An analysis of landscape predictions from a convolutional LSTM trained on the EarthNet2021 dataset

Jeroen Smets

Thesis submitted for the degree of
Master of Science in Artificial
Intelligence, option Engineering and
Computer Science

Thesis supervisor:
Prof. dr. Stef Lhermitte

Assessor:
Renaud Detry

Academic year 2022 – 2023

Preface

I would like to thank Professor Stef Lhermitte for suggesting this interesting topic and for his continuous advice. I very much appreciate his excellent support in writing this Master's thesis.

Jeroen Smets

Contents

Abstract

Seasonal forecasts of the Earth's landscape and vegetation can be very valuable information to overcome modern-day problems such as climate change, hunger, and natural disasters. Satellite images, which are frames of the Earth's surface, can provide a lot of information for making such seasonal forecasts. Moreover, it is highly available data accessible over an extended range of 8 years back in time. Having large-scale and long-term historical data simplifies the process of training and testing artificial neural networks for landscape predictions. Forecasting those images, with the help of additional parameters, can help predict the future conditions of the Earth's surface. EarthNet2021 is an open-source challenge that provides a large dataset suitable for training deep neural networks on this task. It contains 32,000 samples of satellite imagery covering 2.56 x 2.56 km in 128 x 128 pixels, which are already pre-processed and ready to be used as input data for prediction systems. The existing literature, which includes baseline models provided by EarthNet and new extensions of these, provides a starting point for the analysis in this thesis. The thesis first trains the various models and chooses the currently best suitable model for landscape prediction. It then analyses the source of the limited prediction accuracy of the chosen model. The analysis shows that the prediction accuracy varies according to the location of the satellite frames, the greenness variation for each image time series, and the land usage of each area. Central Europe exhibits the lowest prediction accuracy among the studied regions; regions with moderate variation in greenness have a minimal EarthNet prediction score; and landscapes classified as croplands also display poor prediction performance. These findings lead to the recommendation to tune the training and testing of artificial neural networks for landscape prediction on specific landscapes. To complete this thesis, a use case is developed to demonstrate how the landscape is changing according to manipulated weather variables. The model was implemented and analyzed with the use of the programming language Python and expanded with Deep Learning libraries such as PyTorch.

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations

Abbreviations

Artificial Intelligence (AI)
European Digital Elevation Model (EU-DEM)
Convolutional Long Short-term Memory (ConvLSTM)
Stochastic Adversarial Video Prediction (SAVP)
Recurrent Neural Network (RNN)
Red, Green, Blue, and Near-infrared (RGBNIR)
Strongly Guided (SG)
Convolutional Neural Network (CNN)
In-Domain (IID)
Out-Of-Domain (OOD)
Root-Mean-Squared Error (RMSE)
Median Absolute Deviation (MAD)
Ordinary Least Squares (OLS)
Normalized Difference Vegetation Index (NDVI)
Earth Mover Distance (EMD)
Structural Similarity Index Measure (SSIM)
EarthNetScore (ENS)
European Space Agency (ESA)

Chapter 1

Introduction

This chapter introduces the topic of the thesis by first explaining the motivation, followed by the objectives and key questions of this project. It then gives an overview of the existing literature. Finally, it gives the requirements and lays out the structure of this thesis.

1.1 Motivation

Vegetation landscapes are important in ecological systems, climate regulation, and human well-being. Landscape prediction based on Sentinel-2 data and their response to climatic variables is beneficial for effective land management, conservation efforts, and climate change mitigation [17, 7, 9]. The integration of artificial intelligence (AI) models with satellite imagery and weather variables has emerged as a powerful approach to foreseeing vegetation patterns and improving predictive capabilities. However, the existing models have performance limitations as vegetation landscapes are intricate ecosystems influenced by multiple factors including climate, soil properties, land use, and other disturbances [10]. Changing the model's architecture and parameters doesn't seem to achieve high distinctions in performance. This circumstance prompted further investigation into the data and predictions, considering that all models rely on the identical EarthNet dataset. Understanding the cause of the limitations could help to understand what should be focused on to improve the models. Additionally, this process helps to refine the scope of the predictions, narrowing down their objectives. Explaining the source of low accuracies can help to improve the prediction of vegetation patterns. Their responses to climatic variables are essential for assessing ecosystem health and sustainable land management.

Satellite imagery is highly available and large datasets can be obtained by processing open-source images. This project will use the EarthNet2021 dataset, which represents a collection of Sentinel-2 satellite imagery and complementary weather variables [18]. Additionally, this data has been complemented with image masks that represent cloud pixels and the European digital elevation model (EU-DEM) [2]. This dataset provides a rich source of information capturing the spatiotemporal dynamics of vegetation landscapes. Moreover, the inclusion of weather variables, such as precipitation, sea pressure, and temperature enables comprehensive investigations into the correlation between climatic factors and vegetation dynamics.

AI models, particularly those employing machine learning algorithms, exhibit several key advantages in vegetation landscape prediction. Firstly, AI models have the capability to capture complex, nonlinear relationships between vegetation patterns and environmental variables. Secondly, these models can efficiently handle vast amounts of high-dimensional data, enabling the incorporation of multiple satellite image channels and weather variables. Lastly, AI models are adaptable and can learn from diverse datasets, facilitating the transferability of knowledge across different geographical regions and vegetation types.

1.2 Objectives

The main objective of this project is to choose and implement the best-performing model trained on the EarthNet2021 dataset from the literature and analyze its behavior and prediction results. A convolutional long short-term memory (ConvLSTM) model proved to be under the best-performing models and uses a suitable architecture to predict landscapes from Sentinel-2 data with additional weather variables. This thesis will first train the ConvLSTM on Sentinel-2 satellite imagery (time series with additional weather variables) to achieve its best performance. This is followed by an analysis of what causes its limitations in performance. Finally, a use case is demonstrated where the data is manipulated to scenarios that are predicted to be the future climate change. Specifically, the project is divided into two steps: 1) answering the following key question, and 2) demonstrating the use case:

Key question: What causes the limitation of performance in the ConvLSTM model predicting satellite imagery, based on a Sentinel-2 dataset with additional climate variables?
Here the focus of the analysis lies on the predictions of the chosen ConvLSTM model. The outcome of the neural network will be studied on aspects such as location, greenness variability, and land cover of the surfaces. Based on this an explanation will be given, as to why the model performs poorly for certain landscapes with specific properties.

The challenge that comes with this analysis is to find a pattern in the predictions of the model. Instead of checking the overall performance, the accuracy of predictions will be reviewed on different features for various locations across Europe. The performance is analyzed based on different variables such as greenness variation, similarity score, and land cover. This gives an insight into which patterns in landscapes result in poor prediction and influences the overall performance in a negative way.

Overall, the analysis can be split into three different experiments:

- **Location**:
  This experiment investigates the performance of the chosen model according

to its location. Multiple locations across Europe are tested and the score is analyzed.

- **Greenness variation**:
  Here, the variation of the image greenness is taken for every time series sample. This is compared to other variables such as the EarthNet score, location, and similarity index.

- **Land cover**:
  The last analysis investigates the land cover of every location. The type of land usage is compared to the performance, location, and greenness variability.

Use case: How will the landscape change if the climatic variables get manipulated?
This demonstration shows how landscape prediction changes when the climatic input gets modified. For this, two scenarios will be laid out corresponding to KNMI14 climate scenarios [1]. This is a prediction for the Netherlands, how the climate will change in 2050 and 2085. Mainly, the two variables (temperature and precipitation) will be exploited.

1.3 Existing literature

The EarthNet2021 challenge [18] is an ongoing open-source competition to study surface forecasting based on satellite imagery. It provides a pre-processed dataset with multiple components including different landscape information. There are given baseline models which can be used to get started. The goal is to challenge the participants to improve the earth surface prediction with deep learning models. There are guidelines that allow participation in this challenge. The challenge itself published 3 baseline models to help participants to kickstart their projects. Next to that, the competition's leaderboard shows some improved, open-source publications. This section explains the concepts of the existing models and compares their approaches. First, the three baseline models get laid out, after which the published improvements are shown. An overall comparison will be made and a suitable model will be chosen for further analysis.

1.3.1 Baseline Models

In this section, the three baseline models provided by EarthNet2021 are reviewed and compared. The models are called Persistence, Arcon, and Channel-U-Net. These models are very helpful to start off projects related to the EarthNet2021 challenge. As the database structure remains the same, those models can introduce participants to the possibilities of data processing, training, and testing. Additional variables such as the static topography and climatic condition can be used as conditioning parameters, which is also demonstrated in the baseline models.

Persistence

This model is a method that applies simple averaging. The EarthNet2021 toolkit provides a baseline based on NumPy. This method averages all non-cloud classified pixels over the context frames and uses this result as a prediction [18].

Arcon

This baseline is based on stochastic adversarial video prediction (SAVP) [14]. At first, it was used as an unguided/weakly guided deep learning model, but the EarthNet2021 challenge transformed it into a variables-guided model. This was achieved by using the climatic variables as extra video channels. For this, these daily frames had to be recalculated to their corresponding 5-daily mean. Initially, the SAVP model was used for video data, but because this use case includes only static images, all components for the motion prediction were disabled. The Arcon model was trained with the mean absolute error over non-cloud-covered pixels (non-masked L1 loss), corresponding to omitting the adversarial loss.

Channel-U-Net

This solution proposes a U-Net architecture with dense connections between the model's layers [20]. All available input information was stacked as channels and provided to the network. This procedure transforms the U-Net into a Channel-U-Net. The Channel-U-Net specifically used for the EarthNet2021 use case is implemented with an ImageNet [5] pre-trained DenseNet161 encoder [11] provided by the Python library PyTorch. The input of this network includes 10 5-daily context frames, the EU-DEM data with the same resolution, and the upsampled meteorological predictors to match the other's resolution. This all adds up to 191 input channels that are fed into the model. The output on the other hand corresponds to 80 channels including the four color channels of the future 20 5-daily prediction frames. The model was trained by the EarthNet2021 team for 100 Epochs with the following configuration: Masked L1 loss with Adam [12], an initial learning rate of 0.002 (decreased by factor 10 after 40, 70, and 90 Epochs), a batch size of 64 and 4 x V100 16GB GPUs.

1.3.2 Submitted Models

There are two participants in the EarthNet2021 challenge that submitted their models leading to improved results compared to the baseline models. Both use similar approaches by choosing the convolutional Long Short-Term Memory (ConvLSTM) networks as architecture. The first participant, a research group from the University TU Muenchen, submitted a model called Diaconu ConvLSTM [6]. This was followed by a second initiative by a research group at ETH Zuerich introducing two models called SGConvLSTM and SGEDConvLSTM [13]. Both submissions will be studied and analyzed. A final comparison of each other and also to the baselines will finalize this section.

Diaconu ConvLSTM

The ConvLSTM model has been trained on red, green, blue, and near-infrared (RGBNIR) frames, the weather conditions, and the EU-DEM as input. For this, the weather conditions had to be processed in order to make them usable. First, this information was averaged over five days to make the time steps match the satellite imagery. Second, the frames were scaled up to frames of 128 x 128 pixels to match the other input data. The EU-DEM data was attached to each frame and provided as additional input. All frames combined were stacked as channels resulting in input data of the dimension 128x128x10. While iterating over the context frames the fully combined input frames are considered. After exceeding t>c the previous predicted RGBNIR frames are used as input for the current step. This is illustrated in Figure 1.1.

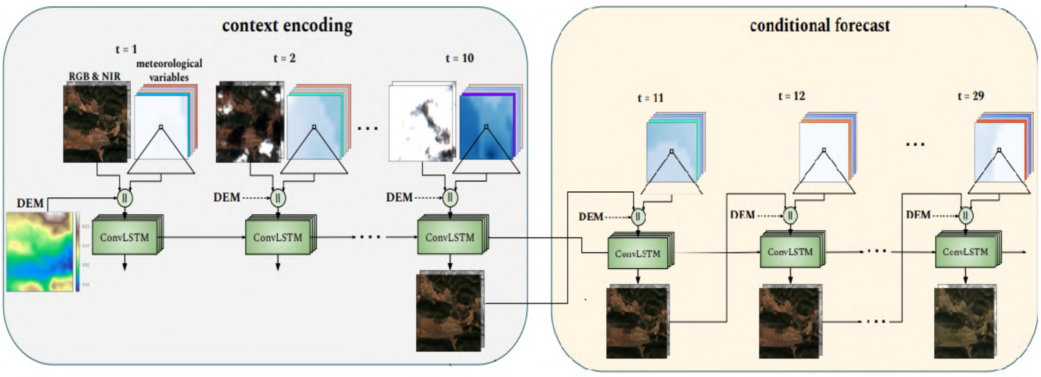


FIGURE 1.1: Training input of the ConvLSTM model: The context steps include the satellite image for each time point, the cloud masks, weather variables, and EU-DEM. In the target time steps the model uses as input the prediction of the previous step, weather variables, and EU-DEM. [6]

The model consists of four stacked layers: The first layer has 10 input channels and 32 output channels, the next two layers have both 32 input and output channels, and the final layer has 32 input and 4 output channels. The final dimension matches the required output size. The kernel size is 3x3 with a padding of one pixel. The resulting total number of parameters is around 200k [6].

This model has proven to be suitable for the EarthNet2021 application for several reasons. As it is a recurrent neural network (RNN), the temporal data that we are dealing with can be processed easily in such an architecture. Next to that, the provided masks for filtering the non-usable satellite imagery pixels can be used more easily to learn the model to ignore these areas. The gating mechanism is suitable for learning when to ignore certain areas through training on the context frames with the included cloud-marked masks. Also, this architecture allows adding the current weather variables in a straightforward manner. Providing weather information as input can also be used as guidance for the prediction steps. This makes it easy to turn this model into a guided network. Finally, it is clear that this model provides much flexibility, and changes in input and variables can easily be adapted.

SGConvLSTM and SGEDConvLSTM

Similar to the previous model two modifications of a convolutional LSTM network are presented. The models were implemented using the deep learning framework PyTorch Lightning, which is built on top of PyTorch and enables improved scalability. The hyperparameters were tuned using an Optuna-based hyperparameter optimization procedure.

The first deep learning architecture is a ConvLSTM inspired by a convolutional LSTM network [21], which is very similar to the architecture of the previous ConvLSTM model [6]. The model is called SGConvLSTM to reflect aspects related to the strongly guided (SG) modeling task. As shown in the following Figure 1.2 the process of guidance with the weather variables is the same during training. The context frames include all channels representing the stacked data of RGBNIR, weather variables, DEM, and cloud masks. In the target time period, the prediction is guided by the weather variable, DEM, and the RGBNIR frames of the previous time step as input.

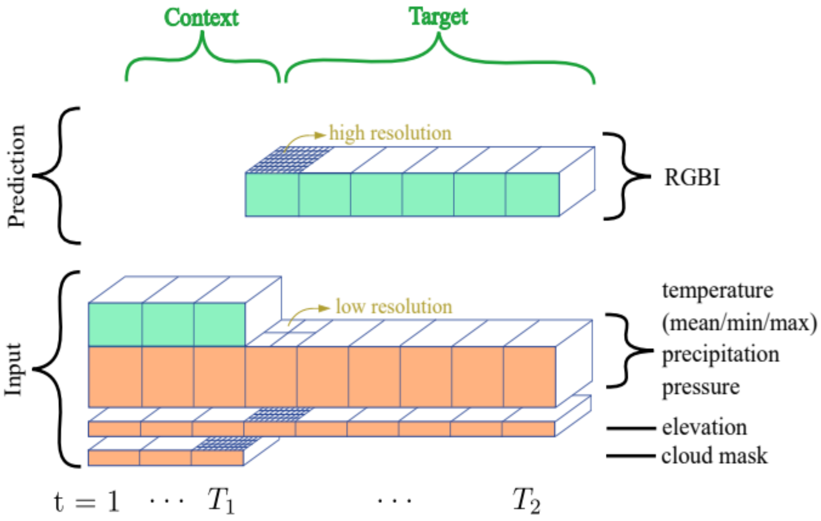


Figure 1.2: Training input SGConvLSTM: The input for the context time period consists of the landscape frames, E-OBS, EU-DEM, and cloud masks. In the target time range only previous predictions, E-OBS, and EU-DEM are fed into the model for prediction. [13]

The second model that was implemented by the same participants was the SGEDConvLSTM, which stands for an Encode-Decoder architecture. Such an architecture consists of two multilayer LSTM networks. The sequential output is fed to the first network (encoder) and as input to a second network (decoder) at each time step. Both networks, encoder, and decoder, require to have the same depth for this architecture. In this manner, another dimension of parameterization is added to the network without having to resort solely to stacking LSTM cells on top of each other. For the SGEDConvLSTM, the same hyperparameters are used as for the SGConvLSTM, except for the number of hidden channels, which was increased to 22 [13].

1.3.3 Overall comparison

The discussed baselines and submitted models are compared according to their scores for the test sets provided by EarthNet2021. The scores for the in-domain (IID) test images and out-of-domain (OOD) robustness test track are the most interesting for this comparison. The general EarthNet score (ENS) contains multiple components (MAD, OLS, EMD, SSIM) and is explained in Chapter 3. It should be mentioned that the last two models SGConvLSTM and SGEDConvLSTM [13] outperform all other models for the Extreme track, which is a special test set containing images of the extreme summer in Germany of 2018. Also, the simple Persistence model performs the best in the seasonal track (see Section 3.2), as generalization over a long period of time seems to be hard to reach for the other models. Overall, the ConvLSTM (Diaconu) is the best-performing one for the main track and the robustness track. The other improvements to the baselines (SGConvLSTM and SGEDConvLSTM) perform slightly worse but still improve the baselines significantly. This is shown in Tables 1.1 and 1.2.

Table 1.1: Overall comparison IID

	IID				
	ENS	MAD	OLS	EMD	SSIM
Persistence	0.2625	0.2315	0.3239	0.2099	0.3265
Channel-U-Net	0.2902	0.2482	0.3381	0.2336	0.3973
Arcon	0.2803	0.2414	0.3216	0.2258	0.3863
Diaconu	0.3266	0.2638	0.3513	0.2623	0.5565
SGConvLSTM	0.3176	0.2589	0.3456	0.2533	0.5292
SGEDConvLSTM	0.3164	0.2580	0.3440	0.2532	0.5237

Table 1.2: Overall comparison OOD

	OOD				
	ENS	MAD	OLS	EMD	SSIM
Persistence	0.2587	0.2248	0.3236	0.2123	0.3112
Channel-U-Net	0.2854	0.2402	0.3390	0.2371	0.3721
Arcon	0.2655	0.2314	0.3088	0.2177	0.3432
Diaconu	0.3204	0.2541	0.3522	0.2660	0.5125
SGConvLSTM	0.3146	0.2512	0.3481	0.2597	0.4977
SGEDConvLSTM	0.3121	0.2497	0.3450	0.2587	0.4887

1.4 Requirements

The requirements that are needed for this thesis are first of all a powerful computer and software for the interpretation of the data. The dataset was downloaded (500Gb) and the AI model was trained locally. The program was executed within an Anaconda3 environment and with the help of the IDE Visual Studio Code. The tools used in this thesis are the programming language Python combined with libraries such as PyTorch, Numpy, and Pandas.

1.5 Structure

The next chapter elaborates on the background theory behind the implemented convolutional LSTM network. The general architecture of these networks and the combination between a convolutional network and an LSTM is explained. This is followed by the third chapter which explains the EarthNet2021 dataset and its different components. This involves the training and testing data, but also the approach for testing and scoring. The fourth chapter contains the methodology and describes the analysis done on the predictions. This is followed by the fifth chapter, which includes the results of the performance of the trained model and the exploration results of the predictions. The sixth chapter then demonstrates the use case and gives the predictions of two climatic scenarios. Finally, the last chapter provides a discussion of the results and an overall conclusion. Also, some suggestions are given for future work.

Chapter 2

Background theory: Convolutional LSTMs

This chapter gives a theoretical explanation of the chosen model. As shown in Section 1.3.3, ConvLSTM is the best-performing one and will be used to analyze the predictions. The architecture of such a network is explained in the following section.

Convolutional LSTMs are a powerful architecture for modeling sequential data. By combining the advantages of LSTM networks and Convolutional Neural Networks (CNNs), ConvLSTMs can learn spatiotemporal features in image time-series data and is able to extract important features from image data at the same time. An additional advantage is that it's trainable on spatiotemporal data with fewer parameters than a fully-connected deep learning network [6].

The ConvLSTM consists of multiple units, where the data flows through. These units consist of three internal gates that control the information. The three gates are the input gate i_t (Eq. 2.1), the forget gate f_t (Eq. 2.2), and the output gate o_t (Eq. 2.3). Every single control gate works with its own assigned weights. Together, the combination delivers a new input X_t, with stored data of the previous state H_{t-1}. This is shown in the Equations 2.1-2.3, where $*$ denotes the convolution operator and σ is the sigmoid function.

$$i_t = \sigma(W_{ix} * X_t + W_{ih} * H_{t-1}) \tag{2.1}$$

$$f_t = \sigma(W_{fx} * X_t + W_{fh} * H_{t-1}) \tag{2.2}$$

$$o_t = \sigma(W_{ox} * X_t + W_{oh} * H_{t-1}) \tag{2.3}$$

The previously described gates are then used to update the long-term cell state C_t and the corresponding hidden state H_t as described in the following Equations 2.4 and 2.5. In these formulas, $\cdot$ stands for the Hadamard product. C_t is obtained with the new input X_t and the output from the previous step H_{t-1} and additionally with Hadamard product between the input gate f_t and the corresponding previous values C_{t-1} filtered by the forget gate. Finally, the hidden state H_t is calculated by the new cell state C_t cleared by the output gate.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_{cx} * X_t + W_{ch} * H_{t-1}) \tag{2.4}$$
$$H_t = o_t \cdot \tanh(C_t) \tag{2.5}$$

As mentioned, the ConvLSTM is a variant of LSTM (Long Short-Term Memory) containing a convolution operation inside the LSTM cell. It is a special kind of recurrent neural network (RNN). In a ConvLSTM the matrix multiplication is replaced with a convolution operation at each gate in the LSTM cell. By doing so, it captures underlying spatial features by convolution operations in multiple-dimensional data. The traditional LSTM input data is one-dimensional, which makes it not suitable for spatial sequence data such as satellite imagery. ConvLSTM is designed for 3-D data as input and makes it therefore suitable for the application of this thesis. In the following Figure 2.1 it is illustrated how one cell in a ConvLSTM network is constructed [23].

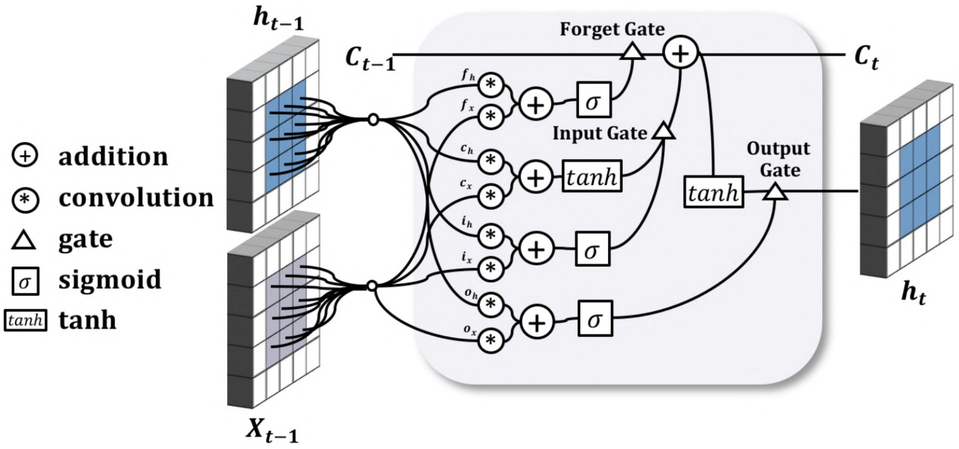


Figure 2.1: ConvLSTM cell: Convolutional operation on input gate, output gate and forget gate. The convolution is calculated between the kernel segment of two images. [23]

Chapter 3

Data

This chapter documents the dataset corresponding to the EarthNet2021 challenge. First, an explanation of the structure of the data is given, followed by a description of the testing tracks and the available scoring mechanism. Finally, the ESA WorldCover data is explained.

3.1 EarthNet2021 dataset

The EarthNet2021 dataset is a collection of processed satellite imagery time series at high resolution. Additionally, this dataset is extended with climatic predictors, which ideally should come from a seasonal weather model. The EarthNet2021 challenge approximates this from the E-OBS forecasts which contain the observation ground truth over several locations in Europe [4]. The pre-processed set of images originates from the public satellite mission Sentinel-2. The time series consists of satellite imagery that is revisiting the same location on the Earth map every 5 days.

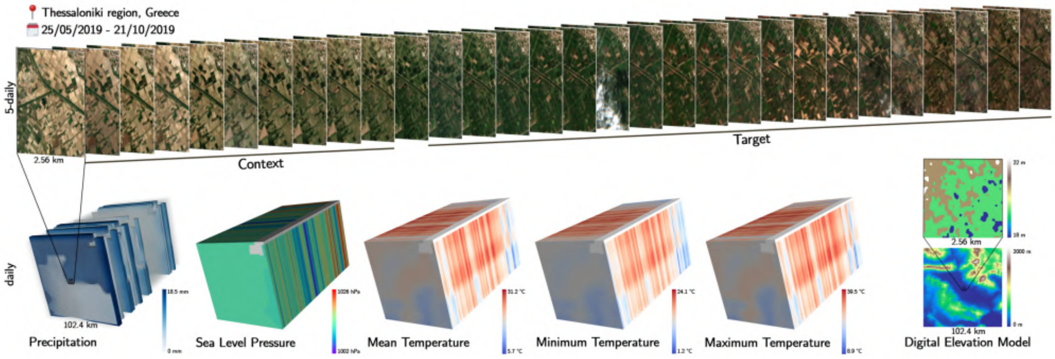


FIGURE 3.1: Structure of a mini-cube: 10 context frames and 20 target frames in 5-daily time steps. Additional weather variables: Precipitation, sea level pressure, mean-, min-, and max-temperature. Extended with EU-DEM data.[18]

Overall, the EarthNet2021 dataset contains 32000 samples, where one sample represents a mini-cube. One mini-cube contains 30 5-daily images of 128 x 128 pixels, which represents an area covering 2.56 km x 2.56 km. These frames consist of four different channels which are the three RGB channels and an additional near-infrared one. Also, each satellite image comes with a representative mask that indicates how many pixels of the frame were covered by clouds. The 5-daily frames are complemented with 150 daily frames of the E-OBS data with 5 meteorological variables: Precipitation, sea level pressure, and mean minimum and maximum temperature. This data is present at a mesoscale resolution where 80 x 80 pixels comply to cover an area of 102.4 km x 102.4 km. Additionally, the EU-DEM elevation model data is present in high- as well as mesoscale resolution. The structure of such a mini cube is illustrated in the Figure 3.1.

Combining these components in a way that they are analysis-ready samples, that can be split into a training and a test set, is very challenging. They need to be processed and prepared to be accessible for deep learning models. First, the Sentinel-2 imagery was downloaded by pre-filtering and only downloading a random subset of 110 tiles. More information about the Sentinel-2 tiles can be found in Section 4.2.1. This data was fused together with the climatic variable E-OBS and the surface model data EU-DEM. This is accomplished by reprojecting, resampling, and cutting the data into the desired shape. After this, the cubes could be generated by creating the data mask and compressing all the data in one array. The following Figure 3.2 shows the process of the fusion between the differentdata inputs.

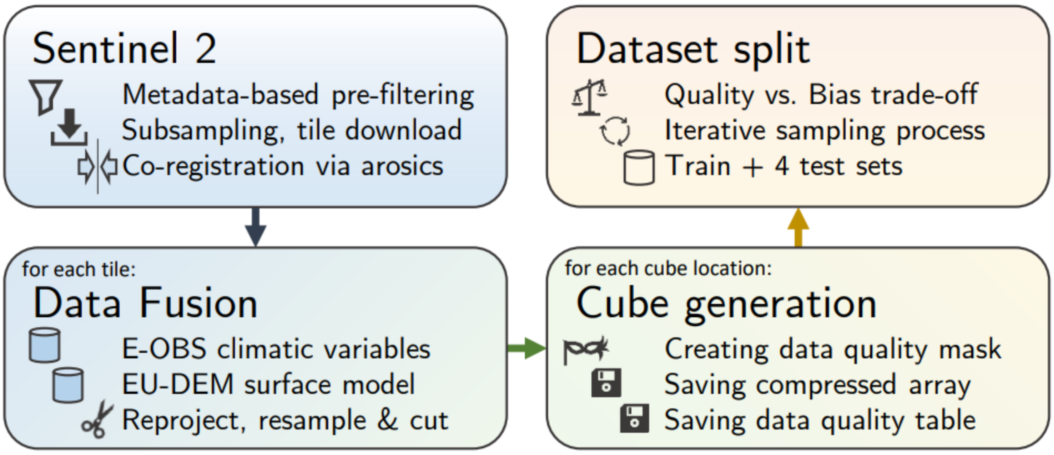

Sentinel 2
Metadata-based pre-filtering
Subsampling, tile download
Co-registration via arosics
Dataset split
Quality vs. Bias trade-off
Iterative sampling process
Train + 4 test sets
for each tile:
Data Fusion
E-OBS climatic variables
EU-DEM surface model
Reproject, resample & cut
for each cube location:
Cube generation
Creating data quality mask
Saving compressed array
Saving data quality table

Figure 3.2: Data fusion: Pre-processing steps to create the mini-cubes. Stepwise, downloading Sentinel-2 data, fusing together with E-OBS and EU-DEM, generating the compressed NumPy array with cloud masks, and splitting it into train and test sets. [18]

3.2 Testing tracks

The EarthNet2021 challenge provides different testing datasets, which allow for testing the accuracy of the trained models. The so-called testing tracks use different criteria to check the performance of the model on different attributes. The following 4 tracks are explained: The main in-domain track (IID), the robust out-of-domain track (OOD), the extreme summer track, and the seasonal cycle track.

- **Main (IID) track**:
  This track checks the model's validity and has very similar test mini-cubes to the training set. It contains around 4000 samples from the same region as the training set. Additionally, it was assured that there was no temporal overlapping between the samples of the same region. The models get 10 context frames of the 5-daily imagery back in time. Furthermore, 150 frames of the static topography and dynamic climate conditions are added 50 days back in time and 100 days ahead of time. The models should output 20 5-daily outputs of RGBNIR satellite imagery for the following 100 days. These can then be evaluated with the ground truth.

- **Robustness (OOD) track**:
  This track checks the robustness of the model when it comes to geo-locations. This means that some test images can be from the different tiles as seen in the training set. It contains a similar amount of test samples, but as these are from different domains it tests the spatial generalization capacity of the models. Also for this track, the context and prediction length respectively are 10 5-daily frames in the past (50 days back) and 20 5-daily frames ahead (100 future days).

- **Extreme summer track**:
  This track includes only test images from the extreme summer of 2018 in northern Germany. This implies that the test set differs in temperature from the training set. In this case, the track includes 20 5-daily context frames from the past, and a prediction is made for the following six months.

- **Seasonal cycle track**:
  This track spreads the test images over a longer period of time to check the prediction performance over different seasons. The context frames are provided over the past year and the prediction time frame is over the next two years. Like this, the models are being evaluated in the capability of the generalization of four seasons.

3.3 EarthNet score

The evaluation of predictions in the case of this challenge is influenced by multiple factors and is not straightforward. The clouds and other unforeseen disturbances might hinder evaluation and lead to untrustworthy predictions. For example, a ranking only using the root-mean-squared error (RMSE) would not consider such disturbances. Due to this reason, EarthNet2021 avoided imperfect predictions by combining multiple components into one EarthNetScore.

- **MAD**:
  The first score uses the median absolute deviation (MAD) between the prediction and test sample. The MAD score is a good evaluation of the distance in pixel space, which is desired to be as close as possible between the predicted and target values. It simply quantifies how close the pixels are in a robust way.

- **OLS**:
  Next, the OLS score is the difference of ordinary least squares (OLS) linear regression slopes of the normalized difference vegetation index (NDVI) time series. This checks if the prediction follows the trend in vegetation change. The NDVI maps are computed for the target and protection series after which the OLS models are fitted over time for each pixel. The comparison between the slopes then gives the score for this evaluation.

- **EMD**:
  Similarly, the third score indicates the earth mover distance (EMD), which is the pixel-wise distance between prediction pixels and the NDVI time series, but over a short span of 20 timesteps. This and the previous OLS score are robust against missing data points.

- **SSIM**:
  The final score is the structural similarity index measure (SSIM) which stands for the similarity of spatial structure between predicted images and target frames. It computes the average SSIM over channel and timestep. The SSIM index is calculated on various windows of an image. The measure between two windows x and y of common size (NxN) is [22]:

  $$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3.1}$$

  Where, μ_x and μ_y are the pixel sample mean of respectively x and y, σ_x and σ_y the variances, σ_{xy} the covariance, $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$, L the dynamic range of the pixel-values $(2^{\#bits\ per\ pixel} - 1)$, $k_1 = 0.01$, and $k_2 = 0.03$

Then the EarthNetScore (ENS) is calculated by the harmonic average of the four mentioned components as shown in Equation 3.2.

$$ENS = \frac{4}{\frac{1}{MAD} + \frac{1}{OLS} + \frac{1}{EMD} + \frac{1}{SSIM}} \tag{3.2}$$

3.4 ESA WorldCover map

In a later step, the analysis will compare the predictions with the ESA WorldCover map. This map gives the first global land cover for 2020 and 2021 at 10 m resolution. It is based on Sentinel-1 and Sentinel-2 data and was developed and validated in near-real time. The map is shown in Figure 3.3.

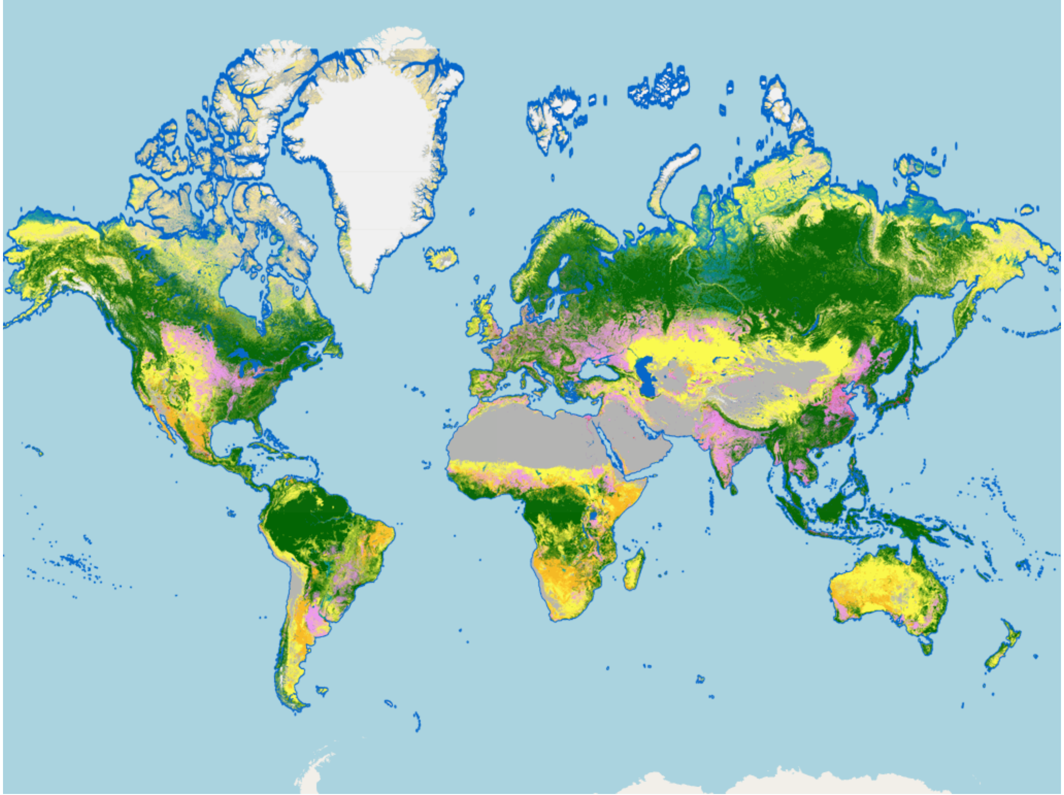


FIGURE 3.3: ESA landcover map: The complete WorldCover showing the globe categorized in land usage. [25]

The demand for precise and up-to-date information regarding land use and its changes has significantly increased in a rapidly evolving environment due to climate change. However, until now, regional or continental land cover maps mostly relied on low-resolution images. To address this, the European Space Agency (ESA) initiated the WorldCover project, inspired by the 2017 WorldCover conference.

One of the key achievements of this project was the release in October 2021 of a freely accessible global land cover map for 2020, consisting of 11 land cover classes. It underwent independent validation by ESA, resulting in a global overall accuracy of approximately 75 %. Given the positive response from users, ESA decided to expand the WorldCover project and tasked the WorldCover consortium with producing an updated version for 2021, aiming for even higher quality. The new WorldCover map for 2021 was made available on October 28, 2022, and achieved a global overall accuracy of 76.7 % [25].

The dataset contains high-resolution images and due to the large image sizes, only the tiles containing predicted samples were downloaded. For this, the grid coordinates of each tile were obtained and the percentages of land cover in this tile were stored in a data frame. The ESA WorldCover map categorizes the landscapes into 11 different sorts of coverage:

- Tree
- Shrub
- Grass
- Crop
- Built Up
- BareSparse
- SnowIce
- Water
- Wetland
- Mangroves
- MossLichen

The images of Figure 3.4 illustrate the frames extracted from the data for respectively the tiles 30TYR and 33VVG. These tiles are based on the military grid reference system and are explained in Section 4.2.1. The first frame shows a high level of agricultural usage. The second tile shows mainly a landscape covered by trees. These tiles will be related to the performance of the ConvLSTM.

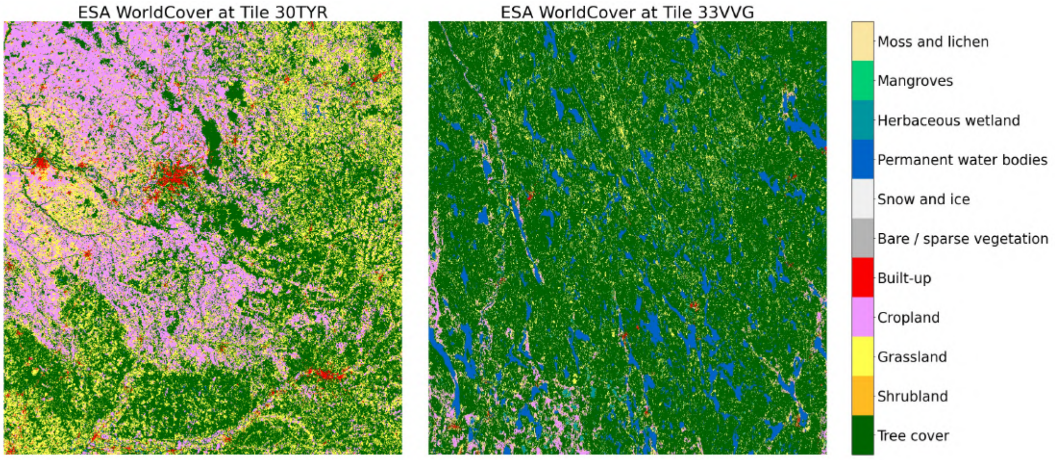


Figure 3.4: ESA WorldCover tiles: The left frame is derived from the coordinates 30TYR and the right tile from 33VVG. The color map indicates all 11 categories of land usage

Chapter 4

Methods

This chapter explains the methodology of this thesis and elaborates on the implementation of the ConvLSTM model. This is followed by the explanation of the experiments.

4.1 ConvLSTM implementation

As mentioned before, the chosen model for implementing satellite imagery prediction is the ConvLSTM model. This model belongs to the top performing models of the submitted improvements in the EarthNet2021 challenge. To analyze the performance on different attributes in a later step, this neural network needs to be trained first. For this, the pre-processed EarthNet2021 was downloaded and the model was trained on the training data.

4.1.1 Parameters

The training process was completed with the following parameters:

- Train batch size = 1
- Validation batch size = 1
- Test batch size = 1
- Number of workers = 4
- Loss: Adam L1 loss (arguments: learningrate = 0.001, Beta=(0.9,0.999))
- Learning rate schedule: MultiStepLR, Gamma = 0.5 (milestones: 10 Epochs, 20 Epochs, 50 Epochs)
- Context length = 10
- Target length = 20
- Input size = (128x128)
- Hidden dimension = 32
- Number of layers = 3
- Kernel size = (3x3)
- GPUs = 1
- Max. Epochs = 60

4.1.2 Setup

The model was implemented in a virtual conda environment and the open-source code of the Diaconu ConvLSTM [6] was used. The following commands describe the workflow of setting up the model:

Prepare the conda environment:

```
conda env create -f enpt111py39.yml
```

Activate the environment:

```
conda activate enpt111py39
```

Data download:

```
python scripts/data_download.py
```

Set the chosen model parameters in the configuration file:

```
./src/models/pt_convlstm/configs/convlstm.yaml
```

Train the model with the Python script:

```
python ./src/models/pt_convlstm/train.py \
--setting="./src/models/pt_convlstm/configs/convlstm.yaml"
```

Predict with the trained model:

```
python ./src/models/pt_convlstm/test.py \\
       --setting=./data/experiments/conv_lstm/
          version_31/settings.yaml \\
       --checkpoint=./data/experiments/conv_lstm/
          version_31/checkpoints/Epoch-epoch=59-ENS-
          EarthNetScore=0.3285.ckpt \\
       --track=iid \\
       --pred_dir=./data/scratch/preds/conv_lstm/
          version_31/iid_test_split
```

The Python version in this environment is 3.9.7 in combination with Anaconda3. The main libraries used are machine learning tools such as Pytorch with the help of computation libraries like NumPy, Pandas, OpenCV, MatPlotlib, and many more. The editor to modify and interact with the environment is the IDE Visual Studio Code. Some modifications to the code were necessary to allow this model to run on a local computer. Initially, the training process was utilizing multiple GPUs and the backend was not Windows compatible. After modification, the code was executable locally. The same root directory was used to create additional notebooks to process and analyze the prediction data. The files can be found on the following GitHub repository: https://github.com/JeroenSmets/An-analysis-of-landscape-predicitions-from-a-convolutional-LSTM-trained-on-the-EarthNet2021-dataset.git

4.1.3 Training

The model was trained on a local computer and therefore was limited to using only one GPU. This resulted in long training times of approximately 3 hours per epoch. Due to restricted memory usage, the batch size of the training, validation, and test processes was also limited to one. The training process therefore was lasting multiple days. In total the model consists of 201 K trainable parameters. These are split over its layers, where the first ConvLSTMCell0 contains 48.5 K, the ConvLSTMCell1 73.9K, the ConvLSTMCell2 73.9 K, and the final ConvLSTMCell3 5.2K parameters. The model layers have the following architecture:

```
ConvLSTM(
      (0): ConvLSTMCell(
         (conv): Conv2d(42, 128, kernelsize=(3, 3), stride
            =(1, 1), padding=(1, 1)))
      (1): ConvLSTMCell(
         (conv): Conv2d(64, 128, kernelsize=(3, 3), stride
            =(1, 1), padding=(1, 1)))
      (2): ConvLSTMCell(
         (conv): Conv2d(64, 128, kernelsize=(3, 3), stride
            =(1, 1), padding=(1, 1)))
      (3): ConvLSTMCell(
         (conv): Conv2d(36, 16, kernelsize=(3, 3), stride
            =(1, 1), padding=(1, 1))))
```

After training the model for 60 epochs it was interrupted automatically as this was configured. Figure 4.1 is showing the validation score for each epoch. It is clear that after some epochs the performance converges to its maximum capability. In the last 10 epoch steps the model was fluctuating only with very small accuracy differences. The maximum validation score reached during training was at epoch 59 with a score of ENS = 0.3285. Further training was not expected to make significant improvements. The trained model was stored for every epoch and finally, the checkpoint at epoch 59 was used for further analysis.

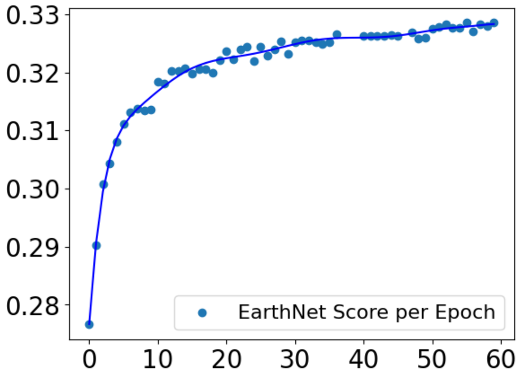


FIGURE 4.1: Validation ENS score after each epoch during training.

4.2 Experiments

This section discusses the approach to studying the predictions of the trained model in the form of three experiments. This analysis has the goal of better understanding where the limitations of the ConvLSTM come from. In the literature, none of the models was able to exceed the score of 0.3425 and make any major improvement. This supported the idea that rather than finding an improving machine learning model, the existing models should be investigated on where the bad performance is coming from. The problem lies more within the data and the focus of the application than the capabilities of the machine learning models. In particular, it is shown that the performance of the model depends on the location, the variance, and the land cover. This finding shows the importance of narrowing down the type of landscape prediction and the related objective.

4.2.1 Location

The first experiment compares the score of every prediction to the location of the satellite images. This is required to see which climate characteristics conform to which score. Our field of interest is Europe and investigating the vegetation of multiple areas shows us that there are many differences.

To investigate the score of every location the first step was to split the scoring calculation for every sample individually. The predictions are sorted by their Sentinel-2 tile coordinates. Every tile is named after the following convention:

N1C1C2, with
N1: Two digit number representing the longitude
C1: One character representing the latitude
C2: Two characters representing the sub quadrant

For Europe: (As shown in Figure 4.2)
N1: 29 to 34 respectively from west to east
C1: S, T, U, V respectively from south to north

Every tile used in the test set consists of a random amount of samples within the tile's area. Previously, the model's performance was checked by iterating over all tiles and calculating the overall score. Now, the score will be calculated for one sample and stored in a data table with its corresponding sample name. Every sample was named and stored as a compressed NumPy array according to the following structure:

```
tilename_daterange_hrcoordinates_mesocoordinates.npz
```

For every tile, the mean was taken from all the scores within this tile. With the help of the bounding coordinates of every tile, the representing score was indicated on a map. Also, the mean scores were checked per latitude in Europe. The mean values were obtained by filtering the sample tile names by the latitude characters. This resulted in four scores for the coordinates S, T, U, and V. This was motivated by the observation that the scores on the map were varying more along the latitude compared to the longitude.

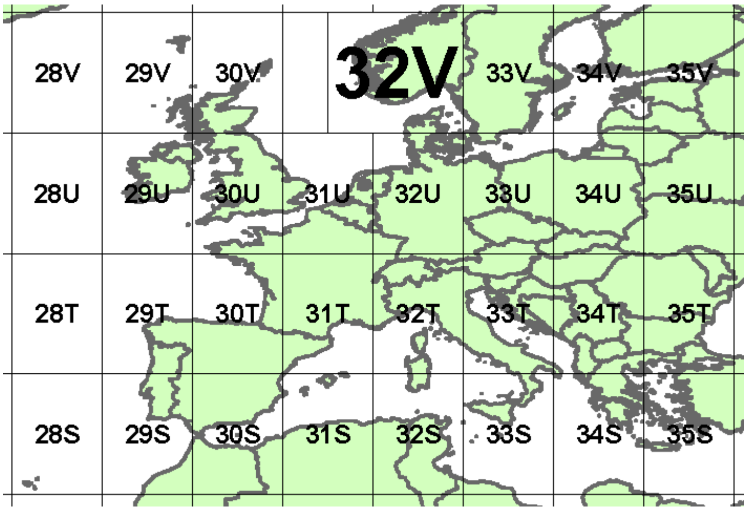


FIGURE 4.2: The HLS tiling system is identical to the Sentinel-2 coordinate system. The system is aligned with the Military Grid Reference System (MGRS) [16]

4.2.2 Variation

The second analysis is studying the change in vegetation of the images. This is important to analyze as the challenge of the model is to predict sudden changes. The performance is therefore also related to the amount of change in the image time series. Exploring the variation corresponding to the performance can help us understand which tiles are influencing the general performance in a negative way.

For this experiment, the NDVI image was extracted from each mini cube. Then, for each frame of the time series, the average NDVI score was calculated, resulting in a representative greenness value for each time point. For every time series, the average greenness and its standard deviation were obtained. Finally, the score was calculated for every sample. This constructed a data table of the 4219 samples including the NDVI value per time step, the mean, the standard deviation, and the ENS. This data helps to see the correlation between the vegetation change and the performance of the predictions. Next to that, the cloud coverage should be brought in relation to the NDVI variance. For this, the cloud masks were extracted from every frame of the mini cubes. Then the percentage of cloud coverage was calculated by division of the number of masked pixels by all pixels. This information was stored with every sample for later processing. This data can easily be compared to the previously mentioned table with the ENS, mean, and NDVI variance data table.

4.2.3 Land usage

This experiment utilizes the ESA WorldCover map [25] to extract the land cover information for each tile consisting of predicted samples. The aim is to represent for each tile the percentages of land usage. The sample locations within a tile were randomly chosen, so the percentages of land usage just indicate a higher chance that one sequence is from a certain land cover category. The kind of land cover has a big influence on the prediction performance. Some vegetation types are easier to predict than others. For example, the changes to a tree-covered area will remain the same and are easy to predict, but agriculture is harder to foresee as the farmer can make unexpected changes to the fields.

As a first step, the bounding coordinates of the appropriate Sentinel-2 tiles were used to download the same tile from the ESA WorldCover map. The high resolution (10m) limited the analysis to the relevant tiles to prevent long computation times. Then these arrays were analyzed on how much percentage of every usage category is present. There are 11 different landscape utilizations and for every array (representing an image of a tile) the number of pixels of every category was divided by the number of array elements. Multiplying this by a hundred gave the percentages of land usage per tile. This resulted in a data frame with data about the land cover for every tile consisting of predictions. To plot this data on a map, the percentages were used to create pie charts, and together with the Sentinel-2 tile coordinates the pie charts could be visualized on the right location of the map. Additionally, the color map of the pie charts was used to create a legend indicating which color represents which land usage.

Chapter 5

Results

This chapter shows the results of the proposed analysis. First, the performance of the ConvLSTM is discussed. This is followed by the results of the experiments.

5.1 ConvLSTM performance

The trained ConvLSTM model was used with the EarthNet2021 dataset consisting of four different testing tracks. Table 5.1 lists the EarthNet score and its components for all four testing tracks. The overall performance for this model is slightly better than the performance of the submission on the challenge. However, this is only a difference of 0.2 %, which gives consideration for analyzing what is causing this model to underperform. Clearly, the optimization of the model parameters cannot make a big difference in prediction accuracy. The main focus of this thesis lies on the main IID testing track. This is because this track consists of the most data spread over Europe and also because the predictions are the most similar to the targets.

Table 5.1: EarthNet scores for trained ConvLSTM

Track	ENS	MAD	OLS	EMD	SSIM
IID	0.3286	0.2654	0.3527	0.2638	0.5626
OOD	0.3213	0.2544	0.3533	0.2672	0.5139
extreme summer	0.2283	0.2251	0.2976	0.1994	0.2127
seasonal	0.2104	0.2161	0.326	0.2113	0.152

The following Figures 5.1 show the frame time-series for one of the samples in the 31UES (Benelux area) tile dated between 2018/03/28 to 2018/08/24. The illustration demonstrates landscape images in 10-daily time steps. The first sequence is representing the target images showing the reality to have a reference image to the predictions. These images are compared to the foreseen images to obtain the scoring. The blacked-out pixels here represent the cloud masks, which are not considered in the model. The predictions are shown in the second time series. At the first steps of the time series, the predictions seem accurate and the landscape patterns look very similar. However, when the landscape changes in later steps the predictions seem less accurate.

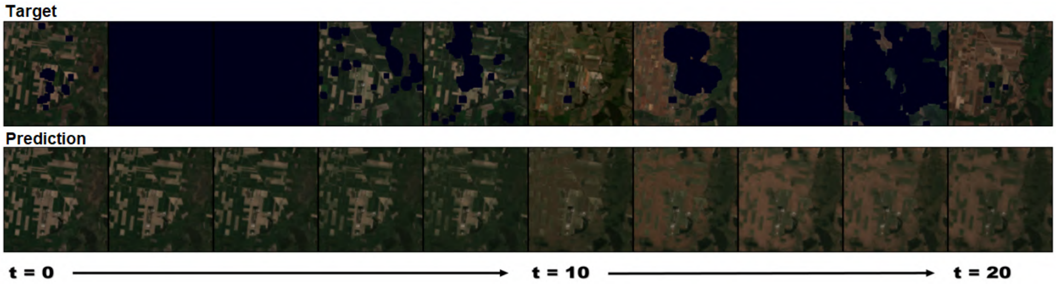


Figure 5.1: The output of the ConvLSTM model in comparison to the reality. RGB time series of target and prediction frames. Target time steps with 10 daily intervals.

Figure 5.2 is showing the vegetation greenness of the same time series, instead of colored images. The color map next to the sequences indicates the score on the NDVI scale. The landscape represents at the beginning of the sequence a high amount of greenness and scores less on the NDVI score towards the end of the time series.

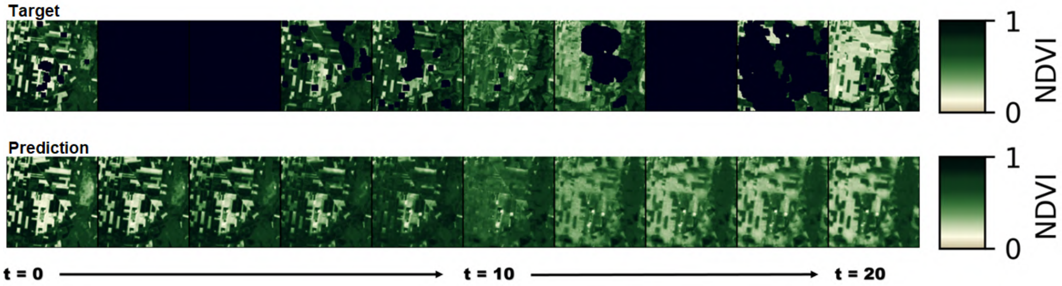


Figure 5.2: NDVI time series of target and prediction frames. Target time steps with 10 daily intervals.

5.2 Experiment outcomes

The analysis conducted on the outcome of the trained model consists of three parts. The prediction results are first studied on their location and the model's performance is split into regions. For this, one region is seen as a Sentinel-2 tile. The second experiment examines the variation in the vegetation greenness of the image time series. Last, the land cover percentages per tile are shown and compared to the performance.

5.2.1 Experiment 1

This experiment investigates the score according to the tile location. As the different areas in Europe have different characteristics it is useful to see how the varying landscapes score. The Sentinel-2 tiles are marked with first a number indicating the longitude, followed by a character describing the latitude, and finally two characters determining the sub-quadrant. For example, the tile 29SND contains samples at longitude 29, latitude S, and sub-quadrant ND.

Figure 5.7a shows the EarthNet score performance for every target tile in Europe. Overall, we see that central Europe doesn't score well and that the Mediterranean and Scandinavian areas perform better. The map indicating the score shows clearly that the performance varies with changing latitude, while the tiles laying in the same longitude are more likely to have the same score. This characteristic was expected as the general weather conditions change more over the latitude rather than the longitude. The score for one tile was calculated as the mean score between the scores of all sample time series in this area. The score for one mini cube was calculated according to Equation 3.2.

Next, some examples will demonstrate which characteristics will represent different locations and how this will be expected to influence the prediction performance.

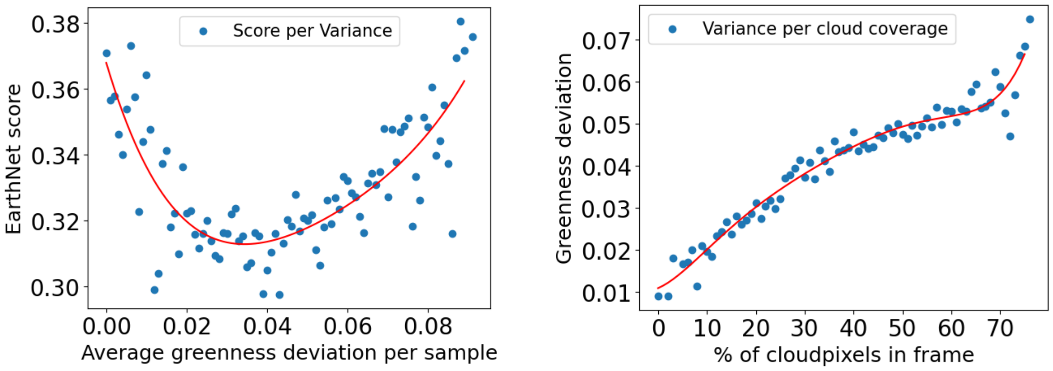
Mediterranean area: The vegetation is drier in the south than in the north of Europe, due to weather circumstances. This feature will be represented in the prediction score, as dry landscapes will have less change in vegetation. Landscapes that do not change much over time are easier to predict. Multiple indicators such as temperature and precipitation (provided as input in the ConvLSTM model) support this behavior. The cloud coverage in this area is low, leading to less rain and high temperatures. A climatic scenario like this leads to lower values on the NDVI scale and results mostly in brown landscapes. The average score of all tiles in this area reaches an ENS performance of: 0.3275 (Latitude S)

Central Europe: This region varies a lot in land usage. In central Europe, there is a lot of agriculture, many grass or tree landscapes, and even mountains. They all have very varying behavior. Additionally, there is generally a high percentage of cloud coverage. The precipitation rate is higher and the vegetation is more likely to change. These multiple options in the same climatic area will be harder to predict, as the input variables remain the same. The combined score of all tiles in this region is: 0.2859 (Latitude U)

Scandinavia: The last zone contains again different characteristics. Mainly this area is covered by trees, shrubland, and rocky landscapes. The landscape does not include many changes as the vegetation stays the same throughout the seasons. The precipitation and cloud coverage is high, but as the vegetation does not have many changes, this area is expected to have a good prediction performance. Together, all tiles in this area add up to an EarthNet score of: 0.3495 (Latitude V)

5.2.2 Experiment 2

The second experiment investigates the variation over the sample time series of the NDVI values for each frame. In Figure 5.3a the scatter plot is shown between the average NDVI variation per sample and the corresponding EarthNet score. The trend shows that a low variation results in high accuracy, medium variance results in low performance, and when the variation increases again the score is increasing significantly. The greenness variation per tile was plotted on a map in Figure 5.7b and can be compared to the score per tile on the image 5.7a. The variation increases with a higher latitude on the map. For the score, we have a minimum in central Europe and maximum performance for the lowest and highest latitude. It is expected that the landscape in the north has less change than in central Europe. However, the graph shows that these frames vary more. This occurrence is related to the cloud coverage in these areas, as there are more clouds in higher latitudes. Relating the amount of cloud-covered pixels to the NDVI greenness variation, as shown in Figure 5.3b, shows that more clouds cause higher greenness variation. This explains the high variance in the north of Europe.



(A) EarthNet score per variance

(B) Variance per cloud coverage

Figure 5.3: Two graphs: (A) represents the EarthNet score against the NDVI variation. (B) Gives the NDVI variation for the percentage of clouds in samples.

The cloud coverage influences the greenness variation of the time series, while a high variation does not directly imply a bad prediction performance. Next to the NDVI variation, the average SSIM index for every predicted time series was studied as an additional indicator for the similarity between the pictures. The SSIM is calculated according to Equation 3.1 and gives a score on how similar the two pictures are. Figure 5.4 gives the average SSIM score for every predicted tile, where the score represents the similarity between the first and the last predicted image in the sample. As we can see, the similarity index shows that the landscapes in central Europe vary the most. As the clouds must be ignored in order to investigate the landscape change of the frames, the predicted samples were used for this investigation. As shown in Appendix A, Figure A.2 gives the average SSIM score per tile latitude. Central Europe performs here poor and the Mediterranean and Scandinavian regions well.

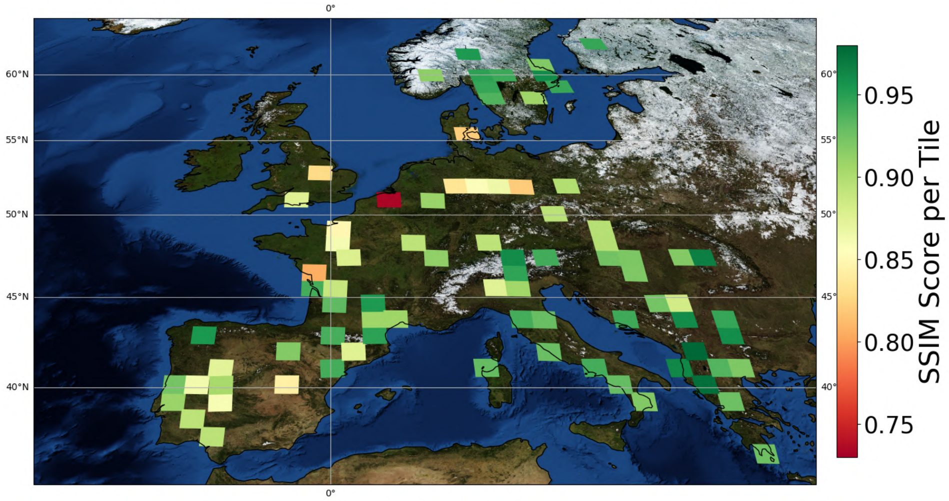


FIGURE 5.4: A map of Europe with the SSIM score, indicated by a color map, per Sentinel-2 tile

To finalize the two previous experiments, three time series are illustrated next to each other in Figure 5.5. In the first column, the time steps are shown. These are the 20 target frames in five daily intervals. Next to the sequences, two bar plots are indicating respectively the EarthNet score (green) and the NDVI variation (blue).

The first sample is from the tile 29SQB and represents a typical landscape in the Mediterranean area. Respectively to the other time series the cloud-masked pixels are very low here. The landscape looks very dried out (brown colored) and the vegetation is not changing a lot. The ENS performance is high and the NDVI variation is low.

The following mini cube gives a frame sequence from the tile 33UXQ, which is located in central Europe. This image includes already many cloud-covered areas. Clearly, this landscape is changing a lot as this is agriculture. The fields change constantly which results in a low similarity index. The overall score of this time series is low and the NDVI variation is medium to high. The frequently changing patterns of the fields cause a lot of unexpected changes which are very hard to predict. This is due to the observation that crop field arrangements are dependent on the individual farmer's choice. The usage of fields relies less on the foreseen weather variables which makes it harder for the model to consider.

The last sequence is from the tile 32VPP and is from the Scandinavian region. Here, many frames are utilized by cloud-marked pixels. The high cloud utilization is represented in the NDVI variation of the time series. This is not related to the similarity between the images and therefore does not influence the overall score in a negative way. The landscape seems to be covered by trees and shrubland. Between the non-cloud-covered areas, there are not many differences. This results in a good ENS performance for these regions.

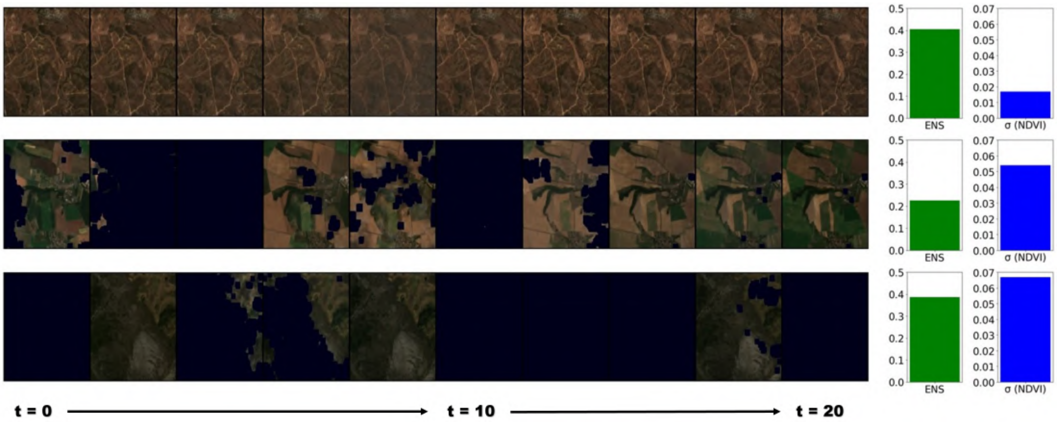


Figure 5.5: Time series with 10 daily time steps of three locations with different characteristics over a time interval from April to September. The first is from a Mediterranean area, the second from central Europe, and the last from Scandinavia. On the right, the bar plots represent their EarthNet score and NDVI variation.

To summarize, this experiment analyzed the predicted satellite imagery on multiple variables such as NDVI variation, SSIM score, and cloud coverage. Combining the greenness variation with the similarity index led to the conclusion that the performance is dependent on unexpected changes represented in the similarity score.

5.2.3 Experiment 3

The last experiment investigates the land usage for each tile. In Figure 5.7c the land cover percentages are illustrated with the pie charts. This map can be compared to the other maps with the EarthNet score and NDVI variation. It stands out that the worst-performing areas have a high percentage in crop fields and agriculture. These landscapes have very unpredictable features and include changing colors and shapes. The change in vegetation of the fields relies on the farmers who decide on the usage of the area. This is independent of the climatic variables and therefore harder to learn for the model, as it relies on the context frames and the climate forecasts. The better scoring areas, such as the tiles containing latitude S and latitude V, have clearly less percentage in crop fields. Those tiles consist mainly of tree coverage, grassland, and shrubland. These types of vegetation zones are completely dependent on the climate and are not exposed to sudden changes caused by humans. For example, a grass landscape will only change its greenness according to the precipitation and temperature variables. This is easier to predict as the climatic variables already indicated a change in climate.

In Figure 5.6 the average percentage of landcover is calculated per tile latitude. Here, only the most important categories are considered. The purple bar is representing agriculture usage. The green bar is the sum of the tree, grass, and shrub landscape. The last yellow bar stands for the combination of herbaceous wetlands, mangroves, and moss lichen. It is clear that the central European region utilizes a lot more crop fields. The average percentage reaches nearly 40 %, which results in a smaller percentage of green vegetation. This increases the probability that a sample from this tile is used for agriculture and results in more samples with lower prediction accuracy. In the better-scoring Mediterranean regions, there is clearly less percentage of cropland usage. Here, the employment lies between 10 % and 20 %. This means that the chance of a sample being taken from cropland is lower. The landscapes such as trees, grass, and shrub have a higher similarity index over time and result in better scores. In the south, a lot of the green landscapes are grass fields. This vegetation still varies over the seasons according to the weather variables. However, this is easier to predict for the models as these changes depend on the climatic input and result in a high score. In the best-scoring Scandinavian region, cropland usage is less than 5 %. The chance that a sample is taken from a landscape occupied by agriculture is very low. In the north, the most vegetation is trees and shrubland. These landscapes have a minimal change and are therefore easy to predict.

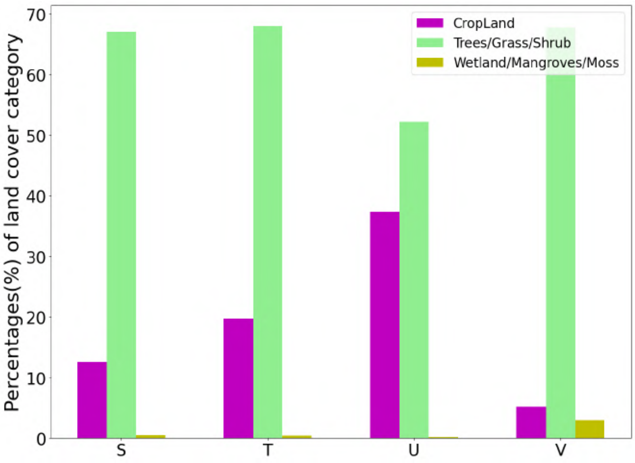
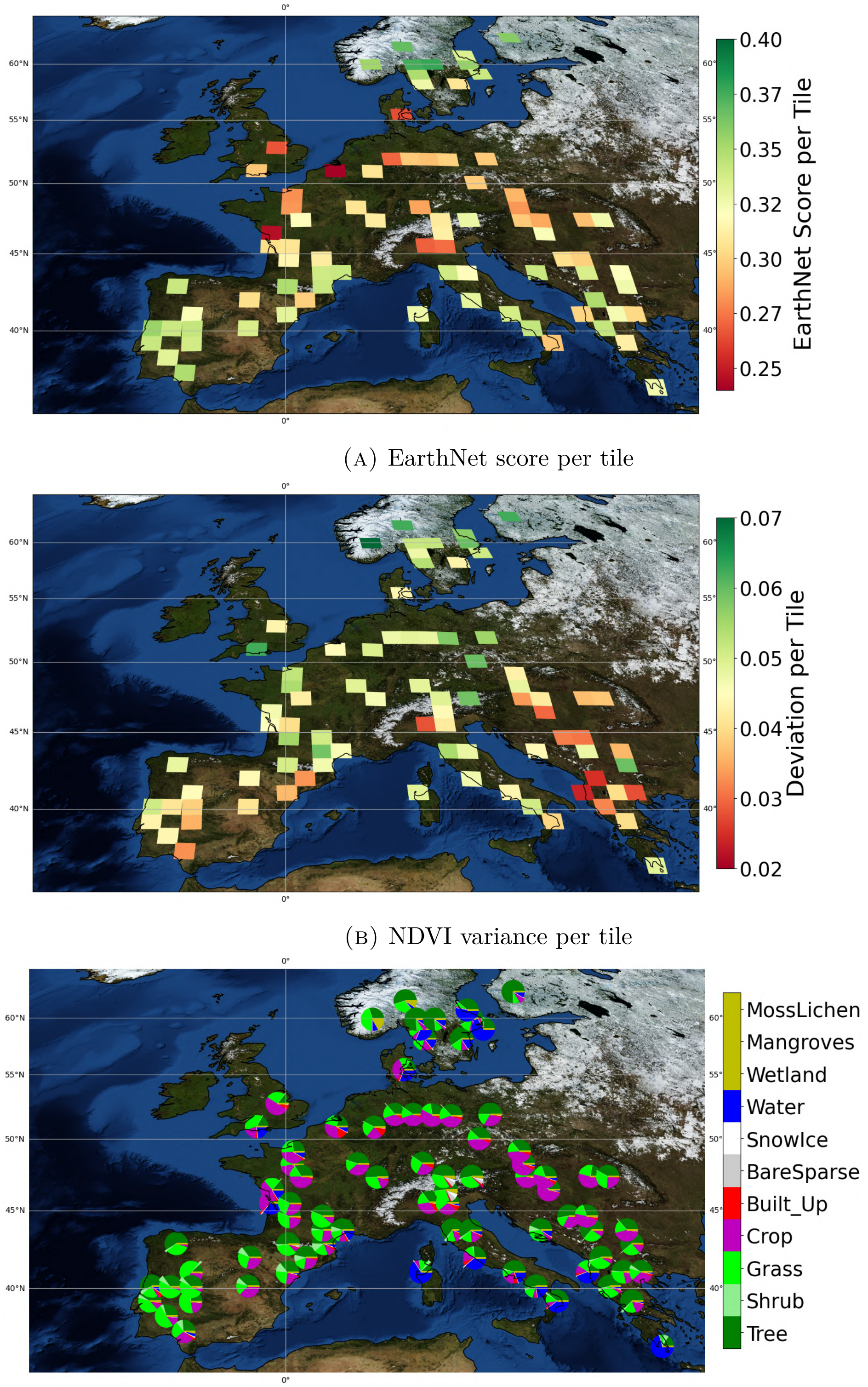


FIGURE 5.6: Average land cover % per tile latitudes S, T, U, and V. Land usage was categorized into (1) cropland, (2) trees, grass, and shrub, and (3) wetlands, mangroves, and moss.

To conclude, the outcome of this experiment is that the model's performance changes by latitude with high contrast. Analyzing the land usage over Europe showed that in certain latitude regions, agriculture is dominating the land utilization. This is reflected in the prediction score, as this land usage is the hardest to predict. The score is highly dependent on the land cover.



(A) EarthNet score per tile

(B) NDVI variance per tile

(C) Land cover per tile

Figure 5.7: Three maps: (A) ENS, (B) NDVI variation, and (C) land cover indicated with a color map in their Sentinel-2 tiles on a map of Europe.

Chapter 6

Use case

This chapter demonstrates the use case of the project. First, the procedure of setting up the use case is explained, followed by the demonstration of the prediction results.

6.1 Procedure

The aim of the use case is to predict the landscape images on the bases of two scenarios for the future. The aim is to manipulate the input data in a way that the prediction is made on future foreseen climatic variables. The scenarios are based on the forecast of KNMI'14: Climate Change scenarios for the 21st century from a Netherlands perspective [1].

In 2014 KNMI developed four scenarios outlining climate changes in the Netherlands by the years 2050 and 2085. These scenarios, known as the KNMI'14 climate scenarios involve variables such as temperature, precipitation, and sea level. Each scenario is accompanied by a distinct narrative, which is influenced by various factors, such as the level of CO2 emissions. The four KNMI'14 scenarios vary in terms of the extent of global warming (Moderate or Warm) and potential change in air circulation patterns (Low or High). These scenarios are illustrated in Figure 6.1. They establish the framework for envisioning potential future climate change in the Netherlands. Their purpose is to assess the implications of climate change and determine the significance and urgency of implementing measures for climate adaptation. The temperature in the Netherlands is projected to continue rising, with the most significant increase expected during winter and the least during spring. This will lead to a decrease in the number of cold winter days and an increase in the number of warm summer days, along with a higher likelihood of heat waves. Additionally, the temperature differences between coastal and inland areas will amplify in summer but diminish in winter.

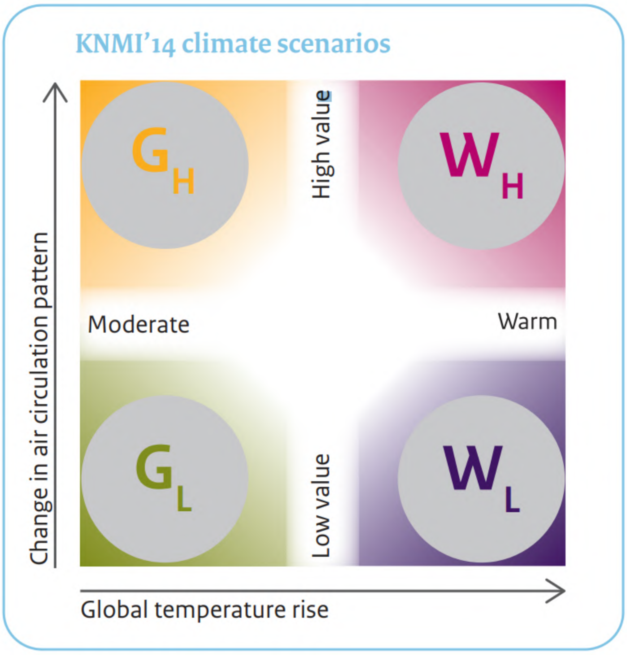
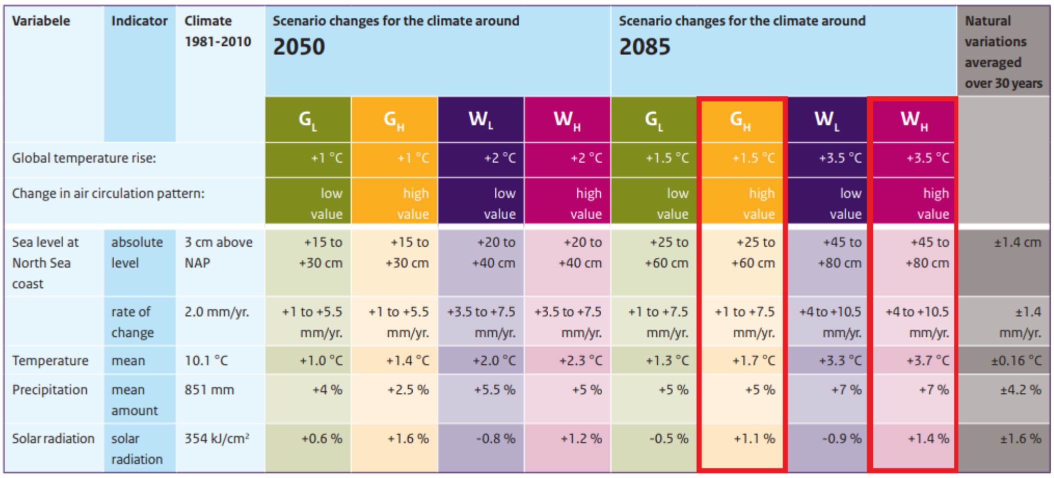


Figure 6.1: KNMI'14 climatic prediction [1] consisting of four different scenarios G_L, W_L, G_H, and W_H. All considering different conditions influenced by temperature rise and air circulation.

More specifically, KNMI'14 predicts the following alterations for the different variables: The precipitation levels are expected to increase. The probability of extreme rain showers accompanied by thunderstorms and hail will rise. However, two scenarios (G_H and W_H) suggest a decrease in mean precipitation during summer. The sea level rise is expected to accelerate and depends a lot on global temperature increases. By 2050, relative to the period between 1981 and 2010, the sea level will rise up to 40 centimeters. By 2085, it could be up to 80 centimeters higher along the Dutch coast. Changes in wind speed are anticipated to be minimal. The number of summer days with southerly to westerly wind directions will decrease across all scenarios, with the greatest reduction in the scenarios that exhibit more significant changes in air circulation patterns (G_H and W_H). Additionally, the G_H and W_H scenarios indicate an increase in westerly winds during winter. Solar radiation has slightly increased in recent decades, partially attributed to reduced air pollution. Clouds have also become more transparent, resulting in greater solar radiation. In the G_H and W_H scenarios, there is a slight decrease in cloudiness expected during future summers due to increased easterly winds.

For the demonstration of the vegetation change two scenarios of the 2085 predictions were chosen. The two most extreme forecasts were chosen to see the changes in a worst-case scenario. As illustrated in Figure 6.2 the climatic variable change was considered for scenarios G_H and W_H. In the first scenario, G_H predicts a mean precipitation increase of 5% and a raise of 1.7°C in temperature. The other scenario foresees a rise in precipitation of 7% and a change in temperature of 3.7°C.



Variabele	Indicator	Climate 1981-2010	Scenario changes for the climate around 2050				Scenario changes for the climate around 2085				Natural variations averaged over 30 years
			G_L	G_H	W_L	W_H	G_L	G_H	W_L	W_H	
Global temperature rise:			+1 °C	+1 °C	+2 °C	+2 °C	+1.5 °C	+1.5 °C	+3.5 °C	+3.5 °C	
Change in air circulation pattern:			low value	high value	low value	high value	low value	high value	low value	high value	
Sea level at North Sea coast	absolute level	3 cm above NAP	+15 to +30 cm	+15 to +30 cm	+20 to +40 cm	+20 to +40 cm	+25 to +60 cm	+25 to +60 cm	+45 to +80 cm	+45 to +80 cm	±1.4 cm
	rate of change	2.0 mm/yr.	+1 to +5.5 mm/yr.	+1 to +5.5 mm/yr.	+3.5 to +7.5 mm/yr.	+3.5 to +7.5 mm/yr.	+1 to +7.5 mm/yr.	+1 to +7.5 mm/yr.	+4 to +10.5 mm/yr.	+4 to +10.5 mm/yr.	±1.4 mm/yr.
Temperature	mean	10.1 °C	+1.0 °C	+1.4 °C	+2.0 °C	+2.3 °C	+1.3 °C	+1.7 °C	+3.3 °C	+3.7 °C	±0.16 °C
Precipitation	mean amount	851 mm	+4 %	+2.5 %	+5.5 %	+5 %	+5 %	+5 %	+7 %	+7 %	±4.2 %
Solar radiation	solar radiation	354 kJ/cm²	+0.6 %	+1.6 %	-0.8 %	+1.2 %	-0.5 %	+1.1 %	-0.9 %	+1.4 %	±1.6 %

FIGURE 6.2: Weather variables predicted changes according to KNMI'14 scenario changes [1]. The forecast is done for 2050 and 2085. All four scenarios predict variables such as Sea level, temperature, precipitation, and solar radiation.

The goal of this use case is to change the climatic variables according to the chosen scenarios of KNMI'14. For this, the trained model should be tested on the original samples, the manipulated context according to G_H 2085, and the changed samples with input variables from W_H 2085. To accomplish this the original climatic variables were extracted from the context mini cubes and changed to the desired values. For the precipitation, the predicted percentage increase was added to the array representing the amount of rain for every pixel. For the temperature, the corresponding arrays were also manipulated by adding to each element the recalculated values of the scenarios. The climatic variables in the ConvLSTM model were re-scaled to lie in a range from 0 to 1. For the precipitation the values are re-scaled according to Equation 6.1 and for the temperature according to Equation 6.2. The manipulated data is then stored back in the mini cubes and saved as a new context dataset.

$$Rain(mm) = 50 * rain \tag{6.1}$$

$$Temperatur(°C) = 5000(2temp - 1) \tag{6.2}$$

6.2 Predictions

The newly created test tracks based on scenarios G_H and W_H are used to make predictions starting from satellite images taken from the past. This use case creates a scenario that investigates how an existing landscape reacts to weather variables predicted for the future. This could be helpful to understand how landscapes could evolve in the future due to certain climatic conditions as predicted by KMNI'14. Starting from a context with certain vegetation characteristics it is possible to analyze the impact of the temperature and precipitation on such landscapes.

In Figure 6.3 three frame scenarios are illustrated from the tile 31UES (Benelux area). To have a comparison, first, the target time series is shown for the same location, which represents reality. All time series are representing 10-daily time steps showing a period of the next 100 days. The first prediction sequence gives the original frames, the second gives the adapted scenario W_H, and the final frames represent the scenario G_H. The new predictions show that scenario W_H, with a rise in precipitation of 7% and a change in temperature of 3.7 °C, predicts a greener landscape than the original time series. Overall, the frame into a monotone green-colored area. The other scenario with an increase of 5% and a rise of 1.7 °Cin temperature converges to a browner landscape, which means that vegetation seems to disappear.

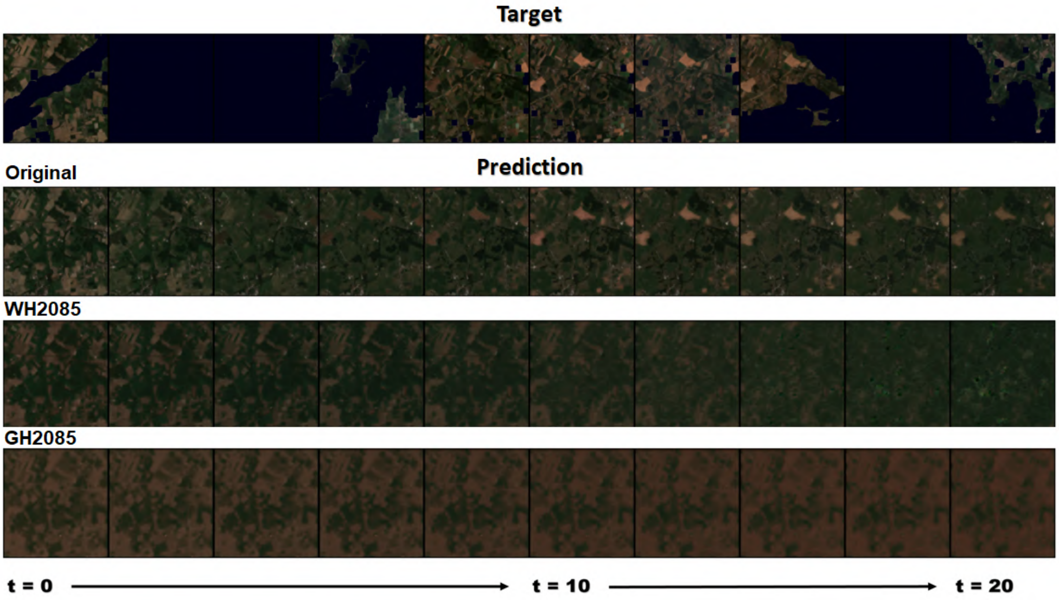


Figure 6.3: Four sequences: First, the target frames for reference, followed by the predicted time series with original climatic input. Last, the two sequences respectively the prediction with manipulated input according to scenarios W_H and G_H.

In Figure 6.4 the vegetation greenness is shown for the different scenarios. The NDVI scale shows the amount of vegetation for each frame. As previously mentioned the scenario WH2085 changes to more vegetation and the scenario GH2085 converges to less greenness.

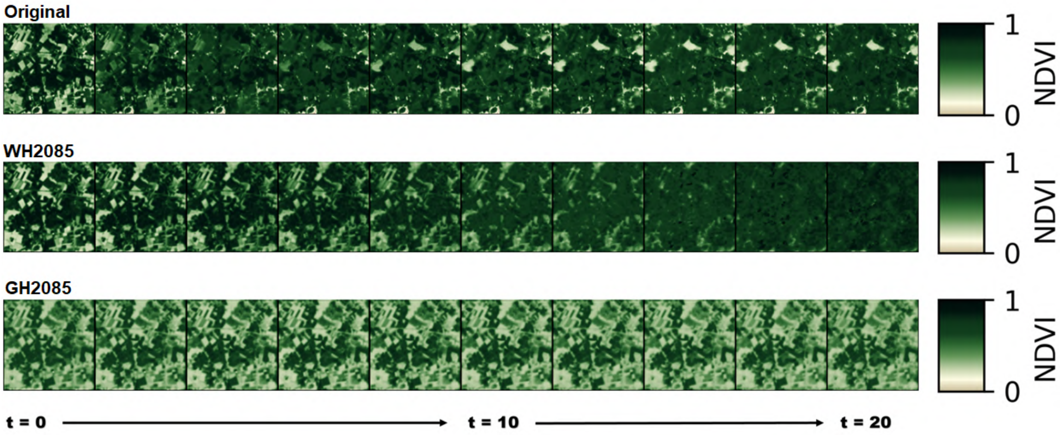


Figure 6.4: Three NDVI greenness time series predictions with the original input and two manipulated inputs according to climatic KNMI'14 scenarios W_H and G_H. All time series have a time step of 10 days.

Chapter 7

Conclusion

In the final chapter a conclusion is drawn from the analysis. First, the results will be elaborated in a discussion by zooming out and seeing the outcome from a bigger perspective. Finally, some future suggestions are given, and how the results of this thesis can help the following research

This project has started by reviewing the EarthNet2021 challenge and studying the existing literature on this topic. Existing models were created to predict satellite imagery based on the EarthNet dataset. This data consists of Sentinel-2 frames and additional climatic variables. Only minor improvements were noticed, comparing the capabilities of the submitted models. The limitations of those models were studied by choosing a dedicated model to analyze its predictions. For this, the model was trained to achieve its best performance. Then, the predictions were made on the EarthNet IID test track. The prediction scores, location, NDVI variance, similarity index, cloud coverage, and land usage were studied and compared to each other. This gave a better understanding of why certain predictions resulted in poor scoring performance.

Several tests were performed, such as plotting the scores of each tile on a map to see where the inaccurate predictions are located. Also, the average scores per latitude coordinate were compared. A second experiment investigated the NDVI variance and plotted these values according to their location. The EarthNet score was compared as a function of the greenness variation. Cloud coverage was brought in relation to the variance and the SSIM index was put in context with the NDVI change. The final analysis included seeing the land usage on the map. The percentages of landscape occupation were compared to each other according to their latitude coordinate.

Checking the locations showed that in the latitude of central Europe, the lowest accuracy was gained. The results showed that a medium NDVI variance and low SSIM index score were performing the worst. This happened to be the landscapes where the most land usage is utilized by agriculture. To summarize, the performance of the prediction was particularly poor for tiles of latitude U. Many frames from this area include crop fields and therefore have unexpected variations in the landscape with the lowest similarity index (not considering the clouds). As these changes are dependent on the farmer's choices and rely less on the climatic variables these images are hard to predict. This results in a bad performance. Regions like these perform approximately 10 % less than good-performing tiles and therefore influence the overall performance in a negative way.

7.1 Discussion

In this section, the findings of this project will be discussed in a broader context. As previously concluded, the results of the tests showed which characteristics of satellite images lead to low-accuracy predictions. This knowledge can improve our understanding of the limitations of the model. Rather than trying to optimize the parameters of the machine learning model and trying to find the best-performing architecture, the approach of this thesis is to specify the application of the models and increase the performance by narrowing down the target.

At first, a model from the literature was chosen on which the experiments have been conducted. The training of the ConvLSTM resulted in an ENS performance of 0.3286 and showed a slight improvement of 0.2% compared to the accuracy of 0.3266 from the original model. Optimizing the model parameters does not result in significantly better results. Relating the performance to the other models in the literature shows that even exchanging model architecture does not improve the accuracy extensively. The best-performing model in the leaderboard of the EarthNet2021 challenge is the EarthFormer with a score of 0.3425. The Earthformer is a space-time transformer using an attention block called Cuboid Attention [8]. Although this system is using a completely different architecture, which is currently seen in the literature as the most capable model for deep learning applications, the performance only increases by 1.59% to the self-trained ConvLSTM. The worst-performing model in the literature is the SGEDConvLSTM with a score of 0.3164. This is only 1.02% accuracy less than the trained model from this thesis. So, it is clear that changing the architecture and its corresponding parameters only results in a performance difference of a total of 2.61%. These differences are barely noticeable and motivated the main question of this thesis to investigate the data and application of the task.

Taking into account that the locations in Europe have different attributes the model shouldn't be biased by its data selection. The first experiment resulted in a categorization of scoring performance according to the latitude coordinates. It should be mentioned that the cloud coverage also varies along this vertical axis. EarthNet2021 pre-filtered their data by randomly choosing 110 tiles with at least 80% land visible for the best-performing tile in the time series together with a minimum of 90% data coverage [18]. The clouds were classified using a deep residual neural network [15] and specifies these pixels as unusable. As there are significantly fewer clouds in the south this results in more processed data for these regions. As the Scandinavian regions can reach up to approximately 75 % of cloud pixels these samples have a lot less computed data. Comparing the regions over their latitude brings many different climatic characteristics and therefore a lot of varying vegetation.

The EarthNet team already discussed the bias-quality trade-off and concluded that most high-quality mini cubes are from summer in the Mediterranean. Also, they have shown that the dataset does not contain samples covering winter in the northern latitudes. This is due to the cloud masking classifying snow as clouds [18]. To compensate for the high data loss in the northern latitudes the cloud coverage should be taken into consideration for the data preparation. The goal is to obtain a model that will perform equally well for vegetation at different latitudes. For this, it should have been trained on equally many time series that include the same amount of pixel data with their corresponding climatic variables. During the dataset creation, the cloud percentages can be taken into account when selecting the number of samples from one region. For example, when the average of double the amount of pixels in a tile, compared to another tile, are cloud marked then it would be helpful to select double the amount of samples from this tile to compensate for the bias. In this manner, we train and create a model which is unbiased and ready to predict the landscape over Europe with higher performance independent from the location.

The second discussion relates the outcome of the NDVI variation analysis to the literature. As a result of the corresponding experiment, the score compared to the greenness variation showed a minimum for medium variation. Here, the cloud coverage increased the variance the more you go north in latitude. The EarthNet challenge already investigated this appearance and showed that the more north you go in latitude, the more cloud coverage you have. This is demonstrated in Figure 7.1 and was underlined with 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" [18]. The experiment demonstrated that more clouds lead to a high variation, but together with the similarity index everything beyond the minimum performing variance showed to be caused by clouds and not influence the performance directly. In order to use this information to achieve better performance the cause should be explained. Tiles with medium greenness variation and a poor similarity index are shown to represent landscapes from agriculture. These are varying extensively and include a lot of changes due to farmer's choices of field occupation. Improving the performance of a model dedicated to predicting vegetation greenness change, therefore should exclude this kind of data.

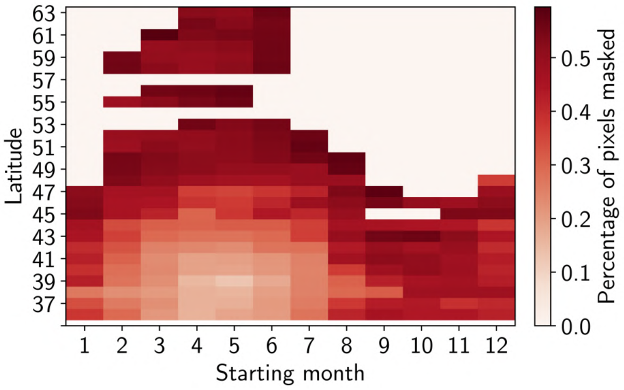


Figure 7.1: Percentage of cloud masked pixels related to its latitude and the month of the year. The color map indicated the percentage of clouds and the x-, and y-axes respectively the starting month and the latitude.[18]

The last point to consider, motivated by the findings of the experiments, is to narrow the application purpose of the predictions by specifying the data. The pre-processed dataset of EarthNet2021 includes randomly selected samples over Europe [18]. As the analysis of the land usage showed, the samples can contain all kinds of landscapes. Re-thinking the data selection could help define the exact goal of the predictions. This reconstruction should be done on as well the training data as the testing data. Currently, the predictions can contain any sort of landscape and this raises the question of what the exact intention of the models is. The models are trained to predict the vegetation change according to the climatic variables, but as different kinds of vegetation have completely different behavior this should be narrowed down. This can be clarified with the following two examples:

- **Example 1:**
  Suppose the intention of the model is to predict the greenness of vegetation according to the changing climate. In that case, it is more applicable to filter the data by landscapes including trees, grassland, and shrubland. Other landscapes, that include vegetation changes not directly relying on the climate, won't gain additional information to this goal. The ESA WorldCover dataset has proven to be suitable to classify which landscapes contain which kind of usage on a high resolution [24]. This can help with the selection process of what kind of data is needed to investigate greenness change.

- **Example 2:**
  If the use case is to predict how agriculture is changing over time, then the model should be trained with only this data. For instance, a seasonal prediction system can be used in a crop yield model to assess the performance and usefulness of such a system for crop yield forecasting [3]. This is a new application and therefore needs to consider its related data. As these changes are less dependent on the climate this underlines the fact that more context frames are required. Landscapes with crop field usage have very unique characteristics and should not be mixed with other data.

By determining the specific goal of the model a higher performance can be achieved and the purpose gets clarified at the same time. The predictions are more useful for further analysis as the user knows what to look for. If the purpose is for example to predict accurate change in greenness, an investigation into future vegetation growth could be made. As the third experiment showed the land usage data is available and can be used for the dataset tile selection. In order to achieve a high performance the training and test frames could exclude agriculture. This results in a good-performing model with the aim to predict vegetation change in green landscapes such as trees, grassland, and shrubland.

The demonstrated use case showed climate scenarios predicted to be the future changes in the Netherlands. As demonstrated in the modification of the annual precipitation and temperature predictors displaying novel landscapes for hypothetical scenarios [19], it is clear that the climatic variable has a strong impact on the vegetation change. Both use cases show that a large increase in precipitation changes the vegetation to a higher percentage of greenness. Lower temperature with lower precipitation however changes the landscape to less greenness.

7.2 Future work

Finally, to complete this thesis some suggestions are given for future work. The limitations of the trained model are discussed and propositions on how to use the result to achieve better performance are given.

The first limitation of the ConvLSTM trained with the EarthNet2021 dataset is that the data is not filtered down to the exact field of application. As mentioned before, the data includes satellite images of landscapes occupied by various usage. If the goal of this model is to predict future changes of a specific landscape type guided by climatic variables, then the dataset should only consist of landscapes covering the same type. The deep learning architecture ConvLSTM proved to be suitable for the prediction of landscapes. One suggestion is to use the same machine learning model (ConvLSTM), but re-think the data pre-processing. More features can be considered while downloading the data such as land usage, cloud coverage, and location. In this way, the model is biased on purpose in the direction of the prediction goal. Including various data from different kinds of landscapes leads to the fact that the model has fewer data to learn from for one distinctive landscape type. So overall, this study showed that rather than focusing too much on the deep learning model it is best to focus on the correct usage of it.

Second, the used sources were limited in computation power. As the model was trained locally and only one GPU was used to train the ConvLSTM there were some limitations of the parameter settings. Also, the amount of data to be processed is bounded, and using stronger computers working with parallel GPUs could shorten the waiting time. In the case of considering including more data fast computing units would be of a big advantage.

Third, the model was only analyzed on one test track. Exploring the different tracks and doing a similar investigation on the performance of the other tracks could help understand new causes of prediction difficulties. Especially, the seasonal testing track includes many changes as the context and target time range last over multiple seasons. Analyzing the predictions of these and checking where the model performs well and badly proves the strengths and weaknesses of the model. Also, the robustness track with test samples from varying locations can further add understanding to the flaws of the model. The predictions were already studied according to their location, but comparing this to an additional prediction set with deviating locations can help to determine how dependent the performance is on its exact latitude.

Last, the model is purely dependent on the context frames and the weather variables. The predictions are not considering the quality of the input data. This means that the model doesn't contemplate how many pixels are masked in the context data and how accurate the weather predictions are. Adding a heuristic could provide an additional parameter that indicates this.

Appendices

Appendix A

A.1 Experiment 1:

While analyzing the score related to its location it was also investigated how many of the samples within a tile performed good, medium, and poor. This is demonstrated in Figure A.1.

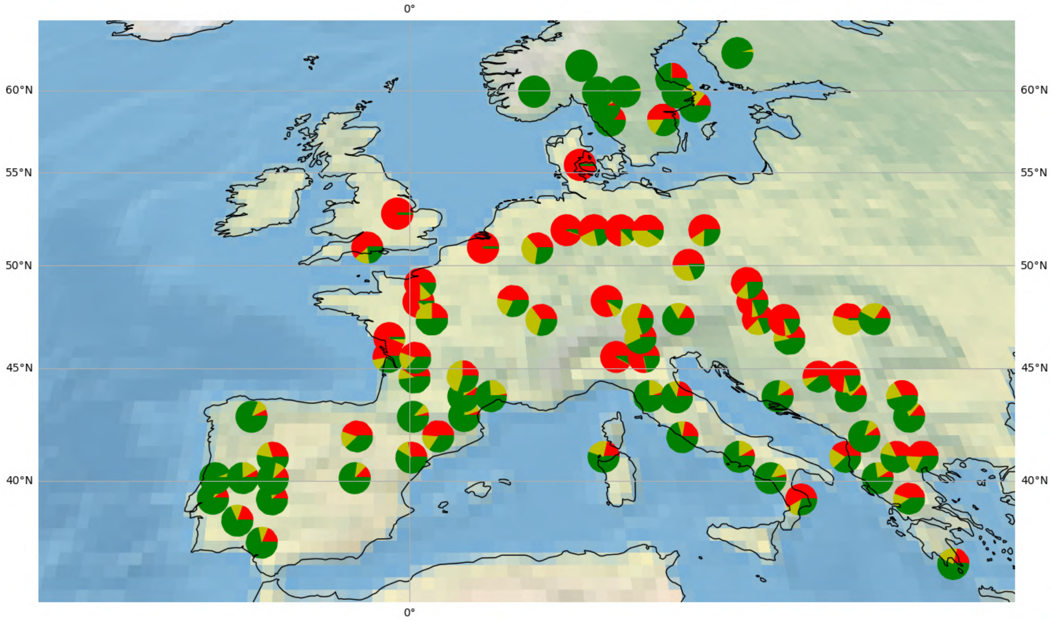


FIGURE A.1: Percentages of sample performance per tile

A.2 Experiment 2:

Figure A.2 gives the average SSIM score for every tile latitude. It demonstrates that the higher we go in latitude, the lower the similarity index is. This appearance is similar like with the NDVI variation and is caused by increasing cloud coverage.

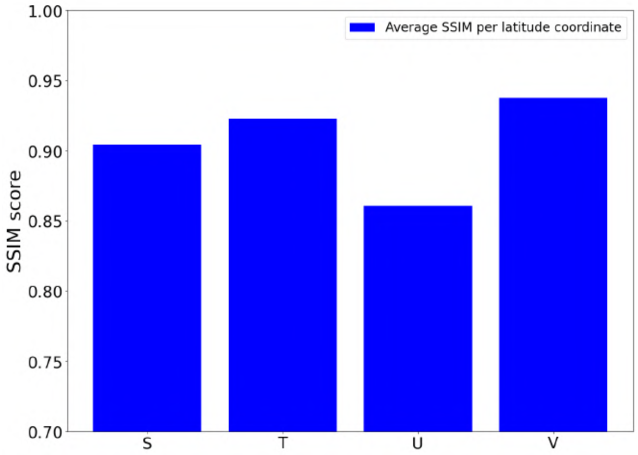


FIGURE A.2: Average SSIM Scores per latitudecoordinate

A.3 Tables and data:

All used tables and data can be found on the personal git repository of this thesis. Also, the notebooks with code are included which was used for analysis and processing.

https://github.com/JeroenSmets/An-analysis-of-landscape-predicitions-from-a-convolutional-LSTM-trained-on-the-EarthNet2021-dataset.git

Bibliography

[1] J. Attema, A. Bakker, J. Beersma, J. Bessembinder, R. Boers, T. Brandsma, H. van den Brink, S. Drijfhout, H. Eskes, R. Haarsma, et al. Knmi14: Climate change scenarios for the 21st century–a netherlands perspective. *KNMI: De Bilt, The Netherlands*, 2014.

[2] 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.

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

[4] R. C. Cornes, G. van der Schrier, E. J. van den Besselaar, and P. D. Jones. An ensemble version of the e-obs temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409, 2018.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] C.-A. Diaconu, S. Saha, S. Günnemann, and X. X. Zhu. Understanding the role of weather data for earth surface forecasting using a convlstm-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1362–1371, 2022.

[7] M. Fauvel, M. Lopes, T. Dubo, J. Rivers-Moore, P.-L. Frison, N. Gross, and A. Ouin. Prediction of plant diversity in grasslands using sentinel-1 and-2 satellite image time series. *Remote Sensing of Environment*, 237:111536, 2020.

[8] Z. Gao, X. Shi, H. Wang, Y. Zhu, Y. B. Wang, M. Li, and D.-Y. Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

[9] S. Getzin, K. Wiegand, and I. Schöning. Assessing biodiversity in forests using very high-resolution images and unmanned aerial vehicles. *Methods in ecology and evolution*, 3(2):397–404, 2012.

[10] S. S. Hasan, L. Zhen, M. G. Miah, T. Ahamed, and A. Samie. Impact of land use change on ecosystem services: A review. *Environmental Development*, 34:100527, 2020.

[11] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] K.-R. W. Kladny, M. Milanta, O. Mraz, K. Hufkens, and B. D. Stocker. Deep learning for satellite image forecasting of vegetation greenness. *bioRxiv*, pages 2022–08, 2022.

[14] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

[15] A. Meraner, P. Ebel, X. X. Zhu, and M. 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.

[16] NASA. Hls tiling system (https://hls.gsfc.nasa.gov/wp-content/uploads/2016/03/MGRS_GZD-1.png), 2018.

[17] B. Peng, K. Guan, M. Pan, and Y. Li. Benefits of seasonal climate prediction and satellite data for forecasting us maize yield. *Geophysical Research Letters*, 45(18):9662–9671, 2018.

[18] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. 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.

[19] C. Requena-Mesa, M. Reichstein, M. Mahecha, B. Kraft, and J. Denzler. Predicting landscapes as seen from space from environmental conditions. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1768–1771. IEEE, 2018.

[20] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[21] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[22] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[23] Z. Yuan, X. Zhou, and T. Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 984–992, 2018.

[24] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.-E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino. Esa worldcover 10 m 2021 v200, Oct. 2022.

[25] D. Zanaga, R. Van De Kerchove, W. De Keersmaecker, N. Souverijns, C. Brockmann, R. Quast, J. Wevers, A. Grosu, A. Paccini, S. Vergnaud, O. Cartus, M. Santoro, S. Fritz, I. Georgieva, M. Lesiv, S. Carter, M. Herold, L. Li, N.-E. Tsendbazar, F. Ramoino, and O. Arino. Esa worldcover 10 m 2020 v100, Oct. 2021.

Use of ChatGPT (or any other AI Writing Assistance) – Form to be completed

Student name: **Jeroen Smets**

Student number: r0676599

Please indicate with "X" whether it relates to a course assignment or to the master thesis:

O This form is related to a **course assignment**.

Course name: ....

Course number: ....

X This form is related to **my Master thesis**.

Title Master thesis: An analysis of landscape predictions from a convolutional LSTM trained on the EarthNet2021 dataset

Promotor: Stef Lhermitte

Please indicate with "X":

O **I did not use** ChatGPT or any other AI Writing Assistance.

X **I did use** AI Writing Assistance. In this case **specify which one** (e.g. ChatGPT/GPT4/...):

ChatGPT

Please indicate with "X" (possibly multiple times) in which way you were using it:

X **Assistance purely with the language of the paper**

- *Code of conduct*: This use is similar to using a spelling checker

○ **As a search engine to learn on a particular topic**

- *Code of conduct*: This use is similar to e.g. a google search or checking Wikipedia. Be aware that the output of Chatbot evolves and may change over time.

○ **For literature search**

- *Code of conduct*: This use is comparable to e.g. a google scholar search. However, be aware that ChatGPT may output no or wrong references. As a student you are responsible for further checking and verifying the absence or correctness of references.

○ **For short-form input assistance**

- *Code of conduct*: This use is similar to e.g. google docs powered by generative language models

○ **To let generate programming code**

- *Code of conduct*: Correctly mention the use of ChatGPT and cite it. You can also ask ChatGPT how to cite it.

○ **To let generate new research ideas**

- *Code of conduct*: Further verify in this case whether the idea is novel or not. It is likely that it is related to existing work, which should be referenced then.

○ **To let generate blocks of text**

- *Code of conduct*: Inserting blocks of text without quotes from ChatGPT to your report or thesis is not allowed. According to Article 84 of the exam regulations in evaluating your work one should be able to correctly judge on your own knowledge. In case it is really needed to insert a block of text from ChatGPT, mention it as a citation by using quotes. But this should be kept to an absolute minimum.

○ **Other**

- *Code of conduct*: Contact the professor of the course or the promotor of the thesis. Inform also the program director. Motivate how you comply with Article 84 of the exam regulations. Explain the use and the added value of ChatGPT or other AI tool: ….

Further important guidelines and remarks

- ChatGPT cannot be used related **to data or subjects under NDA agreement.**
- ChatGPT cannot be used related **to sensitive or personal data due to privacy issues**.
- **Take a scientific and critical attitude** when interacting with ChatGPT and interpreting its output. Don’t become emotionally connected to AI tools.
- As a student you are responsible to comply with Article 84 of the exam regulations: your report or thesis should reflect your own knowledge. Be aware that plagiarism rules also apply to the use of ChatGPT or any other AI tools.
- **Exam regulations Article 84**: “Every conduct individual students display with which they (partially) inhibit or attempt to inhibit a correct judgement of their own knowledge, understanding and/or skills or those of other students, is considered an irregularity which may result in a suitable penalty. A special type of irregularity is plagiarism, i.e. copying the work (ideas, texts, structures, designs, images, plans, codes , ...) of others or prior personal work in an exact or slightly modified way without adequately acknowledging the sources. Every possession of prohibited resources during an examination (see article 65) is considered an irregularity.”
- **ChatGPT suggestion about citation**: “Citing and referencing ChatGPT output is essential to maintain academic integrity and avoid plagiarism. Here are some guidelines on how to correctly cite and reference ChatGPT in your Master's thesis: 1. Citing ChatGPT: Whenever you use a direct quote or paraphrase from ChatGPT, you should include an in-text citation that indicates the source. For example: (ChatGPT, 2023). 2. Referencing ChatGPT: In the reference list at the end of your thesis, you should include a full citation for ChatGPT. This should include the title of the AI language model, the year it was published or trained, the name of the institution or organization that developed it, and the URL or DOI (if available). For example: OpenAI. (2021). GPT-3 Language Model. https://openai.com/blog/gpt-3-apps/ 3. Describing the use of ChatGPT: You may also want to describe how you used ChatGPT in your research methodology section. This could include details on how you accessed ChatGPT, the specific parameters you used, and any other relevant information related to your use of the AI language model. Remember, it is important to adhere to your institution's specific guidelines for citing and referencing sources in your Master's thesis. If you are unsure about how to correctly cite and reference ChatGPT or any other source, consult with your thesis advisor or a librarian for guidance.”

Additional reading

ACL 2023 Policy on AI Writing Assistance: https://2023.aclweb.org/blog/ACL-2023-policy/

KU Leuven guidelines on citing and referencing Generative AI tools, and other information:https://www.kuleuven.be/english/education/student/educational-tools/generative-artificial-intelligence

Dit formulier werd opgesteld voor studenten in de Master of Artificial intelligence. Ze bevat een code of conduct, die we bij universiteitsbrede communicatie rond onderwijs verder wensen te hanteren.

Deze template samen met de code of conduct zal in de toekomst nog verdere aanpassingen behoeven. Het schept alvast een kader voor de 2de en de 3de examenperiode van 2022-2023.