Self-Supervised Transformers for Long-Term Prediction of Landsat NDVI Time Series

Ido Faran[1], Nathan S. Netanyahu[1,2], Elena Roitberg[3] and Maxim Shoshany[3]
[1]*Dept. of Computer Science, Bar-Ilan University, Ramat Gan 5290002, Israel*
[2]*Dept. of Computer Science, College of Law and Business, Ramat Gan 5257346 , Israel*
[3]*Faculty of Civil and Environmental Engineering, Technion Israel Institute of Technology, Haifa 3200003, Israel*

Keywords: Deep Learning, Transformers, Self-Supervised Learning, Remote Sensing.

Abstract: Long-term satellite image time-series (SITS) analysis presents significant challenges in remote sensing, especially for heterogeneous Mediterranean landscapes, due to complex temporal dependencies, pronounced seasonality, and overarching global trends. We propose Self-Supervised Transformers for Long-Term Prediction (SST-LTP), a novel framework that combines self-supervised learning, temporal embeddings, and a Transformer-based architecture to analyze multi-decade Landsat data. Our approach leverages a self-supervised pretext task to train Transformers on unlabeled data, incorporating temporal embeddings to capture both long-term trends and seasonal variations. This architecture effectively models intricate temporal patterns, enabling accurate predictions of the Normalized Difference Vegetation Index (NDVI) across diverse temporal horizons. Using Landsat data spanning 1984–2024, SST-LTP achieves a Mean Absolute Error (MAE) of 0.0338 and an R^2 value of 0.8337, outperforming traditional methods and other neural network architectures. These results highlight SST-LTP as a robust tool for long-term environmental monitoring and analysis.

1 INTRODUCTION

Image motion and sequence prediction have attracted significant attention in recent years (Verma et al., 2013; Mo et al., 2025). Time-series prediction aims to uncover temporal patterns that are often hidden in spatially complex scenes, where short, medium, and long-term processes occur and interact simultaneously. Self-supervised machine learning approaches offer unique advantages for such tasks. They operate without constraints or assumptions and, most importantly, do not require labeled data. By learning from past time series, these methods inherently capture the representation of "predicted images". Applying this technique to environmental time series is crucial for understanding ecosystems' responses to climatic and anthropogenic changes. Our study evaluates this approach in a desert fringe environment of the southeastern Mediterranean region, which is severely threatened by desertification.

Earth observation satellites have become invaluable tools for analyzing such dynamic environmental processes, collecting data about our planet's surface and ecosystems for over half a century. These platforms are crucial for monitoring global environmental changes, including vegetation patterns and long-term ecological trends. The Landsat TM mission, operational since 1984, has been pivotal in this field, enabling continuous monitoring of vegetation health, land use changes, and ecosystem dynamics across diverse landscapes. Landsat satellites capture multispectral imagery globally every 16 days at 30 [m] resolution. These images are organized into Satellite Image Time Series (SITS), which provide a temporal dimension to Earth observation data. SITS allows researchers to analyze changes over time, revealing patterns and trends that might be invisible by human visual interpretation. This temporal aspect is particularly valuable for land cover classification, change detection, and predictive modeling of global environmental trends (Zhu et al., 2019).

The primary objective of time- series analysis in remote sensing is to estimate future values accurately based on historical observations. These capabilities can be used to forecast future images, reconstruct missing data due to cloud cover or sensor malfunctions, and facilitate data fusion across multiple sources. Time-series analysis enhances the detection of abrupt and gradual changes in land cover and land use, providing crucial insights into long-term Earth surface processes (Gómez et al., 2016).

However, time-series analysis in remote sensing faces unique challenges. It requires consideration of complex temporal dependencies, including seasonal-

Faran, I., Netanyahu, N. S., Roitberg, E. and Shoshany, M.
Self-Supervised Transformers for Long-Term Prediction of Landsat NDVI Time Series.
DOI: 10.5220/0013381700003905

In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 542-552
ISBN: 978-989-758-730-6; ISSN: 2184-4313

ity in natural systems and abrupt changes in human activities. Global trends like climate change introduce gradual shifts that are difficult to distinguish from natural variability. Additionally, data inconsistencies due to cloud cover and sensor limitations complicate the development of robust predictive models. Addressing these challenges is crucial for accurate long-term environmental monitoring and change detection (Zhu, 2017; Kennedy et al., 2018). See Figure 1 for an example of Normalized Difference Vegetation Index (NDVI) time-series analysis across different locations.

We introduce Self-Supervised Transformers for Long-Term Prediction (SST-LTP), a novel approach for Landsat time-series analysis in diverse Mediterranean landscapes. This method combines self-supervised learning, temporal embedding techniques, and a Transformer-based architecture to predict NDVI values over both short-term (1–2 years) and long-term (5–10+ years) horizons. SST-LTP is built on three key components: (1) A self-supervised pretext task that trains the Transformer model to infer NDVI values from historical observations, (2) a temporal embedding strategy designed to capture persistent trends and seasonal patterns, and (3) a robust Transformer architecture optimized to handle complex temporal dependencies, long-range interactions, and seasonal variability. By leveraging the temporal dynamics inherent in satellite data, SST-LTP provides accurate and reliable predictions, addressing critical challenges in long-term environmental monitoring and land-use analysis.

Our main contributions are as follows:

1. Presentation of a self-supervised training method for long-term SITS data, capable of learning from unlabeled multi-decade data.
2. Introduction of a temporal embedding technique that captures both long-term trends and seasonal patterns to enhance the model's ability to make accurate predictions across different temporal horizons.
3. Experimental evaluation of our method's performance on Landsat data from Mediterranean regions, demonstrating its prediction capability of future short-term and long-term NDVI values based on varying lengths of historical data sequences.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work, covering traditional statistical methods, deep learning approaches, and self-supervised learning techniques for time-series analysis. Section 3 describes the proposed Self-Supervised Transformers for Long-Term Prediction (SST-LTP) framework, detailing its architecture, temporal embedding strategies, and self-supervised training methodology. Section 4 presents the experimental setup, including the study area, dataset, and implementation details, followed by an in-depth evaluation of the model's performance. Section 5 compares the proposed method with baseline models to highlight its advantages and limitations. Finally, Section 6 concludes the paper by summarizing the findings and outlining directions for future research.

2 RELATED WORK

2.1 Traditional Statistical Methods

Time-series prediction in remote sensing has traditionally relied on techniques such as Cellular Automata Markov Chain (CA-Markov), Random Forests (RFs), and Autoregressive Integrated Moving Average (ARIMA) models (Gómez et al., 2016). Additionally, models that explicitly incorporate seasonality, such as Seasonal Autoregressive Integrated Moving Average (SARIMA) (Box et al., 2015) (Yan et al., 2022) and Facebook Prophet (Taylor and Letham, 2018), have been widely utilized for temporal forecasting tasks in various domains, including vegetation index prediction and phenology analysis. These models are particularly adept at handling periodic patterns and trend decomposition but may struggle with the complex, non-linear relationships and missing data inherent in satellite imagery time series.

More recently, hybrid approaches combining traditional statistical methods with machine learning concepts have emerged. For example, hybrid SARIMA-ANN models have shown potential in leveraging the strengths of statistical seasonality modeling and data-driven learning (Ruiz-Aguilar et al., 2014).

2.2 Deep Learning for Time-Series Analysis

Advancements in deep learning have further revolutionized time-series analysis. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have demonstrated efficacy in capturing complex patterns in remote sensing time-series data (Zhu, 2017).

Transformers, initially developed for natural language processing (Vaswani et al., 2017), have also

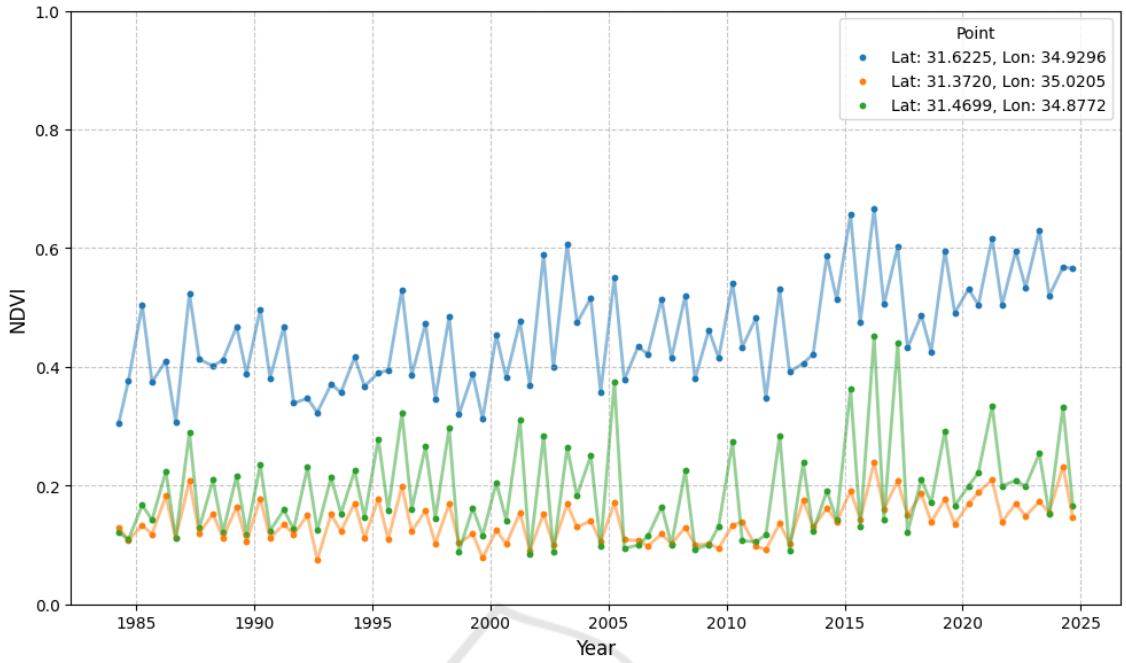

Figure 1: Time series of NDVI values from 1985 to 2024 for three samples. The plot highlights seasonal variability and long-term trends, showing NDVI growth across all samples, with the first sample exhibiting the most noticeable increase, indicating significant vegetation growth.

been adapted for time-series analysis due to their ability to model long-term dependencies effectively. Key advancements include the Informer model (Zhou et al., 2021), which introduces a sparse self-attention mechanism to improve scalability for long sequence time-series forecasting while maintaining the ability to capture complex temporal patterns. Similarly, the Temporal Fusion Transformer (TFT) by (Lim et al., 2021) provides an interpretable framework for multi-horizon time-series forecasting by integrating local and global context information with a focus on feature-level attention, emphasizing both scalability and interpretability.

The Crossformer model (Zhang and Yan, 2023) further enhances Transformer capabilities by addressing cross-dimension dependencies in multivariate time-series data, enabling more accurate modeling of interrelated features. Additionally, the iTransformer model (Liu et al., 2023) adopts an inverted Transformer architecture to better capture multivariate correlations with improved computational efficiency, highlighting the evolution of Transformer designs for time-series analysis.

In the domain of remote sensing, Transformers have shown significant potential for modeling spatiotemporal data. The Earthformer model (Gao et al., 2022) extends Transformer architectures by incorporating a cuboid attention mechanism, which segments data into smaller, manageable units for efficient spatiotemporal dependency modeling. This design enables an Earthformer to capture the intricate interactions between spatial and temporal dimensions in remote sensing tasks. Similarly, the RingMo foundation model (Sun et al., 2022) uses masked image modeling to bridge the gap between natural and remote sensing images, enhancing feature extraction and generalization. Building on this foundation, RingMoSense (Yao et al., 2023) introduces a triple-branch architecture for spatiotemporal evolution disentangling, enabling effective spatial and temporal pattern extraction for remote sensing applications.

Drawing on architectural innovations, the above Transformer models highlight their potential to handle remote sensing time-series data's complex, multi-dimensional nature, bridging the gap between general-purpose time-series analysis and domain-specific requirements. Indeed, such innovations have been critical for improving the scalability and accuracy of time-series forecasting, particularly in handling long-term dependencies and multi-modal inputs. This makes them especially relevant for remote sensing applications.

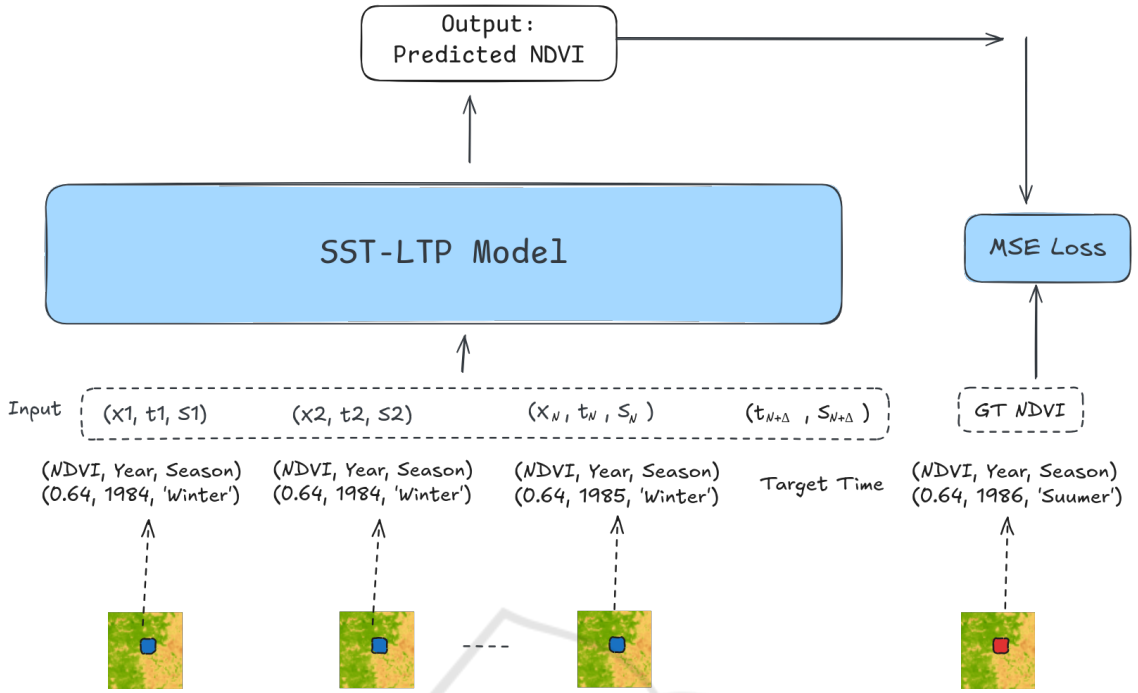

Figure 2: The training framework of the proposed Self-Supervised Transformers for Long-Term Prediction (SST-LTP) model. The input consists of a sequence of (NDVI, Year, Season) data for a single pixel across multiple timestamps and a specified target time. The model predicts the NDVI value for the target time, which is compared against the ground truth NDVI using MSE loss. Model weights are updated through backpropagation.

2.3 Self-Supervised Learning Approaches

Self-supervised learning has emerged as a promising approach to address one of the key challenges in remote sensing and satellite image analysis, i.e., the collection of labeled data. Unlike traditional datasets, labeling satellite imagery requires expert knowledge, is labor-intensive, and is often infeasible for large-scale, diverse geographical areas. Additionally, changes in land cover, climate, and sensor types can further complicate the creation of consistent labels across time and regions. Self-supervised learning mitigates this issue by leveraging abundant unlabeled data to create pseudo-supervised tasks (Miller et al., 2024). Recent advancements include the self-supervised training scheme for SITS classification (Yuan and Lin, 2020) and the Presto model (Tseng et al., 2023), a lightweight pre-trained transformer designed for pixel-time-series that leverages multi-modal data.

However, current self-supervised methods typically focus on shorter time sequences, e.g., one-year sequences of Sentinel-2 data (Yuan and Lin, 2020; Moskolaï et al., 2021), and are often restricted to agricultural areas (Rußwurm and Körner, 2018). Furthermore, many of these approaches rely on autoregressive prediction techniques, where the model predicts the next value in a sequence based on prior observations. While effective for short-term predictions, these methods face significant challenges when extended to long-term forecasting. Autoregressive models require iterative predictions to reach farther into the future, leading to error accumulation, as inaccuracies in earlier predictions propagate and compound over time. Additionally, this iterative process incurs high computational cost and time complexity, as the model must be repeatedly activated for each step in the sequence, making it inefficient for long-term analyses. These limitations are particularly pronounced in the context of analyzing long-term time series in heterogeneous Mediterranean landscapes, which involve complex seasonality, human-induced changes, climate trends, sensor variations, and data quality issues (Zhu et al., 2019).

While these advancements have propelled time-series analysis in remote sensing, significant challenges remain when aiming to capture extended historical ranges and the complex seasonality inherent in remote sensing data. This work addresses these gaps by introducing a method specifically tailored to long-term NDVI data from a Mediterranean setting. Our approach learns inherent temporal patterns directly

from the data, enabling the capture of persistent trends and seasonal cycles across multiple decades. Unlike prior methods that emphasize short-duration sequences or broader, homogeneous regions, our framework focuses on modeling the extended-range evolution of a single, heterogeneous landscape.

In contrast to autoregressive methods that rely on iterative predictions to extend into the future, our approach avoids repeated model activations by directly forecasting long-term temporal patterns in a single step. This design minimizes computational overhead and avoids the error accumulation typical of autoregressive techniques. By focusing on domain-specific temporal embeddings and leveraging a robust architecture, this work advances the understanding of how ecosystems transform over extended periods, addressing critical gaps left by existing methods in time-series analysis.

3 PROPOSED METHOD

Figure 2 illustrates our proposed training method, which treats time-series prediction as a self-supervised learning task. We leverage the temporal nature of long-term satellite imagery to train a deep learning model without explicitly labeled data. The process involves feeding the model with a sequence of past satellite images, from which it predicts subsequent (NDVI) values. These predictions are then compared against the actual observed values from the time-series data. The model is subsequently trained to minimize the difference between its predictions and the true values. This approach harnesses the inherent temporal structure of satellite imagery, enabling the model to learn patterns and trends without manual labeling. The model refines its forecasting capabilities over time by continuously predicting and adjusting future states based on real observations.

Formally, we define our input as a time-series sequence

$$O = \{(x_1, t_1, s_1), \ldots, (x_N, t_N, s_N)\} \tag{1}$$

for a single pixel, where N is the number of observations. Each tuple (x_i, t_i, s_i) represents an observation, where x_i is the NDVI value, t_i is the year, and $s_i \in \{\text{“Winter”}, \text{“Summer”}\}$ is the season. Our model's task is to predict the NDVI value for a specified future time point, defined by $N+\Delta$, where $\Delta > 0$. The prediction can be expressed as

$$\hat{x}_{N+\Delta} = f(O, t_{N+\Delta}, s_{N+\Delta}) \tag{2}$$

where f is our deep learning model that learns to forecast NDVI values based on past patterns and the desired future time point.

To train the model, we utilize the inherent temporal structure of the satellite imagery data. The model learns to predict future NDVI values based on the sequence of past observations. We compare the model's predictions $\hat{x}_{N+\Delta}$ against actual NDVI values $x_{N+\Delta}$ using a loss function $L(\hat{x}_{N+\Delta}, x_{N+\Delta})$. We update the model's parameters through iterative backpropagation to minimize this loss. This process continues, progressively improving the model's ability to capture and forecast NDVI patterns over time.

Figure 3 illustrates the architecture of the Transformer-based deep learning model. It consists of the following three main parts: (1) Observation Embedding, (2) Transformer Encoder, and (3) Regression Decoder.

3.1 Observation Embedding

The observation embedding layer projects the input sequence $\{(x_1, t_1, s_1), ..., (x_N, t_N, s_N)\}$ into a higher-dimensional feature space, preserving intrinsic data relationships. This embedding comprises three components:

1. NDVI Embedding: A linear dense layer projects the NDVI sequence $X_1, ..., X_N$ into a high-dimensional vector space.
2. Temporal Encoding: A continuous embedding space represents years and seasons, ensuring that temporally close points have similar representations while accounting for seasonal cycles. For a time point with year t and season s, we compute a normalized time value
$$t' = \frac{2(t - t_{\text{start}}) + s}{2(t_{\text{end}} - t_{\text{start}} + 1)} \tag{3}$$
where t_{start} and t_{end} are the dataset's temporal bounds, and $s \in \{0, 1\}$ denotes the season. The final temporal embedding is calculated via:
$$E(t') = \big[t', \sin(2\pi t'), \cos(2\pi t'), \ldots, \sin(2\pi k t'), \cos(2\pi k t')\big] \tag{4}$$
where d is the dimensional encoding and $k = \lfloor (d-1)/2 \rfloor$.
3. Positional Encoding (PE): The temporal order of years and seasons $(t_1, s_1), ..., (t_N, s_N)$ is encoded via PE (Devlin, 2018).

The final observation embedding O_i for each time point i is the element-wise sum of these components:

$$O_i = \text{NDVI}_i + \text{PE}_i + E(t_i) \tag{5}$$

This formulation captures both long-term trends and seasonal patterns in NDVI data, with proximate time points having similar representations in the embedding space.

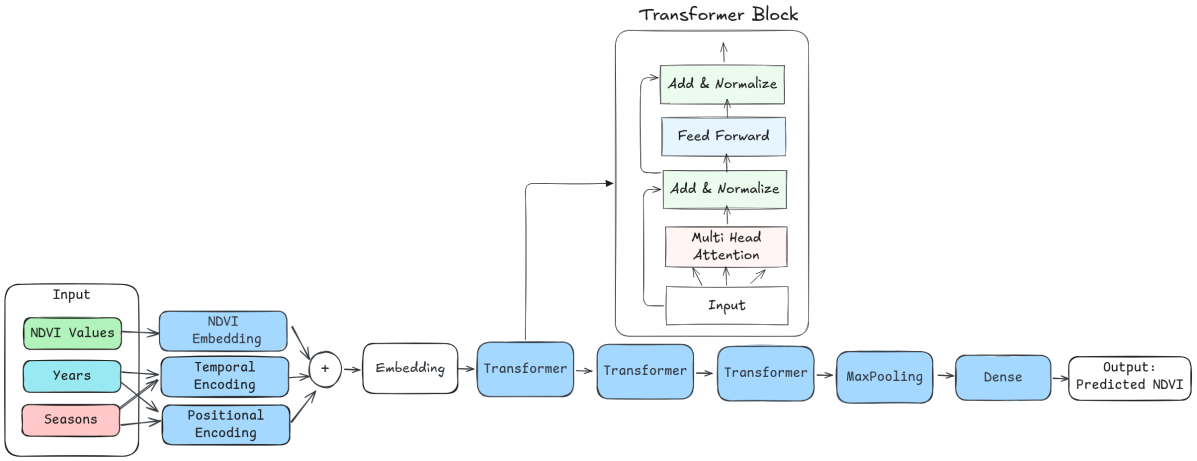

Figure 3: Architecture of the Self-Supervised Transformers for Long-Term Prediction (SST-LTP) model. Input (NDVI, Year, Season) is processed through three embedding channels: NDVI Embedding (to capture NDVI patterns), Temporal Encoding (to model year/seasonal patterns), and Positional Encoding (to represent sequence order). The combined embedding passes through Transformer blocks, followed by MaxPooling and a dense layer for aggregation and final NDVI prediction.

3.2 Transformer Encoder

The embedded time series is processed through stacked Transformer blocks, similar to the BERT architecture (Devlin, 2018), employing multi-head attention mechanisms. Each block generates progressively higher-level representations, building upon the output of the previous block. This iterative process yields encoded features that capture local and global temporal dependencies, effectively representing the complex patterns in the NDVI and temporal data across the entire sequence.

3.3 Regression Decoder

The output from the Transformer encoder is processed through a regression decoder to predict the NDVI for the target year and season. This decoder employs two key components: A MaxPooling layer and a Dense (linear) layer. The MaxPooling layer aggregates the most important features across the temporal dimension, reducing the sequence to a single vector representation. This pooled vector is then fed into the Dense layer, which maps it to a single scalar value representing the predicted NDVI.

4 EXPERIMENTAL RESULTS

4.1 Study Area

The study area is located in the southeastern corner of the Mediterranean basin, along a gradient transitioning from Mediterranean to arid climate zones (see

Figure 4: Study area in the southeastern Mediterranean basin, illustrating the transition from Mediterranean to arid climate zones, as visualized using Google Earth.

Figure 4). The rainfall varies between 450 mm/year to 250 mm/year resulting in a transition from shrublands to phrygana (Bata) to bare desert. Frequent annual rainfall fluctuations, droughts, and human disturbance to natural ecosystems create a mosaic of highly-variable vegetation, soil, and rock patterns.

Temporal landscape changes are significantly influenced by fires and periods of low rainfall. As shown in Figure 1, the NDVI time series recorded in the study area exhibit distinctive annual and seasonal fluctuations, which pose challenges to predicting future NDVI maps based on past NDVI sequences (Roitberg and Shoshany, 2024; Mozhaeva and Shoshany, 2022).

We evaluated the proposed method using Landsat images (Missions 5, 7, 8, and 9) from 1984 to 2024 over Israel. These data represent Mediterranean regions with high spatial and temporal variability, characterized by long dry spells and short, intense rainfalls (Faran et al., 2020).

The dataset spans an area of 53.94×37.35 km^2, represented by a scene of 1798×1245 pixels, with a spatial resolution of 30 [m] per pixel. It was obtained from the Google Earth Engine L2 products, with two seasonal composites created annually by averaging NDVI values from scenes with less than 20% cloud cover (October–April for "Winter" and May–September for "Summer"). The resulting dataset, comprising a total of 2,238,510 pixels, was divided into 80% (1,790,808 samples) for training, 10% (223,851 samples) for validation, and 10% (223,851 samples) for testing. A windowing approach was applied to extract training sequences, where each window consisted of N consecutive NDVI values (e.g., $N = 10$ for 5 years of past data with two seasons each), with one additional value serving as the target timestamp to predict.

4.2 Implementation and Parameters

The model employs an embedding dimension of 256, followed by three encoding Transformer blocks, each with eight attention heads. A dropout rate of 0.2 was applied after the embedding layer and each Transformer block. The training was conducted over 200 epochs using an initial learning rate of 1×10^{-4}, with a 10-epoch warm-up period followed by an exponential decay. The mean square error (MSE) or L_2 loss served as the objective function, optimized using the Adam optimizer. Model performance was evaluated using MSE, Mean Absolute Error (MAE) or L_1, and the coefficient of determination, R^2.

The dataset and implementation code are available in a public repository*.

4.3 Time-Series Evaluation

We evaluated the proposed model's prediction capabilities, using various N and Δ values for past observation lengths and future time horizons, respectively. This also helped assess our method's capability of capturing time-series patterns and seasonality. The results are presented in Figure 5. A past observation sequence of 10 years yielded the most significant improvement in prediction accuracy, with noticeable gains compared to shorter sequences. Beyond 10 years, further gains are less pronounced, with 15 to 20 years offering the lowest L_1 loss. The model performs best for next-year predictions, with performance gradually declining as the prediction horizon increases. Separating the analysis by land-cover class might yield different results, as some classes (e.g., building areas, bare ground) have static NDVI values.

*https://github.com/FaranIdo/SST-LTP

Table 1: Comparative performance of previous methods and our SST-LTP model for next-year prediction using 10-year input sequences. The proposed SST-LTP model outperforms all other methods examined.

Model	L_1 **(MAE)**	L_2 **(MSE)**	$\mathbf{R^2}$
SVM	0.0456	0.0041	0.7459
CNN-1D	0.0391	0.0035	0.7810
Fully Connected	0.0383	0.0033	0.7953
LSTM	0.0363	0.0031	0.8088
SST-LTP	**0.0338**	**0.0027**	**0.8337**

In contrast, other classes (e.g., grass, shrub) exhibit more temporal variability.

Figure 6 demonstrates the model's output using an input sequence of 10 years, applied to a sample of 12 points. As shown, the model successfully captures the overall trends in NDVI values over time, aligning with the patterns observed in the input sequence. However, it struggles to predict unexpected outliers, which may arise from sudden events or changes in the environment. These deviations highlight the challenges of modeling abrupt anomalies in a predominantly trend-focused framework.

4.4 Comparison with Other Models

We benchmarked our proposed model against various spectral time-series analysis methods, including Support Vector Machines (SVM), Fully-Connected Neural Networks, 1D Convolutional Neural Networks (CNN-1D), and Long Short-Term Memory (LSTM) networks. The SVM model was implemented with a Radial Basis Function (RBF) kernel, a commonly used configuration for time-series regression. The Fully-Connected Neural Network consisted of three layers with a hidden size of 128 neurons each and ReLU activation, tailored to process the 10-sample (i.e., a 5-year) input sequence. For the CNN-1D model, we utilized three convolutional layers, each with 64 filters and a kernel size of 3, along with padding to preserve sequence length, followed by a linear prediction layer. The LSTM network was designed with a 3-layer stacked architecture, each with 64 hidden units, to capture hierarchical temporal dependencies in the data.

The results presented in Table 1 demonstrate the superior performance of our proposed model. A key distinction is our model's ability to incorporate the target year as an input parameter, to obtain multiyear predictions directly. In contrast, traditional methods typically predict only the next item in the sequence, requiring repeated inference calls for long-term predictions. This architectural advantage of our model, which allows direct future-year predictions without

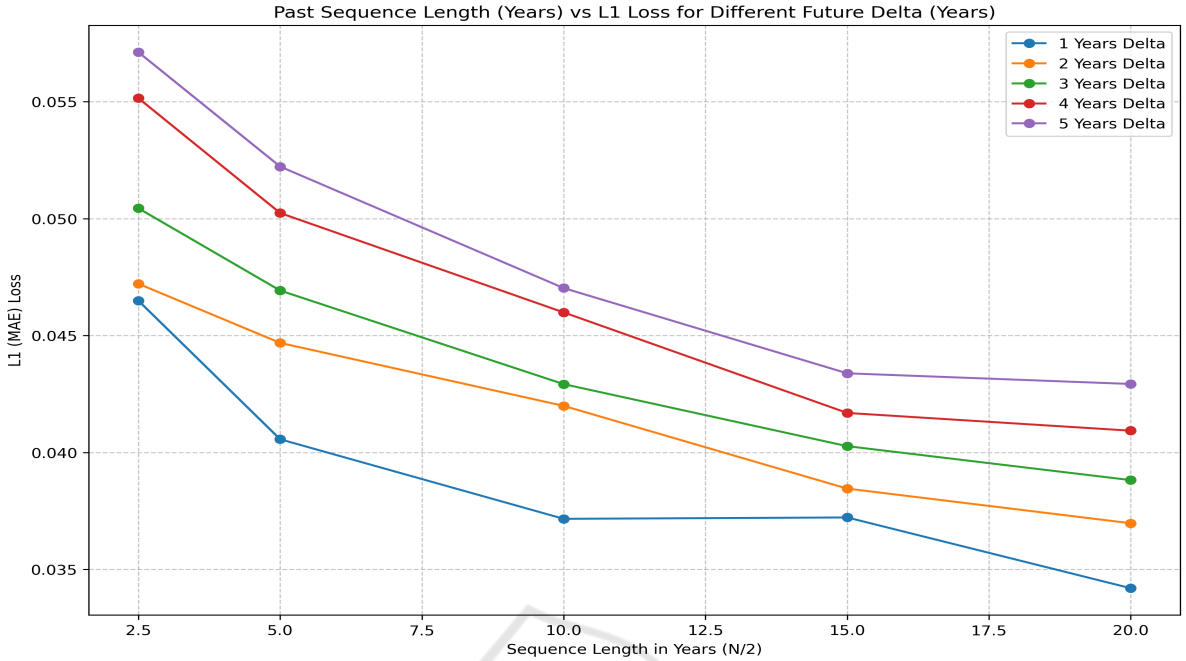

Figure 5: L_1 loss vs. past sequence length for different future prediction horizons; x-axis represents the duration of historical data used for prediction (i.e., 2.5–20 years), and y-axis displays L_1 (MAE) loss. Each curve corresponds to a different future prediction horizon (between 1–5 years). The graph demonstrates that longer historical sequences generally improve prediction accuracy, with diminishing returns beyond 10 years; shorter prediction horizons yield lower L_1 loss across all sequence lengths.

sequential inference and improved accuracy, highlights the efficiency and effectiveness of our approach for long-term time-series forecasting.

Figure 7 further illustrates the comparative performance of the different models for three representative NDVI samples, using a 10-step (i.e., a 5-year) input sequence. The plots show the ground truth (blue), input sequences (dashed green), and predictions from each model. Notably, the proposed SST-LTP model consistently tracks the trends of the ground truth better than the other methods, especially in areas with higher variability. In contrast, the Fully-Connected and CNN models tend to exhibit greater deviations, particularly in regions with more complex temporal patterns. The LSTM predictions show some alignment with ground truth but appear smoother, potentially due to limitations in capturing fine-grained fluctuations over longer horizons. These results emphasize the ability of our proposed model to adapt to intricate patterns in the data while maintaining accuracy across different time-series samples.

5 CONCLUSION

This study introduces a novel approach for analyzing long-term satellite image time series, leveraging self-supervised learning, temporal embeddings, and a Transformer-based architecture. The proposed Self-Supervised Transformers for Long-Term Prediction (SST-LTP) framework excels at modeling complex temporal dynamics and seasonal variations in Mediterranean landscapes. Experimental results using Landsat data from Mediterranean regions demonstrate the method's strong performance for both short-term and long-term predictions, outperforming traditional statistical and neural network-based models.

Future research will focus on integrating spatial correlations between adjacent pixels, adapting the framework for change detection and time-series classification tasks, and incorporating precipitation data to enhance the understanding of seasonal patterns. Additionally, techniques for handling missing data will be developed, and the model will be evaluated across diverse land-cover types to assess its adaptability to varying landscapes.

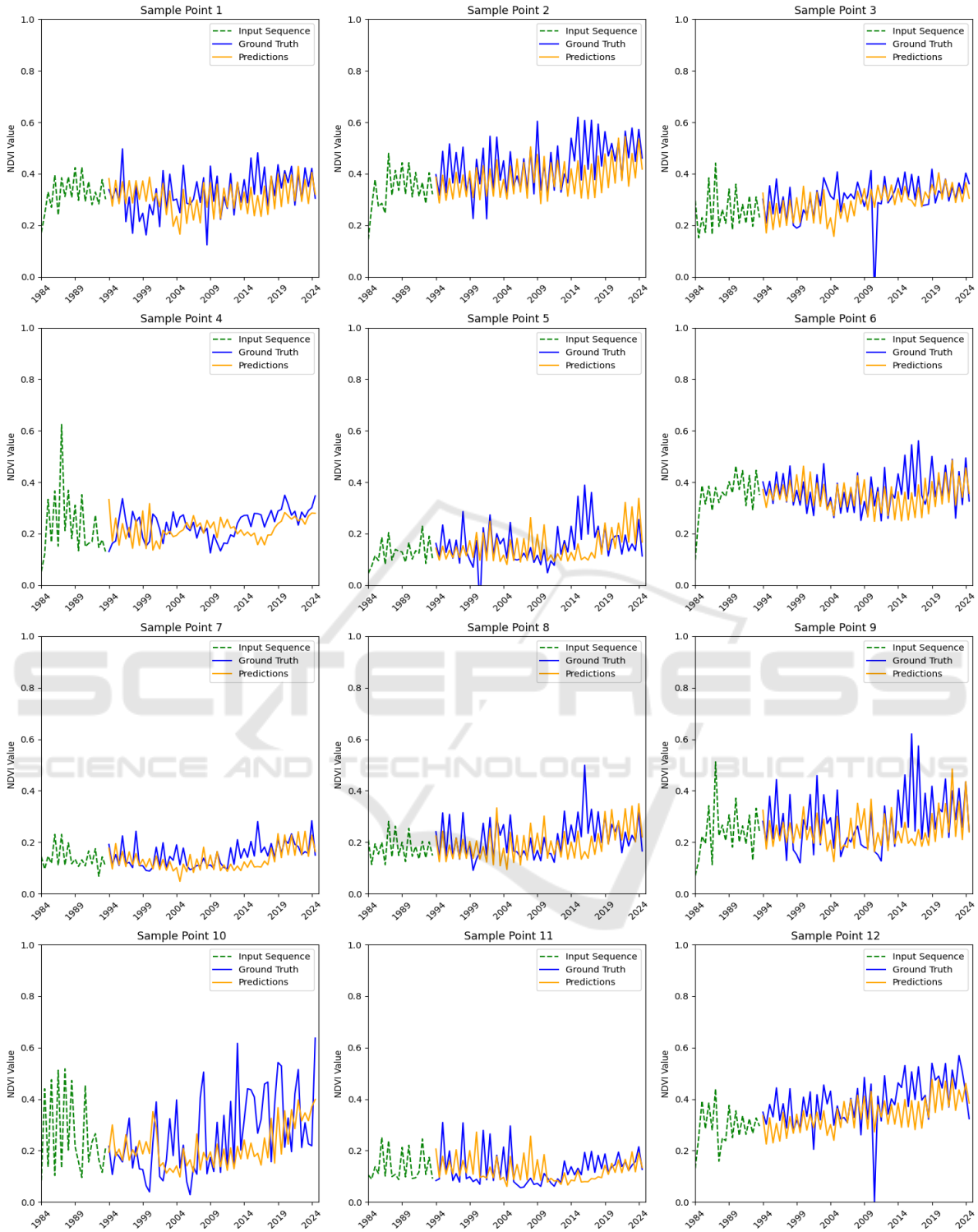

Figure 6: Demonstration of the SST-LTP model's performance using an input sequence of 20 samples (i.e., 10 years) across 12 sample points. Each subplot compares the input sequence, ground truth, and SST-LTP model predictions, illustrating the model's ability to capture seasonal variability and long-term trends in NDVI values. The results highlight the SST-LTP model's effectiveness in accurately predicting NDVI values across diverse temporal patterns.

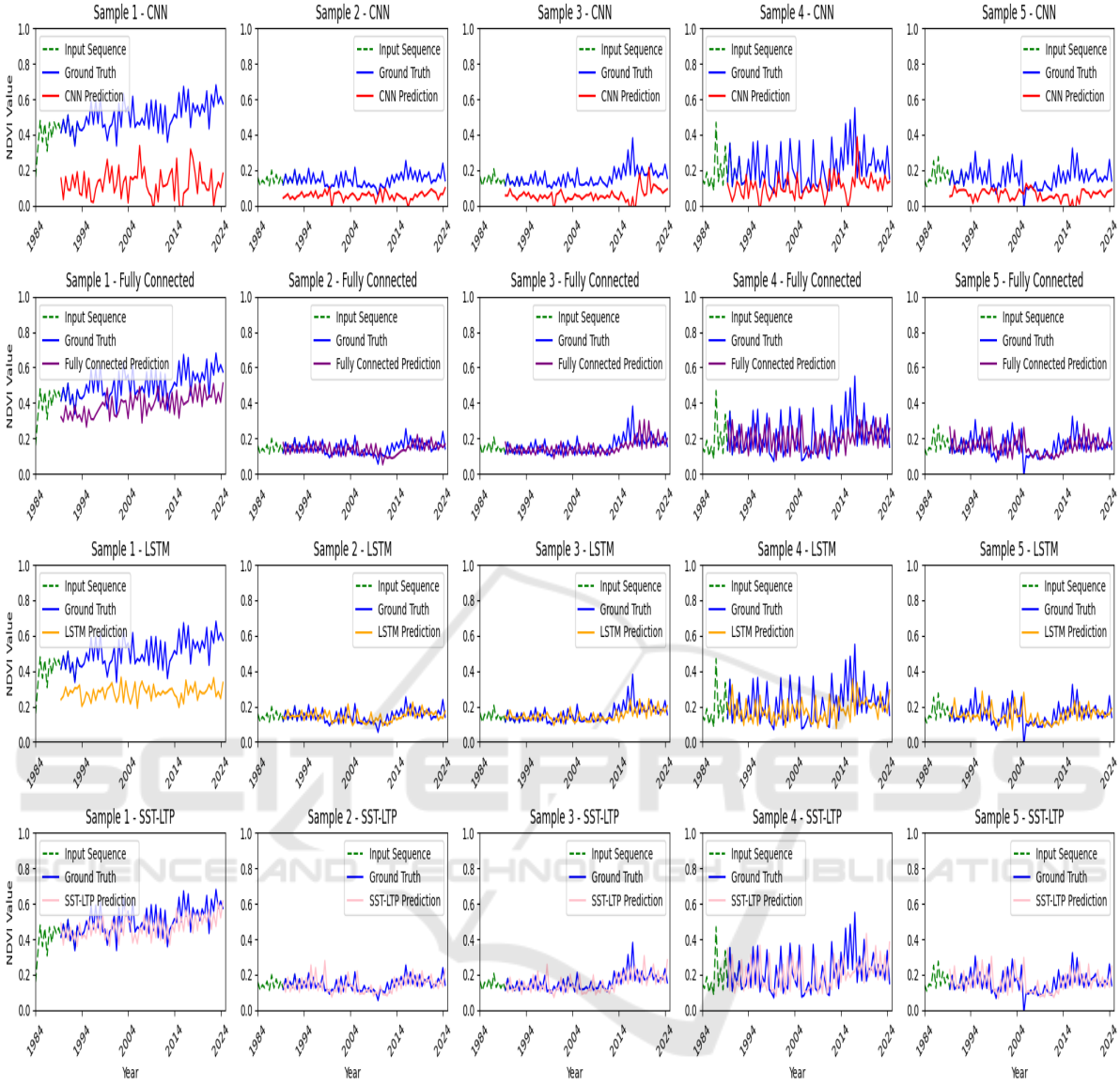

Figure 7: Performance evaluation of SST-LTP and other deep learning methods (i.e., CNN, Fully Connected, and LSTM) on five NDVI samples for 10-step (i.e., 5-year) input sequence. Each subplot illustrates the input sequence, ground truth, and predicted values for various models. SST-LTP predictions demonstrate improved alignment with the ground truth, highlighting its effectiveness in capturing long-term trends and variability.

REFERENCES

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Devlin, J. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Faran, I., Netanyahu, N. S., David, E., Rud, R., and Shoshany, M. (2020). Multi seasonal deep learning classification of VENuS images. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 6754–6757.

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

Gómez, C., White, J. C., and Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *Journal of Photogrammetry and Remote Sensing*, 116:55–72.

Kennedy, R. E., Andréfouët, S., Cohen, W. B., Gómez, C., Griffiths, P., Hais, M., et al. (2018). Bringing an ecological view of change to Landsat-based remote sensing. *Frontiers in Ecology and the Environment*, 16(6):340–348.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. (2023). iTransformer: Inverted Transformers are effective for time series forecasting. *arXiv:2310.06625*.

Miller, L., Pelletier, C., and Webb, G. I. (2024). Deep learning for satellite image time series analysis: A review. *arXiv:2404.03936*.

Mo, F., Huang, Y., Wu, M., Zhu, X., and Zhang, C. (2025). Mmsisp: A satellite image sequence prediction network with multi-factor decoupling and multi-modal fusion. *Pattern Recognition*, 221–236.

Moskolaï, W. R., Abdou, W., Dipanda, A., and Kolyang (2021). Application of deep learning architectures for satellite image time series prediction: A review. *Remote Sensing*, 13(23):4822.

Mozhaeva, S. and Shoshany, M. (2022). Relationships between vegetation indices and rainfall and PET at different time-lags: A study at a Mediterranean to arid gradient. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:939–944.

Roitberg, E. and Shoshany, M. (2024). Primary productivity and woody growth: a 35 years Landsat TM NDVI time series investigation across desert-fringe in the south-eastern Mediterranean. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 2867–2870.

Ruiz-Aguilar, J., Turias, I., and Jiménez-Come, M. (2014). Hybrid approaches based on sarima and artificial neural networks for inspection time series forecasting. *Transportation Research Part E: Logistics and Transportation Review*, 67:1–13.

Rußwurm, M. and Körner, M. (2018). Self-attention for raw optical satellite time series classification. *Journal of Photogrammetry and Remote Sensing*, 169:421–435.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., He, Q., Yang, G., Wang, R., Lu, J., and Fu, K. (2022). RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15.

Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.

Tseng, G., Cartuyvels, R., Zvonkov, I., Purohit, M., Rolnick, D., and Kerner, H. (2023). Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv:2304.14065*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Keiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Verma, N. K., Bansal, A., and Singh, S. (2013). Generation of future image frames for an image sequence. In *International Conference on Intelligent Interactive Technologies and Multimedia*, pages 154–162. Springer.

Yan, B., Mu, R., Guo, J., Liu, Y., Tang, J., and Wang, H. (2022). Flood risk analysis of reservoirs based on full-series arima model under climate change. *Journal of Hydrology*, 610:127979.

Yao, F., Lu, W., Yang, H., Xu, L., Liu, C., Hu, L., Yu, H., Liu, N., Deng, C., Tang, D., Chen, C., Yu, J., Sun, X., and Fu, K. (2023). RingMo-Sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–21.

Yuan, Y. and Lin, L. (2020). Self-supervised pretraining of transformers for satellite image time series classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:474–487.

Zhang, Y. and Yan, J. (2023). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115.

Zhu, Z. (2017). Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *Journal of Photogrammetry and Remote Sensing*, 130:370–384.

Zhu, Z., Zhang, J., Yang, Z., Aljaddani, A. H., Cohen, W. B., Qiu, S., and Zhou, C. (2019). Continuous monitoring of land disturbance based on Landsat time series. *Remote Sensing of Environment*, 238:111116.