Understanding the Role of Weather Data for Earth Surface Forecasting using a ConvLSTM-based Model

Codruț-Andrei Diaconu[1], Sudipan Saha[2], Stephan Günnemann[2], Xiao Xiang Zhu[1,2]
[1]German Aerospace Center (DLR) [2]Technical University of Munich
{codrut-andrei.diaconu, xiaoxiang.zhu}@dlr.de, sudipan.saha@tum.de, guennemann@in.tum.de

Abstract

Climate change is perhaps the biggest single threat to humankind and the environment, as it severely impacts our terrestrial surface, home to most of the living species. Inspired by video prediction and exploiting the availability of Copernicus Sentinel-2 images, recent studies have attempted to forecast the land surface evolution as a function of past land surface evolution, elevation, and weather. Further extending this paradigm, we propose a model based on convolutional long short-term memory (ConvLSTM) that is computationally efficient (lightweight), however obtains superior results to the previous baselines. By introducing a ConvLSTM-based architecture to this problem, we can not only ingest the heterogeneous data sources (Sentinel-2 time-series, weather data, and a Digital Elevation Model (DEM)) but also explicitly condition the future predictions on the weather. Our experiments confirm the importance of weather parameters in understanding the land cover dynamics and show that weather maps are significantly more important than the DEM in this task. Furthermore, we perform generative simulations to investigate how varying a single weather parameter can alter the evolution of the land surface. All studies are performed using the EarthNet2021 dataset. The code, additional materials and results can be found at https://github.com/dcodrut/weather2land.

1. Introduction

Climate change has recently emerged as an important area of research as the Earth's climate is in a significant transition. This poses a serious threat to our existence since it has a pronounced and complex impact on our terrestrial surface that hosts the majority of our living world [2]. Consequences are already prominent, starting from sea ice melting [14] to increased fire events [39, 41]. Remarkably, some impacts are different in different areas of the world and both global and regional impacts are yet to be fully understood [1].

A better understanding of the regional impacts of the different weather scenarios can be useful for many downstream tasks, *e.g.* estimating the crop yield [25, 29], and this becomes even more critical when droughts are expected [4]. The reason is that the weather has a heterogeneous impact, depending on many local factors (*e.g.* vegetation, soil, or topography) [18, 28]. Moreover, due to the computational limits of the numerical models, seasonal forecasts are provided at a relatively low spatial resolution, up to $0.25°$ (*i.e.* $\approx$27 km on the Equator) [12, 23]. This motivates the use of additional inputs to be able to downscale these predictions to a resolution that can allow for downstream analyses.

The last decade has seen a significant increase in the number of satellite sensors, thus making Earth observation data available at an unprecedented scale. The Copernicus program of the European Space Agency provides up to 10m/pixel (Sentinel-2) data at a high temporal resolution of 5 days. Such increased temporal availability has made possible dense predictions/analyses at a reasonable spatial resolution [19, 36] that were not previously feasible. As such, satellite images are a good candidate for being additional input to the models analyzing the regional impact of climate change. Towards predicting the evolution of terrestrial land surface, a promising approach is to combine the cues provided by dense satellite time-series images and the weather data.

Following this direction, Requena-Mesa *et al.* [28] formulated the Earth surface forecasting task as a video prediction of the satellite imagery guided by mesoscale weather projections and released a dataset (called EarthNet2021) for supporting this task. The aim is to build a model that uses a context of 10 satellite images and infers the following 20, given weather information available for both the context frames and the future ones. Additionally, a static high-resolution Digital Elevation Model (DEM) is also used. These additional inputs should guide the predictions of the model, thus the similarity to a guided video-prediction problem. Three different baseline models are proposed in [28], including a naive model which simply makes a constant prediction computed as the average of

the context frames. The best-performing model among the three baselines uses U-Net [30] and it achieves the highest evaluation scores even though it does not explicitly model the temporal dependence. Moreover, the performance gap between U-Net and the naive model is small, which suggests that improvements could be potentially achieved by exploiting architectures that suitably incorporate the temporal context. This motivates us to use the Convolutional LSTM architecture which is known for its capability to learn effective spatio-temporal features [35].

Requena-Mesa *et al*. [28] do not provide any detailed insights regarding the relevance of the additionally provided inputs, *i.e*. the meteorological projections and the DEM, w.r.t. to the target variable, *i.e*. the land surface. We analyze the performance improvement as an indicator that these additional parameters are indeed valuable for land surface forecasting. Additionally, a generative approach can be particularly useful here, to visualize and track the changes in the predicted land surface w.r.t. varying weather parameters. Towards this, we make a detailed investigation under different rainfall scenarios.

Our contributions can be summarized as follows:

- We propose a new model which, despite using a much smaller number of parameters, achieves a significantly better performance than the baseline models.
- We study the importance of the DEM and the weather variables through an ablation study.
- To further validate our model, we evaluate it under various simulated rainfall scenarios which can also serve as an example of an interesting practical use-case of such a land surface prediction model.

2. Related Work

In this section, we first discuss existing works that jointly use weather and remote sensing time-series data. Then, considering that ConvLSTM is the primary component of the proposed method, we briefly discuss some of its existing applications in remote sensing time series analysis.

2.1. Weather Data and Remote Sensing Time-Series

The problem of weather prediction remains an important challenge for both classical modeling approaches and pure data-driven models built using ML [3, 32]. A recent data-driven approach has proven to be very successful in precipitation nowcasting (*i.e*. predicting the precipitation probabilities up to 90 min lead time), achieving state-of-the-art results in comparison to both physics-based methods and other ML approaches [26].

Analyzing time-series optical data from the satellites can be useful in many fields, a prominent one being agriculture where ML/DL can play an important role given the large amount of available data and its complexity [13]. Classical examples are crop yield prediction [7] and crop type classification [31]. For this type of problems, the models can also benefit from including weather-related information as input given that crop fields are affected by it. In previous works, the weather is used as input to directly infer the desired outcome, *e.g*. the crop yield [15, 33]. An alternative to directly predicting the desired outcome is to first forecast the evolution of the Earth surface (discussed in detail in Section 3.1) and then use the resulting predictions in downstream tasks.

2.2. ConvLSTM in Remote Sensing Time-Series

ConvLSTM was first used in [35] for precipitation nowcasting. Since then, it has been used in many works related to remote sensing time-series analysis. Shen *et al*. [34] use ConvLSTM for semi-supervised time-series land cover classification. Moskolai *et al*. [24] compare different variants of LSTM for next-frame forecasting in the Sentinel-1 time series, although ConvLSTM did not perform well when the length of the sequence became higher. CNN and ConvLSTM are combined in [6] for spatio-temporal feature extraction. In [31] a bi-directional ConvLSTM model is employed for land cover classification.

3. EarthNet2021 - Dataset and Challenge

Next, we will provide some details regarding EarthNet2021, which forms the base of our work. We start by elaborating the idea of directly forecasting the Earth surface and its advantages. Then we describe the components of the dataset, followed by the evaluation sets and scores and in the end, we provide a summary of the three proposed baseline models.

3.1. Earth Surface Forecasting

Predicting the Earth surface was recently framed as a guided-video prediction task as follows: infer the future satellite imagery conditioned on the past and also on the weather for the entire period [28]. In other words, for predicting the next h frames, a model uses the context of c frames and the corresponding weather conditions and also guides its future predictions based on a given weather scenario, as depicted in Eq. (1).

$$X_{c+1,\ ...,\ c+h} = F(X_{1,\ ...,\ c}, W_{1,\ ...\ ,c+h}) \tag{1}$$

where X_t and W_t are the satellite imagery and the meteorological conditions at time t, respectively; c and h are the context and the horizon lengths, respectively, and F denotes a model mapping.

Framing the problem in this manner offers some advantages. First, the resulting predicted imagery can be later used for various downstream tasks (*e.g*. predicting vegetation indices, crop yield *etc*.) instead of developing a model

independently for each target. Second, the training in this scenario is self-supervised, therefore a large quantity of satellite data is available since no labels are required. However, dealing with such large-scale datasets can also be very challenging as in general it requires more computational resources and automatic quality control. A challenge platform associated with this task was also released[1].

3.2. Dataset Description

The EarthNet2021 dataset contains approximately 32000 samples, each of them consisting of a sequence of 30 Sentinel-2 images, with a temporal resolution of 5 days. Each image has four bands (red, green, blue, and near-infrared) with a pixel resolution of 128x128px and a 20m ground resolution. Additionally, weather-related variables are provided: precipitation, sea level pressure, and temperature (min, max and mean). These variables are obtained from the observational dataset E-OBS [5] as an alternative to actual seasonal meteorological predictions for the following reasons: first, it is computationally expensive to obtain these forecasts for multiple starting points and, second, using observation-based data provides an ideal testbed for the land surface prediction task [28].

To facilitate the evaluation (and potentially the training), binary quality masks were also created with the same resolution as the imagery. These masks indicate whether a pixel is covered by clouds, shadows or the value is missing. They are used in the evaluation procedure in order to ignore these areas when computing the scores since it is desired that the models produce clean, cloud-free, images. According to these quality masks, approximately 40% of the pixels from this dataset are masked out, which illustrates the difficulty of the task.

3.3. Evaluation Sets

There are four evaluation tracks, denoted by **Main (IID)**, **Robustness (OOD)**, **Extreme summer**, and **Seasonal cycle**. The IID set contains ≈4000 samples from the same regions as the training set, where one region corresponds to a Sentinel-2 tile, *i.e.* ≈100x100km. However, if two samples capture exactly the same area, then it was ensured that there is no temporal overlapping between them. The OOD set contains a similar amount of samples but the regions are completely different, therefore it additionally evaluates the spatial generalization capability of the model. For these two tracks, the context and prediction lengths are 10 and 20, respectively. The Extreme set contains only samples from summer 2018 in northern Germany, a region that faced an extreme heat in that period. In this case, 20 frames (from February to the end of May) are provided as input and the task is to predict the following six months. Last, the seasonal track includes samples with much longer time periods, *i.e.* one year as context and two as target, with the aim of capturing the entire vegetation cycle.

[1] www.earthnet.tech/

3.4. Evaluation Scores

The final evaluation metric, called EarthNetScore (ENS), is the harmonic mean of four different evaluation metrics, calculated only on the non-masked pixels.

- **Median Absolute Deviation (MAD) score**. It computes the median absolution deviation between target pixels and the predicted ones to simply quantify how close are they in a robust manner.
- **Ordinary Least Square (OLS) score**. It measures if the trend in the vegetation is correctly captured in the predictions. First, NDVI maps are computed for both the target and predicted series, then OLS models are fitted over time for each pixel and finally, the slopes are measured and compared to get the OLS score.
- **Earth Mover Distance (EMD) score**. Similar to the previous score but focused on the distribution of the pixels. It computes the Wasserstein-1 pixelwise distance between the target and the predicted NDVIs.
- **Structural Similarity Index (SSIM) score**. It captures the perceptual similarity by computing the average SSIM [38] over channel and timestep.

Each of these intermediate scores is non-linearly scaled in [0, 1], where 1 corresponds to a perfect prediction, in order to make them comparable and to be able to compute an average score based on them.

3.5. Baselines

Three baseline models proposed in [28] are briefly described below, omitting some of the training details.

- **Persistence**. It is a naive model that computes the pixel-wise average of the context frames and uses it as a constant prediction for any lead time.
- **Arcon**. This model is based on the Stochastic Adversarial Video Prediction (SAVP) [22]. Originally, the model combines latent variables and an adversarial loss for predicting high-quality images. However, to avoid predicting cloud-covered images, the authors disabled the adversarial loss and used a masked L1 loss instead.
- **Channel-U-Net**. This baseline used the U-Net [30] architecture. It stacks over channel all the input information, *i.e.* the RGBNIR bands, the meteorological data and the DEM, from all available timesteps, resulting in a 191-dimensional input. The model is then trained to predict a map with 80 channels which is reshaped to produce four-dimensional images, one for each of the 20 horizon steps.

4. Proposed Method

Convolutional LSTMs. The Long Short-Term Memory networks have already a longstanding success in modeling sequential data. Using the memory cells lead to a better gradient flow thus addressing the problem of vanishing gradients encountered in the standard RNNs and also allowing to learn long-range temporal dependencies [11, 37]. By stacking multiple LSTM units, many powerful architectures were built for addressing real-world problems like machine translation [37], handwriting recognition [8] and speech recognition [9] among many others [10].

Similarly, Convolutional Neural Networks also became a standard component for building architectures suitable for image feature extraction due to their trainable filters, with real-world applications ranging from handwritten digits classification [21] to biomedical segmentation [30]. A survey of many well-established architectures and their applications is provided in [16].

The Convolutional LSTM combines both the advantages of the CNNs as powerful image feature extractors and those of the LSTMs that have the capability to learn temporal correlations. The architecture was proposed as a method for learning spatiotemporal features in image time-series data and applied for the problem of precipitation nowcasting [35]. It has been shown that ConvLSTMs perform better than the standard fully-connected version when training on spatiotemporal data and it also leads to fewer parameters when using deep models.

The information flow through a single ConvLSTM unit at time t is controlled by three internal gates, the input gate i_t, the forget gate f_t and the output gate o_t, each with its own weight matrices. They combine the new input X_t with the information stored in the previous hidden state H_{t-1}, as described in Eq. (2).

$$\begin{aligned} i_t &= \sigma(W_{ix} * X_t + W_{ih} * H_{t-1}) \\ f_t &= \sigma(W_{fx} * X_t + W_{fh} * H_{t-1}) \\ o_t &= \sigma(W_{ox} * X_t + W_{oh} * H_{t-1}) \end{aligned} \tag{2}$$

where $*$ denotes the convolution operator and σ the sigmoid function. Using these gates, the new long-term cell state C_t and the hidden state H_t are updated according to Eq. (3).

$$\begin{aligned} C_t &= f_t \cdot C_{t-1} + i_t \cdot tanh(W_{cx} * X_t + W_{ch} * H_{t-1}) \\ H_t &= o_t \cdot tanh(C_t) \end{aligned} \tag{3}$$

where $\cdot$ denotes the Hadamard product. Hence C_t is computed based on parts of the new input X_t and the output from the previous step H_{t-1}, controlled by the input gate, in combination with its previous values C_{t-1} which are not cleared by the forget gate. Finally, the hidden state H_t is updated based on the new cell state filtered by the output gate.

In the context of EarthNet2021, we employ a ConvLSTM-based model for multiple reasons. First, being a RNN, it naturally fits temporal data, with a strong inductive bias. Second, the imagery captured from the satellite can be partially or completely affected by clouds (or their shadows) and as a result, many frames do not contain any useful information. Therefore the model should learn to ignore these areas when iterating over the context frames and the ConvLSTMs are a suitable candidate due to their gating mechanism. In this direction, in [31] it has been shown that such a model can automatically learn to internally filter the clouds without any supervision thus avoiding the need of pre-processing the data. Third, when lopping over the input frames we can directly provide as input the current weather information which should guide the next predictions. In this manner we explicitly constraint the model to learn the temporal evolution of the surface conditioned on the weather by exploiting the recurrent inductive bias. Last, the model allows making inferences on a horizon of variable lengths without additional changes.

Training procedure. In order to train the model using the RGBNIR frames, the weather conditions, and the DEM as input, we use the following strategy: the weather variables are averaged over periods of five days and then upscaled to 128x128 to match both the temporal and spatial resolution of the satellite imagery. Then the DEM is attached to each frame. All these inputs are stacked over channel resulting in a large frame with a size of 128x128x10. At time t ($1 \leq t \leq c + h$), the model takes one such combined frame as input and outputs the next RGBNIR frame. Since we have only a context of c RGBNIR frames available, for $t > c$ we use the previous predicted RGBNIR frame when stacking the inputs. The procedure is illustrated in Fig. 1.

Architecture. We use four stacked ConvLSTM layers. The first layer has 10 input channels and 32 output channels. The next two hidden layers have both input and output sizes of 32. The last layer has an input of 32 and an output of 4 which to match the dimensionality of the target imagery. For all layers the kernel size is fixed to 3x3 with padding of one pixel, to preserve the dimensionality of the intermediate states. The resulting total number of parameters is around 200k which makes the model relatively small.

5. Experiments

In the following, we will present our experiments and discuss the results. We begin with comparing our model to the baselines from [28], followed by an ablation study to evaluate the importance of the input variables. Finally, through a simulation setup, we investigate if the model

	IID					OOD				
	ENS	**MAD**	**OLS**	**EMD**	**SSIM**	**ENS**	**MAD**	**OLS**	**EMD**	**SSIM**
Persistance (baseline-1)	0.2625	0.2315	0.3239	0.2099	0.3265	0.2587	0.2248	0.3236	0.2123	0.3112
Channel-U-Net (baseline-2)	0.2902	0.2482	0.3381	0.2336	0.3973	0.2854	0.2402	0.3390	0.2371	0.3721
Arcon (baseline-3)	0.2803	0.2414	0.3216	0.2258	0.3863	0.2655	0.2314	0.3088	0.2177	0.3432
ConvLSTM	**0.3266**	**0.2638**	**0.3513**	**0.2623**	**0.5565**	**0.3204**	**0.2541**	**0.3522**	**0.2660**	**0.5125**

Table 1. Comparison of our model with the baseline models reported in [28] on the IID and OOD test sets. For our model we report the average scores over five runs with different random initializations.

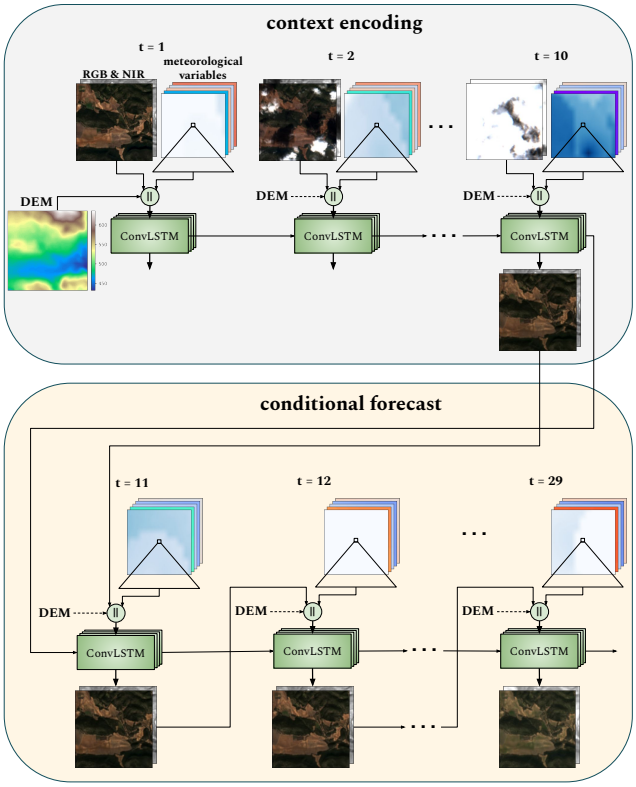

Figure 1. ConvLSTM training procedure: the 10 four-bands context images are encoded together with the additional inputs (i.e. the five meteorological inputs, which are cropped and then upscaled, and the DEM, which is repeatedly added as input). Based on the encoded context, the next 20 images are predicted one by one, also conditioned on the two provided inputs.

learned the weather → land surface relationship which can also serve as an important practical use-case of such a model.

5.1. Comparison Against the Baselines

In the first experiment, we compare our model to the baseline models proposed in [28]. For this, we use as input the RGBNIR frames, all the weather variables, and the DEM. We retrain the model five times with different random initializations and evaluate them on both the IID and OOD test sets. We report the average scores using all proposed metrics in Tab. 1.

The results show improved performance on all the evaluation metrics when compared to all three baselines, with a gap of approximately 0.35 units between our model and the best baseline (Channel-U-Net). The ENS score difference between the IID and OOD is ≈ 0.06, similar to the U-Net-based model (≈ 0.05). We can also notice that the SSIM score has the largest improvement although it is hard to compare the differences due to the non-linear scaling of each individual score. Similar to [28] we show three predictions, together with the corresponding NDVIs, in Fig. 2.

5.2. Variable Importance Analysis

The core idea behind the EarthNet2021 task is that a model should guide its predictions based on a given weather scenario. To validate if our model learns to extract information from the weather conditions, we perform an ablation study by training three types of models in an increasing data complexity order: first using only the optical imagery, then adding the DEM, and lastly including the weather variables. For each of these scenarios, we evaluate the models on the same datasets as in the previous experiment. The results are shown in Tab. 2.

The first conclusion of this experiment is that the model benefits more from using the weather information and less from the DEM. This supports the idea behind the proposed task that weather information should be taken into account when predicting the evolution of the Earth landscapes. Another important aspect is that the standard deviation is relatively low showing that the model is stable w.r.t. its initialization. Analyzing the standard deviations also shows that the model remains stable when including additional data modalities. Last, one can also note that our model already achieves better results only with the optical imagery as input (see the baselines in Tab. 1) suggesting that it is better suited for this particular task.

5.3. Simulations

While the quantitative results already show that the land surface evolution is dependent on the weather, in this Section we further investigate this aspect using a generative approach. For this, we chose one of the weather variables (*i.e.*

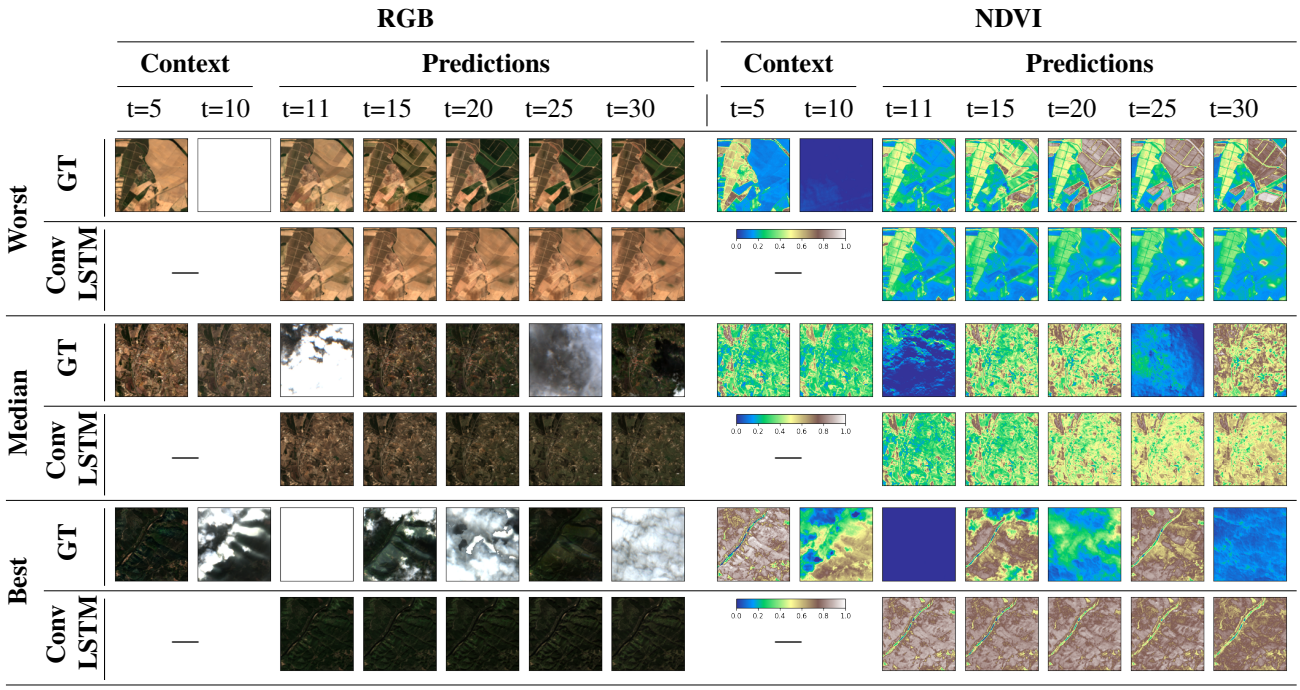

Figure 2. RGB and NDVI predictions for three samples (worst, median and best according to EarthNetScore over the IID test set, as in [28])

Test set	Input data	ENS	MAD	OLS	EMD	SSIM
IID	**RGBNIR**	0.3151 ± 0.0004	0.2576 ± 0.0002	0.3424 ± 0.0004	0.2530 ± 0.0005	0.5162 ± 0.0015
	RGBNIR + DEM	0.3156 ± 0.0003	0.2579 ± 0.0001	0.3424 ± 0.0005	0.2533 ± 0.0006	0.5183 ± 0.0009
	RGBNIR + WEATHER + DEM	0.3266 ± 0.0004	0.2638 ± 0.0002	0.3513 ± 0.0001	0.2623 ± 0.0004	0.5565 ± 0.0017
OOD	**RGBNIR**	0.3078 ± 0.0005	0.2484 ± 0.0001	0.3426 ± 0.0008	0.2547 ± 0.0007	0.4709 ± 0.0016
	RGBNIR + DEM	0.3084 ± 0.0004	0.2482 ± 0.0003	0.3433 ± 0.0008	0.2564 ± 0.0009	0.4703 ± 0.0019
	RGBNIR + WEATHER + DEM	0.3204 ± 0.0002	0.2541 ± 0.0002	0.3522 ± 0.0006	0.2660 ± 0.0004	0.5125 ± 0.0010

Table 2. Ablation study for investigating the importance of the additional inputs (DEM and the weather variables). For each evaluation score we report the average score and the standard deviation over five runs with different random initializations, on both the IID and OOD test sets. Note that the last line from each test set corresponds to the previously reported averaged results from Tab. 1.

rainfall) and artificially increased/decreased the actual values randomly such that on average a certain difference is imposed, if possible (the lower limit for rainfall is zero). All the other variables are kept the same. The idea is illustrated in Fig. 3 where for one sample we show two artificially generated rainfall scenarios (one with less amount of rainfall and one with an increased amount) together with the original one. We can observe that some of the crops become more healthy (as measured by the NDVIs) for an increased rainfall. Second, the model has an integrative effect relative to the vegetation greenness: the differences between the last frames (*i.e.* for t=30) under the three scenarios are much larger than those on earlier steps. This can be explained by the fact that a vegetation area needs a certain period of time until it uses the precipitation accumulated in the soil and, assuming that the amount of rainfall remains high and within a healthy regime, we would expect that the vegetation greenness increases over time.

To validate this quantitatively, we follow the same strategy for the entire IID and OOD test sets and then evaluate the predictions under all scenarios using the same evaluation scores as in the previous experiments. The results are included in Tab. 3. As expected, the best performance is achieved when the actual rainfall scenario is used and decreases with the increasing absolute change. This suggests again that knowing the weather can help the model to guide its prediction.

Simulation use-cases. We would like to further emphasize the idea of using such a model to perform simulations. As mentioned in Sec. 3.2, EarthNet2021 relies on weather measurements obtained from E-OBS [5] which is an observational dataset. However, in practice, we would have to make use of actual forecasts for the meteorologi-

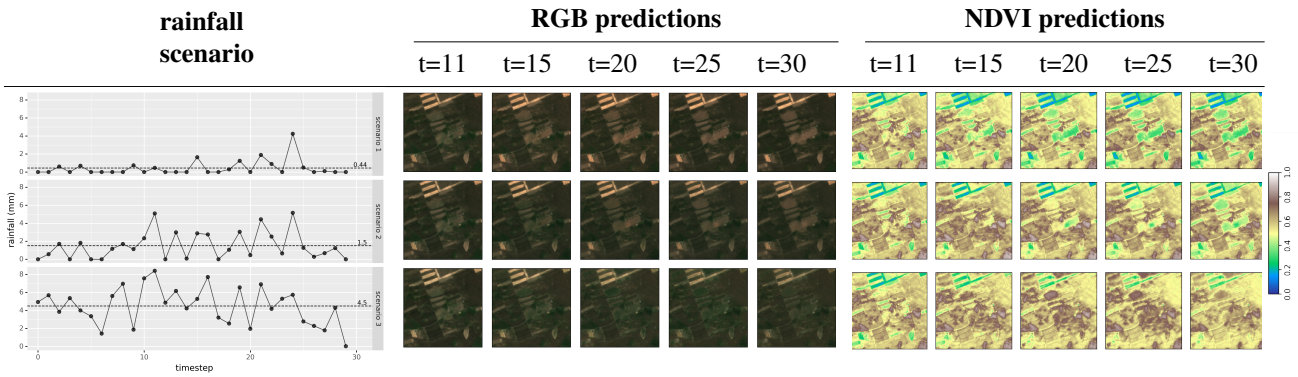

Figure 3. RGB and NDVI predictions under three different rainfall scenarios for a subset of future steps. The row in the middle corresponds to the actual scenario. The average rainfall (in mm) is shown on top of the dash line. All other weather conditions (i.e. sea level temperatures and temperature) are kept the same.

average daily rainfall change (mm)	IID					OOD				
	ENS	**MAD**	**OLS**	**EMD**	**SSIM**	**ENS**	**MAD**	**OLS**	**EMD**	**SSIM**
-0.8	0.3187	0.2591	0.3449	0.2564	0.5286	0.3130	0.2493	0.3479	0.2610	0.4848
-0.4	0.3244	0.2624	0.3498	0.2606	0.5482	0.3181	0.2523	0.3519	0.2646	0.5025
+0.0	**0.3262**	**0.2637**	**0.3512**	**0.2617**	**0.5547**	**0.3203**	**0.2539**	**0.3530**	**0.2659**	**0.5110**
+1.0	0.3163	0.2596	0.3404	0.2522	0.5294	0.3054	0.2476	0.3364	0.2517	0.4727
+2.0	0.3062	0.2558	0.3307	0.2422	0.5001	0.2896	0.2433	0.3183	0.2344	0.4363
+3.0	0.2988	0.2528	0.3247	0.2353	0.4764	0.2807	0.2408	0.3087	0.2252	0.4133

Table 3. Influence of five artificially generated rainfall scenarios on the evaluation scores, using a single model. The first column shows the average difference (over the entire dataset) between the original values and the perturbed ones. The row in bold corresponds to the actual scenario.

cal conditions. Given that at the moment long-term weather predictions are almost impossible to obtain [27], one could use such a model for performing multiple simulations as previously shown in our experiments. For instance, we can build the worst and best-case scenarios, based on the observations from the previous decades or by transferring a certain scenario from another region with similar characteristics to the one of interest. Furthermore, this can also reveal which segments of the surface react more to a specific change. Last, using this kind of counterfactual experiments we can also study the influence of each of the weather variables onto the land surface. As an example, this can help in estimating the vegetation time-lag effects on meteorological conditions which can provide a better understanding of the vegetation dynamics [40].

Weather and seasonal forecasts. Although we cannot obtain weather predictions for a long horizon, we could instead rely on seasonal forecasts in the context of the land surface prediction task. However, these have a much lower temporal resolution, usually monthly, but they still have many sources of uncertainty, *e.g*. measurement errors in the initial conditions or poorly modeled processes [12]. The predictions are usually provided by an ensemble of models with different initializations. Following the idea from [4] we can propagate these forecasts through a land surface prediction model and then analyze the spread in the projections which can provide additional insights.

Training Details

Throughout all our experiments, we trained our models for 60 epochs with a masked L1 loss, using Adam [17], a batch size of 32 and an initial learning rate of 0.001 halved at epochs 10, 20, and 50. The best model is saved based on the performance on a small validation set, *i.e*. 1% of the training samples.

Additional Results

The model development and all the experiments previously described are based on the case when the model receives a context of 10 frames and predicts the following 20, under normal weather conditions, *i.e*. similar to those encountered in the training set.

We additionally evaluated our model on the two special use-cases, *i.e*. one with extreme weather conditions and another one for seasonal evaluation, for which also the context

	Extreme					Seasonal				
	ENS	**MAD**	**OLS**	**EMD**	**SSIM**	**ENS**	**MAD**	**OLS**	**EMD**	**SSIM**
Persistance (baseline-1)	0.1939	0.2158	0.2806	0.1614	0.1605	**0.2676**	**0.2329**	**0.3848**	**0.2034**	**0.3184**
Channel-U-Net (baseline-2)	**0.2364**	**0.2286**	**0.2973**	**0.2065**	**0.2306**	0.1955	0.2169	0.3811	0.1903	0.1255
Arcon (baseline-3)	0.2215	0.2243	0.2753	0.1975	0.2084	0.1587	0.2014	0.3788	0.1787	0.0834
ConvLSTM	0.2140	0.2137	0.2906	0.1879	0.1904	0.2193	0.2146	0.3778	0.2003	0.1685

Table 4. Comparison of our model with the baseline models reported in [28] on the Seasonal and Extreme test sets. For our model we report the average scores over five runs with different random initializations.

and prediction lengths are different (see Sec. 3.3). We report the results in Tab. 4. The Channel-U-Net still performs the best on the extreme set, whereas, for the seasonal set, our model outperforms the Channel-U-Net and Arcon models. However, for the latter evaluation set, the performance of all ML models is still lower than the naive model. Overall, the results show that these testing scenarios would require to design special architectures to address their particularities. As an example, we observed that predictions towards the end of the horizon tend to be more blurred, especially for regions with high spatial variations (e.g. crop fields), therefore this can impact the performance when extending the horizon to much larger values. Coupling the model with an attention mechanism may alleviate the problem [20].

6. Conclusions

In this paper, we propose a ConvLSTM-based architecture for the task of land surface prediction conditioned on elevation and observational meteorological data. Taking advantage of the recurrent inductive bias of the LSTMs, we can explicitly condition future predictions on the given weather information. With a relatively small number of parameters, our model achieves significantly better performance compared to the previous methods on both data drawn from the same spatial distribution as the training data (IID) and data sampled from different regions (OOD), under normal weather conditions. However, a more adapted architecture would be needed to model special situations like extreme weather conditions or long-term seasonal variations.

Other than finding a better model for the task, a major focus of our work was on investigating if the model learned the relationship between the weather conditions and the evolution of the Earth surface. First, we perform an ablation study and show that the model benefits from conditioning its prediction on the meteorological data whereas adding the elevation information did not bring any performance improvement. Furthermore, we build a simulation setup to visually and quantitatively analyze how the predictions change under various perturbations of the rainfall, showing that the model performs best when the actual conditions are provided. This experiment also provides an example of how such a model could be used in practice for generating multiple scenarios or for a better understanding of the interaction between the weather and the land surface. In the end, we also discuss how to make use of seasonal forecasts in this context.

Future work Our model provides a single prediction irrespective of the quality of the context frames or the uncertainty of the weather inputs. An important extension would be to further modify the model to deal with uncertainty in the input data. Towards this, a first step would be to provide multiple predictions which should capture these sources of uncertainty. First, we should expect a high predictive uncertainty when most of the context frames are either missing or completely covered by clouds. Second, the uncertainty should also grow with the horizon length. And last, the model should integrate the uncertainty coming from the seasonal forecasts as discussed in Sec. 5.3.

Another question that deserves further investigation is how to deal with the unexpected changes in the evolution of the Earth surface due to external interventions. An example of such a sudden change is crop harvesting. Such disruptions can hinder the learning process and may also lead to erroneous evaluation scores, for instance those that measure the quality of the learned vegetation trends since the underlying assumptions are not satisfied anymore. This issue can potentially be addressed by introducing a specially-designed component in the model that can detect and treat these discontinuities accordingly.

Finally, we would like to further investigate possible extension of the model to achieve better performance when applied to special cases like those with extreme weather conditions for which also the context and horizon lengths can differ significantly.

Acknowledgements. This work is jointly support by Helmholtz Association in the framework of Munich Data Science Research School (MUDS), by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by German Federal Ministry of Economics and Technology in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C).

References

[1] Nigel W Arnell, Jason A Lowe, Andrew J Challinor, and Timothy J Osborn. Global and regional impacts of climate change at different levels of global temperature increase. *Climatic Change*, 155(3):377–391, 2019. 1

[2] Yinon M Bar-On, Rob Phillips, and Ron Milo. The biomass distribution on earth. *Proceedings of the National Academy of Sciences*, 115(25):6506–6511, 2018. 1

[3] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015. 2

[4] 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. 1, 7

[5] 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. 3, 6

[6] Gael Kamdem De Teyou. Convlstm for spatio-temporal feature extraction in time-series images. *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*, 2020. 2

[7] Diego Gómez, Pablo Salvador, Julia Sanz, and Jose Luis Casanova. Potato yield prediction using machine learning techniques and sentinel 2 data. *Remote Sensing*, 11(15):1745, 2019. 2

[8] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2008. 4

[9] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013. 4

[10] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016. 4

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[12] Stephanie J Johnson, Timothy N Stockdale, Laura Ferranti, Magdalena A Balmaseda, Franco Molteni, Linus Magnusson, Steffen Tietsche, Damien Decremer, Antje Weisheimer, Gianpaolo Balsamo, et al. Seas5: the new ecmwf seasonal forecast system. *Geoscientific Model Development*, 12(3):1087–1117, 2019. 1, 7

[13] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018. 2

[14] Vladimir M Kattsov, Vladimir E Ryabinin, James E Overland, Mark C Serreze, Martin Visbeck, John E Walsh, Walt Meier, and Xiangdong Zhang. Arctic sea-ice change: a grand challenge of climate science. *Journal of Glaciology*, 56(200):1115–1121, 2010. 1

[15] ZH Khalil and SM Abdullaev. Neural network for grain yield predicting based multispectral satellite imagery: comparative study. *Procedia Computer Science*, 186:269–278, 2021. 2

[16] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53(8):5455–5516, 2020. 4

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[18] Felix N Kogan. Remote sensing of weather impacts on vegetation in non-homogeneous areas. *International Journal of remote sensing*, 11(8):1405–1419, 1990. 1

[19] Katja Kowalski, Cornelius Senf, Patrick Hostert, and Dirk Pflugmacher. Characterizing spring phenology of temperate broadleaf forests using landsat and sentinel-2 time series. *International Journal of Applied Earth Observation and Geoinformation*, 92:102172, 2020. 1

[20] Bernard Lange, Masha Itkina, and Mykel J. Kochenderfer. Attention augmented ConvLSTM for environment prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1346–1353. IEEE, 2020. 8

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4

[22] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 3

[23] C MacLachlan, Alberto Arribas, K Andrew Peterson, A Maidens, D Fereday, AA Scaife, M Gordon, M Vellinga, A Williams, RE Comer, et al. Global seasonal forecast system version 5 (glosea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, 141(689):1072–1084, 2015. 1

[24] Waytehad Moskolaï, Wahabou Abdou, Albert Dipanda, and Dina Taiwe Kolyang. Application of lstm architectures for next frame forecasting in sentinel-1 images time series. *arXiv preprint arXiv:2009.00841*, 2020. 2

[25] 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. 1

[26] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021. 2

[27] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. 7

[28] Christian Requena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim Denzler. Earthnet2021: A

large-scale dataset and challenge for earth surface forecasting as a guided video prediction task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1132–1142, 2021. 1, 2, 3, 4, 5, 6, 8

[29] Michael J Roberts, Wolfram Schlenker, and Jonathan Eyer. Agronomic weather measures in econometric models of crop yield with implications for climate change. *American Journal of Agricultural Economics*, 95(2):236–243, 2013. 1

[30] 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. 2, 3, 4

[31] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4):129, 2018. 2, 4

[32] MG Schultz, Clara Betancourt, Bing Gong, Felix Kleinert, Michael Langguth, LH Leufen, Amirpasha Mozaffari, and Scarlet Stadtler. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194):20200097, 2021. 2

[33] 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. 2

[34] Jing Shen, Chao Tao, Ji Qi, and Hao Wang. Semi-supervised convolutional long short-term memory neural networks for time series land cover classification. *Remote Sensing*, 13(17):3504, 2021. 2

[35] 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. *Advances in Neural Information Processing Systems*, 28, 2015. 2, 4

[36] Yady Tatiana Solano-Correa, Francesca Bovolo, Lorenzo Bruzzone, and Diego Fernández-Prieto. A method for the analysis of small crop fields in sentinel-2 dense time series. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):2150–2164, 2019. 1

[37] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 4

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[39] B Mike Wotton, Charles A Nock, and Mike D Flannigan. Forest fire occurrence and climate change in canada. *International Journal of Wildland Fire*, 19(3):253–271, 2010. 1

[40] Donghai Wu, Xiang Zhao, Shunlin Liang, Tao Zhou, Kaicheng Huang, Bijian Tang, and Wenqian Zhao. Time-lag effects of global vegetation responses to climate change. *Global change biology*, 21(9):3520–3531, 2015. 7

[41] Massimo Zanetti, Sudipan Saha, Daniele Marinelli, Maria Lucia Magliozzi, Massimo Zavagli, Mario Costantini, Francesca Bovolo, and Lorenzo Bruzzone. A system for burned area detection on multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 1