: Atmospheres*, 113(D9), 2008.

[40] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[41] 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. Spacebooks Online, 2016.

[42] Björn Lütjens, Brandon Leshchinskiy, Christian Requena-Mesa, Farrukh Chishtie, Natalia Díaz-Rodriguez, Océane Boulais, Aaron Piña, Dava Newman, Alexander Lavin, Yarin Gal, et al. Physics-informed gans for coastal flood visualization. *arXiv preprint arXiv:2010.08103*, 2020.

[43] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[44] Jorge Baño Medina, José Manuel Gutiérrez, and Sixto Herrera García. Deep neural networks for statistical downscaling of climate change projections. In *XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018) 23-26 de octubre de 2018 Granada, España*, pages 1419–1424. Asociación Española para la Inteligencia Artificial (AEPIA), 2018.

[45] 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.

[46] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in neural information processing systems*, pages 2863–2871, 2015.

[47] Haekyung Park, Kyungmin Kim, et al. Prediction of severe drought area based on random forest: Using satellite image and topography data. *Water*, 11(4):705, 2019.

[48] Bin Peng, Kaiyu Guan, Ming Pan, and Yan Li. Benefits of seasonal climate prediction and satellite data for forecasting us maize yield. *Geophysical Research Letters*, 45(18):9662–9671, 2018.

[49] Pierre Ploton, Nicolas Barbier, Pierre Couteron, CM Antin, Narayanan Ayyappan, N Balachandran, N Barathan, J-F Bastin, G Chuyong, Gilles Dauby, et al. Toward a general tropical forest biomass prediction model from very high res-

olution optical satellite images. *Remote sensing of environment*, 200:140–153, 2017.

[50] 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.

[51] Stephan Rasp and Nils Thuerey. Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13(2):e2020MS002405, 2021.

[52] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[53] Christian Requena-Mesa, Markus Reichstein, Miguel Mahecha, Basil Kraft, and Joachim Denzler. Predicting landscapes from environmental conditions using generative networks. In *German Conference on Pattern Recognition*, pages 203–217. Springer, 2019.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[55] Marc Rußwurm and Marco Korner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.

[56] Daniel Scheffler, André Hollstein, Hannes Diedrich, Karl Segl, and Patrick Hostert. Arosics: An automated and robust open-source image co-registration software for multi-sensor satellite data. *Remote Sensing*, 9(7):676, 2017.

[57] Sebastian Scher and Gabriele Messori. Weather and climate forecasting with neural networks: using general circulation models (gcms) with different complexity as a study ground. *Geoscientific Model Development*, 12(7):2797–2809, 2019.

[58] Raí A Schwalbert, Telmo Amado, Geomar Corassa, Luan Pierre Pott, PV Vara Prasad, and Ignacio A Ciampitti. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern brazil. *Agricultural and Forest Meteorology*, 284:107886, 2020.

[59] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[60] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[61] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019.

[62] Pham Huy Thong et al. Some novel hybrid forecast methods based on picture fuzzy clustering for weather nowcasting from satellite image sequences. *Applied Intelligence*, 46(1):1–15, 2017.

[63] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[64] Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, and Auroop R Ganguly. Deepsd: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 1663–1672, 2017.

[65] Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, and Auroop R Ganguly. Generating high resolution climate change projections through single image super-resolution: An abridged version. In *International Joint Conferences on Artificial Intelligence Organization*, 2018.

[66] Mathieu Vrac, Michael Stein, and Katharine Hayhoe. Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate Research*, 34(3):169–184, 2007.

[67] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[68] Jonathan A Weyn, Dale R Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, 2020.

[69] 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.

[70] Zhan Xu, Jun Du, Jingjing Wang, Chunxiao Jiang, and Yong Ren. Satellite image prediction relying on gan and lstm neural networks. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2019.

[71] Xiaoxiang Zhu, Chunping Qiu, Jingliang Hu, Andrés Camero Unzueta, Lukas Kondmann, Aysim Toker, Giovanni Marchisio, and Laura Leal-Taixé. Dynamicearthnet challenge at earthvision 2021, http://www.classic.grss-ieee.org/earthvision2021/challenge.html.

[72] Yuejian Zhu. Ensemble forecast: A new approach to uncertainty and predictability. *Advances in atmospheric sciences*, 22(6):781–788, 2005.

[73] Zhe Zhu, Curtis E Woodcock, Christopher Holden, and Zhiqiang Yang. Generating synthetic landsat images based on all available landsat data: Predicting landsat surface reflectance at any given time. *Remote Sensing of Environment*, 162:67–83, 2015.