GLOBAL VEGETATION MODELING WITH PRE-TRAINED WEATHER TRANSFORMERS

Pascal Janetzky, Florian Gallusser, Simon Hentschel, Andreas Hotho, Anna Krause
Data Science Chair, Center for Artificial Intelligence and Data Science (CAIDAS),
University of Würzburg
{janetzky,gallusser,hentschel,hotho,
anna.krause}@informatik.uni-wuerzburg.de

ABSTRACT

Accurate vegetation models can produce further insights into the complex interaction between vegetation activity and ecosystem processes. Previous research has established that long-term trends and short-term variability of temperature and precipitation affect vegetation activity. Motivated by the recent success of Transformer-based Deep Learning models for medium-range weather forecasting, we adapt the publicly available pre-trained FourCastNet to model vegetation activity while accounting for the short-term dynamics of climate variability. We investigate how the learned global representation of the atmosphere's state can be transferred to model the normalized difference vegetation index (NDVI). Our model globally estimates vegetation activity at a resolution of $0.25°$ while relying only on meteorological data. We demonstrate that leveraging pre-trained weather models improves the NDVI estimates compared to learning an NDVI model from scratch. Additionally, we compare our results to other recent data-driven NDVI modeling approaches from machine learning and ecology literature. We further provide experimental evidence on how much data and training time is necessary to turn FourCastNet into an effective vegetation model. Code and models are available at https://github.com/LSX-UniWue/Global-Ecosystem-Modeling.

1 INTRODUCTION

Environmental changes affect the dynamics of terrestrial vegetation, which is involved in controlling water, energy and CO_2 fluxes (Richardson et al., 2013), and is thus crucial for providing ecosystem services such as food, fiber and fuel (Piao et al., 2020). Hence, a profound understanding of the complex interplay of climate system variables and vegetation changes is desirable to achieve sustainable ecological management.

Previous studies have shown that observed changes in vegetation can be attributed to both long-term and short-term changes in temperature and precipitation, i.e., climate change and climate variability (Burrell et al., 2020; Chen et al., 2019; Higgins et al., 2023; Liu et al., 2022; Seddon et al., 2016; Zhu et al., 2016). While the spatial arrangement of vegetation on a large scale is primarily dictated by climatic factors, the interplay between climate variability and the short-term dynamics of vegetation introduces a higher level of complexity (Papagiannopoulou et al., 2017; Pelletier et al., 2015). Different Machine Learning (ML) approaches have been suggested to capture the complex nonlinear dynamics of those short-term dynamics. However, the employed models are either limited to a specific region (Robin et al., 2022; Smith et al., 2023) or use a coarse global resolution with one pixel covering at least $0.5°(\approx 55\,\mathrm{km})$ (Chen et al., 2021; Kraft et al., 2019). While there are statistical approaches that globally quantify the effect of climate variability on vegetation change on a finer spatial resolution up to $0.083°$, they only consider meteorological data on a coarse time scale, *e.g.*, one data point per month (Burrell et al., 2020; Seddon et al., 2016). Nonetheless, the availability of long-term weather reanalysis datasets such as ERA5 (Hersbach et al., 2020), which comprises hourly high-resolution measurements of $0.25°$ ($\approx 27\,\mathrm{km}$) per pixel, provide the opportunity to model dependencies of short-term changes in meteorological variables on vegetation activity on a fine spatial and temporal resolution.

Recently, Deep Learning (DL) models have demonstrated the capability to efficiently parse and exploit those vast amounts of meteorological data in the context of medium-range weather forecasting. Architectural improvements and increased compute availability have led to DL-based weather models that now perform on par with commonly used numerical weather systems (Bi et al., 2023; Lam et al., 2023; Pathak et al., 2022). These approaches learn a spatial representation of the atmosphere's state by forecasting future atmospheric states. Previous studies have already shown that their trained atmospheric models can be finetuned to effectively solve other climate-related tasks such as statistical downscaling and climate projections (Lessig et al., 2023; Nguyen et al., 2023).

Based on these advances, this work investigates how the pre-trained weather forecasting model FourCastNet (FCN) (Pathak et al., 2022) can be adapted for globally modeling the normalized difference vegetation index (NDVI) (Tucker & Sellers, 1986; Vermote, 2019), a commonly used index for approximating vegetation activity (Ferchichi et al., 2022). We outline an approach building upon a state-of-the-art DL architecture for processing spatio-temporal data, which enables global modeling of the NDVI at a high spatial ($0.25°$) and temporal (daily) resolution with a single model. We investigate how to utilize FCN's atmospheric knowledge by comparing a finetuned model versus a model trained from scratch. Additionally, we analyze the training time and data needed to make FCN an effective vegetation model in three ablation studies.

2 Pre-trained weather models for vegetation modeling

Dataset For our study, we use daily global weather data from ERA5 (Hersbach et al., 2020) at a resolution of $0.25°$ (720 x 1440 pixel) from the years 1982 to 2013. We use the same 20 predictor variables as Pathak et al. (2022): zonal and meridional wind velocity ($10\,\mathrm{m}$ above ground, at $1000\,\mathrm{hPa}$, $850\,\mathrm{hPa}$ and $500\,\mathrm{hPa}$), temperature ($2\,\mathrm{m}$ above ground, at $850\,\mathrm{hPa}$ and $500\,\mathrm{hPa}$), geopotential (at $1000\,\mathrm{hPa}$, $850\,\mathrm{hPa}$, $500\,\mathrm{hPa}$, and $50\,\mathrm{hPa}$), relative humidity (at $850\,\mathrm{hPa}$ and $500\,\mathrm{hPa}$) surface pressure, mean sea level pressure, and total column water vapor. One sample has the dimensionality (20 x 720 x 1440). The NDVI data (Vermote, 2019) is our target variable, regridded linearly from originally $0.083°$ to ERA5's $0.25°$ resolution. This vegetation index is computed from satellite observations as the normalized difference between the spectral reflectances in the near-infrared and red wavebands (Tucker & Sellers, 1986). It ranges from -1 to 1. Negative values indicate water, positive values around zero indicate barren land, and values close to one indicate dense vegetation. NDVI data after 2013 was not considered for this study, as our analysis (*cf.* Fig. 5) shows a noticeable data shift beginning in 2014

Method To investigate the applicability of pre-trained weather models for globally modelling vegetation activity, we use the FCN Deep Learning model, whose pre-trained weights are publicly available (Pathak et al., 2023). FCN is a comparatively lightweight weather model (*cf.* Bi et al. (2023); Chen et al. (2023); Lessig et al. (2023); Nguyen et al. (2023)) with a total of 73 million parameters distributed over 12 Transformer-like (Vaswani et al., 2017) encoder blocks, see Fig. 3 for an architectural overview. Each of these blocks has 5 million parameters and uses an Adaptive Fourier Neural Operator layer (Guibas et al., 2021) replacing the attention mechanism (Bahdanau et al., 2015). We adopt the FCN to the NDVI modeling task by replacing the original weather prediction head with a randomly initialized fully-connected layer with the $\tanh$ activation function. For modeling the effects of short-term climate variability on vegetation activity, FCN is trained on modeling the NDVI for the same timestep as the daily input weather variables. For finetuning, we initialize the original FCN model with the pre-trained weights, while for training FCN from scratch, we freshly initialize all model weights.

Comparison models As a simple baseline, we designed and hyperparameter-optimized a convolutional neural network (CNN) with the details given in Appendix C. We further compare us with two recent data-driven models from ecology literature: The first approach is a global long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) model by Kraft et al. (2019), trained on single-location time-series of meteorological variables at a $15\,\mathrm{d}$ temporal resolution and a $0.5°$ spatial resolution, half of our resolution, with globally shared weights. This approach reflects the so-called memory effect of vegetation, i.e., that preceding vegetation states can have longer-lasting effects on vegetation activity (De Keersmaecker et al., 2015). The second data-driven approach

Table 1: Test year results. Left: Latitude weighted global evaluation of the finetuned FCN, FCN trained from scratch, our baseline CNN and the LSTM (Kraft et al., 2019). Right: Unweighted averages for local evaluation on 100 locations for comparison with Higgins et al. (2023). †: local model with global weight-sharing with different variables at $0.5°$ resolution. ‡: local models with different variables. See Section 2 and Appendix C for details.

Model	FCN finetune	FCN scratch	CNN	LSTM†	FCN finetune	SSM‡
Evaluation	global, 15-daily				local, 7-daily	
RMSE	0.0403	0.0512	0.0431	0.017	0.0547	0.0548
R^2	0.6331	0.4977	0.6061	0.904	0.5151	0.4038

trains separate local, weekly state space model (SSM) on 100 locations across the globe, guided by equations describing the interplay between the used climate-forcing data (Higgins et al., 2023).

Experimental setup The dataset is split into a training (1982-2010), validation (2011-2012), and test (2013) set. Training details for FCN can be found in Appendix B. We perform three individual ablation studies: In study I, we vary the number of finetuning epochs between 1 to 200 epochs. The number of frozen parameters during finetuning is varied in study II, where we freeze between one and eight of FCN's Transformer blocks in ascending order. Lastly, in study III, the number of finetuning data is modified by selecting a random $10\,\%$ to $90\,\%$ subset of the training years.

Evaluation We evaluate all models by computing the root mean squared error (RMSE) and R^2 score on the test set. The R^2 score measures the goodness of fit of the model proportional to the temporal variation of the target NDVI values at a given pixel and ranges from 1 (best) to $-\infty$. For global evaluation in comparison with the LSTM results reported by Kraft et al. (2019), we match their evaluation scheme, thereby removing the same noisy pixels and accounting for varying pixel areas through latitude-weighting (*cf.* Appendix D). Also, our models' outputs and target values are aggregated to 15-day averages to match their temporal resolution. For local evaluation in comparison with the SSMs results provided by Higgins et al. (2023), we evaluate our model locally at the same 100 locations. At these locations, FCN model outputs and target values are aggregated to 7-day averages as in Higgins et al. (2023), while evaluating the SSMs exclusively on the 2013 test year.

3 Results and Discussion

Finetuning the learned atmospheric representation of FCN for vegetation modeling outperforms an NDVI model trained from scratch, as Table 1 shows. Here, the scratch model reaches an R^2 of 0.4977 (RMSE: 0.0512). Finetuning the same model strongly improves NDVI modeling performance up to an R^2 of 0.6331, which is higher than the strong hyperparameter-optimized CNN

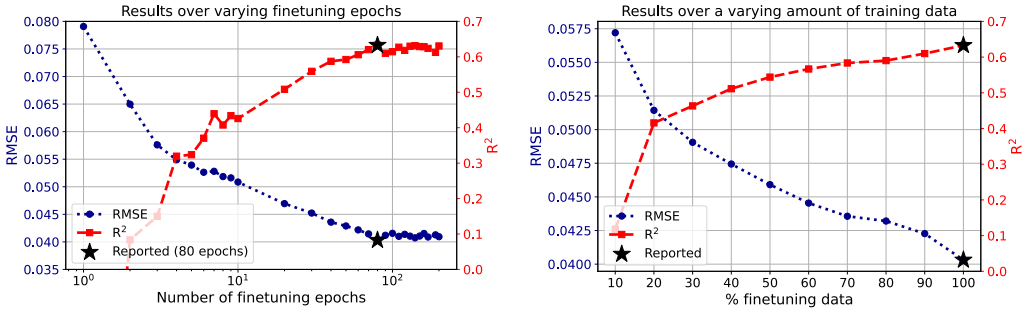

Figure 1: Results for ablation studies I & II. Left: varying number of finetuning epochs. Right: Varying amount of training data. Results reported in Table 1 are highlighted in both plots.

baseline with an R^2 of 0.6061 (RMSE: 0.0431), For the FCN, a parameter search was not considered in this study but could enhance the global ecosystem model further.

To contextualise our performance, the pixel-wise LSTM model by Kraft et al. (2019) reaches both the highest R^2 of 0.904 and the lowest RMSE of 0.017 but was trained at both lower temporal and spatial resolution than our approach. Also, as described in Section 2, the time-series approach uses past meteorological data which allows it to model the so-called memory effect of vegetation. Considering this memory effect seems to be important for modeling the NDVI from weather data and incorporating the respective information into our model might close the observed performance gap compared to the LSTM model.

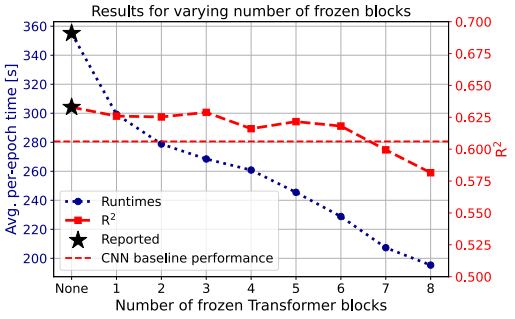

Figure 2: Results for ablation study III: varying number of frozen Transformer blocks during finetuning. Runtimes are averaged over five epochs.

The average performance on the 100 locations selected by Higgins et al. (2023) shows that our finetuned FCN reaches a $28\,\%$ higher R^2 score than the SSMs. A closer analysis of these locations (see Appendix Table 2) shows that a single global model can generally learn biome-specific vegetation patterns, but performance is higher in forested regions than in regions with mainly barren land (*cf.* Appendix Fig. 4). Here, low- to mid-latitude ranges in the Northern Hemisphere are generally well-modelled, while the performance in the Southern Hemisphere and high-latitude regions is worse. This diminished performance may stem from limited data availability towards the poles.

In ablation study I, performance most strongly rises until 80 finetuning epochs, as shown in Fig. 1a. Afterwards, performance stagnates at around an R^2 of 0.62, as more finetuning epochs do not lead to further improvements.

The results for study II in Fig. 1b show that finetuning FCN on more data improves modeling performance. The largest jump occurs when doubling the amount of finetuning data from $10\,\%$ to $20\,\%$. Beyond, further increases still lead to performance improvements albeit at a slower rate. Extrapolating Fig. 1b, this trend suggests that additional data may still enhance performance.

Freezing up to three Transformer blocks in the FCN model results in only minor performance loss, as the results for study III in Fig. 2 show. When more blocks are frozen, the R^2 scores can drop below the CNN baseline. However, more frozen blocks reduce the average per-epoch runtime. It drops from $355\,\mathrm{s}$ for the full model to $195\,\mathrm{s}$ when finetuning only the newly added vegetation-modeling head. These results suggest that selectively freezing a moderate number of Transformer blocks can provide a speedup compared to training the full model, while the performance decreases marginally.

4 Conclusion

In this work, we investigated how a pre-trained weather model can be adapted for globally modeling vegetation activity as measured by the normalized difference vegetation index. We finetuned FourCastNet to model the NDVI from 20 meteorological variables from the ERA5 dataset and reach a globally averaged test set R^2 of 0.6331. This indicates that a weather model finetuned for modeling vegetation activity from high temporal and spatial resolution meteorological data can capture substantial amounts of the NDVI's variability. Our results further show that training from scratch performs worse than finetuning a pre-trained weather model. This suggests that during its pre-training phase, the weather model acquires structural knowledge about the atmosphere, which is beneficial for vegetation modeling and which probably is not attained when training from scratch.

While meteorological data partially reflects the impact of climate variability on vegetation, other factors like atmospheric carbon dioxide, soil-related properties and the so-called memory effect are known to be part of the complex interplay between environmental driving forces and vegetation activity De Keersmaecker et al. (2015); Piao et al. (2020). Hence, incorporating further relevant variables into the model while preserving the information in the pre-trained weather models about

atmospheric dynamics is an area of future work. Lastly, we want to highlight that explainable artificial intelligence techniques allow examining attributions of the model's input to its output, such that Deep Learning models can contribute to enhancing our understanding of how globally changing environmental factors affect local ecosystems.

ACKNOWLEDGEMENT

This research was conducted in the BigData@Geo 2.0 project which is co-financed by the European Regional Development Fund (ERDF). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project ID b214cb. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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

A. L. Burrell, J. P. Evans, and M. G. De Kauwe. Anthropogenic climate change has driven over 5 million km2 of drylands towards desertification. *Nature Communications*, 11(1):3853, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17710-7.

Chen Chen, Bin He, Wenping Yuan, Lanlan Guo, and Yafeng Zhang. Increasing interannual variability of global vegetation greenness. *Environmental Research Letters*, 14(12):124005, November 2019. ISSN 1748-9326. doi: 10.1088/1748-9326/ab4ffc.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

Zhiting Chen, Hongyan Liu, Chongyang Xu, Xiuchen Wu, Boyi Liang, Jing Cao, and Deliang Chen. Modeling vegetation greenness and its climate sensitivity with deep-learning technology. *Ecology and Evolution*, 11(12):7335–7345, 2021. ISSN 2045-7758. doi: 10.1002/ece3.7564.

Wanda De Keersmaecker, Stef Lhermitte, Laurent Tits, Olivier Honnay, Ben Somers, and Pol Coppin. A model quantifying global vegetation resistance and resilience to short-term climate anomalies and their relationship with vegetation cover. *Global Ecology and Biogeography*, 24(5):539–548, 2015. ISSN 1466-8238. doi: 10.1111/geb.12279.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Aya Ferchichi, Ali Ben Abbes, Vincent Barra, and Imed Riadh Farah. Forecasting vegetation indices from spatio-temporal remotely sensed data using deep learning-based approaches: A systematic literature review. *Ecological Informatics*, 68:101552, May 2022. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2022.101552.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803.

Steven I Higgins, Timo Conradi, and Edward Muhoko. Shifts in vegetation activity of terrestrial ecosystems attributable to climate trends. *Nature Geoscience*, 16(2):147–153, 2023.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030, 2022.

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

Basil Kraft, Martin Jung, Marco Körner, Christian Requena Mesa, José Cortés, and Markus Reichstein. Identifying dynamic memory effects on vegetation state using recurrent neural networks. *Frontiers in big Data*, 2:31, 2019.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Christian Lessig, Ilaria Luise, Bing Gong, Michael Langguth, Scarlet Stadler, and Martin Schultz. Atmorep: A stochastic model of atmosphere dynamics using large scale representation learning. *arXiv preprint arXiv:2308.13280*, 2023.

Cuiyan Liu, Jianyu Liu, Qiang Zhang, Hui Ci, Xihui Gu, and Aminjon Gulakhmadov. Attribution of NDVI Dynamics over the Globe from 1982 to 2015. *Remote Sensing*, 14(11):2706, January 2022. ISSN 2072-4292. doi: 10.3390/rs14112706.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Christina Papagiannopoulou, Diego G. Miralles, Stijn Decubber, Matthias Demuzere, Niko E. C. Verhoest, Wouter A. Dorigo, and Willem Waegeman. A non-linear Granger-causality framework to investigate climate–vegetation dynamics. *Geoscientific Model Development*, 10(5):1945–1960, May 2017. ISSN 1991-959X. doi: 10.5194/gmd-10-1945-2017.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. FourCastNet: pretrained weights. `https://github.com/NVlabs/FourCastNet?tab=readme-ov-file#version-notes`, 2023. [Online; last accessed February 2024].

Jon D. Pelletier, A. Brad Murray, Jennifer L. Pierce, Paul R. Bierman, David D. Breshears, Benjamin T. Crosby, Michael Ellis, Efi Foufoula-Georgiou, Arjun M. Heimsath, Chris Houser, Nick Lancaster, Marco Marani, Dorothy J. Merritts, Laura J. Moore, Joel L. Pederson, Michael J. Poulos, Tammy M. Rittenour, Joel C. Rowland, Peter Ruggiero, Dylan J. Ward, Andrew D. Wickert, and Elowyn M. Yager. Forecasting the response of Earth's surface to future climatic and land use changes: A review of methods and research needs. *Earth's Future*, 3(7):220–251, 2015. ISSN 2328-4277. doi: 10.1002/2014EF000290.

Shilong Piao, Xuhui Wang, Taejin Park, Chi Chen, Xu Lian, Yue He, Jarle W. Bjerke, Anping Chen, Philippe Ciais, Hans Tømmervik, Ramakrishna R. Nemani, and Ranga B. Myneni. Characteristics, drivers and feedbacks of global greening. *Nature Reviews Earth & Environment*, 1(1):14–27, January 2020. ISSN 2662-138X. doi: 10.1038/s43017-019-0001-x.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Andrew D Richardson, Trevor F Keenan, Mirco Migliavacca, Youngryel Ryu, Oliver Sonnentag, and Michael Toomey. Climate change, phenology, and phenological control of vegetation feedbacks to the climate system. *Agricultural and Forest Meteorology*, 169:156–173, 2013.

Claire Robin, Christian Requena-Mesa, Vitus Benson, Lazaro Alonso, Jeran Poehls, Nuno Carvalhais, and Markus Reichstein. Learning to forecast vegetation greenness at fine resolution over Africa with ConvLSTMs, November 2022.

Alistair W. R. Seddon, Marc Macias-Fauria, Peter R. Long, David Benz, and Kathy J. Willis. Sensitivity of global terrestrial ecosystems to climate variability. *Nature*, 531(7593):229–232, March 2016. ISSN 1476-4687. doi: 10.1038/nature16986.

Michael J Smith, Luke Fleming, and James E Geach. Earthpt: a foundation model for earth observation. *arXiv preprint arXiv:2309.07207*, 2023.

Compton J Tucker and PJ Sellers. Satellite remote sensing of primary production. *International journal of remote sensing*, 7(11):1395–1416, 1986.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Eric Vermote. Noaa climate data record (cdr) of avhrr normalized difference vegetation index (ndvi), version 5, 2019. Downloaded from `https://www.ncei.noaa.gov/data/land-normalized-difference-vegetation-index/access/`, last accessed February 2024.

Zaichun Zhu, Shilong Piao, Ranga B. Myneni, Mengtian Huang, Zhenzhong Zeng, Josep G. Canadell, Philippe Ciais, Stephen Sitch, Pierre Friedlingstein, Almut Arneth, Chunxiang Cao, Lei Cheng, Etsushi Kato, Charles Koven, Yue Li, Xu Lian, Yongwen Liu, Ronggao Liu, Jiafu Mao, Yaozhong Pan, Shushi Peng, Josep Peñuelas, Benjamin Poulter, Thomas A. M. Pugh, Benjamin D. Stocker, Nicolas Viovy, Xuhui Wang, Yingping Wang, Zhiqiang Xiao, Hui Yang, Sönke Zaehle, and Ning Zeng. Greening of the Earth and its drivers. *Nature Climate Change*, 6(8): 791–795, August 2016. ISSN 1758-6798. doi: 10.1038/nclimate3004.

A EXTENDED RESULTS AND SUPPLEMENTARY FIGURES

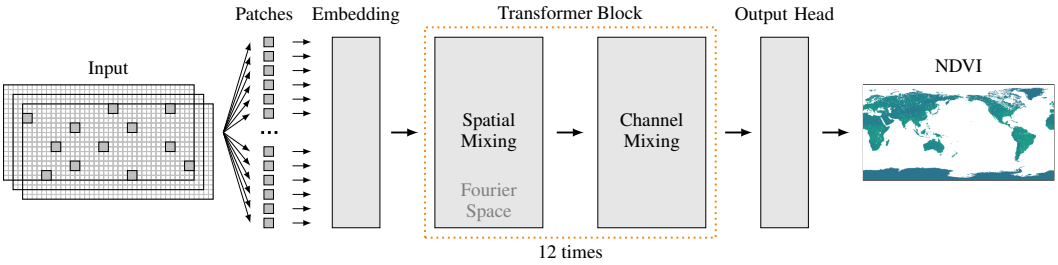

Figure 3: Overview of the used architecture based on the FourCastNet (FCN) model Pathak et al. (2022). Following the common VisionTransformer (ViT) Dosovitskiy et al. (2020) approach, FCN first extracts quadratic, 8 by 8 patches from the 720 x 1440 weather data input. Each patch is embedded as a 768-dimensional embedding vector. Subsequently, 12 Transformer-like encoder blocks process the patched data. Each block contains the Adaptive Fourier Neural Operator (AFNO) Guibas et al. (2021) as an attention module responsible for spatial mixing, followed by channel (*i.e.*, variable) mixing. Finally, an output head consisting of a linear layer with *tanh* activation function performs patch recovery to the input 720 x 1440 resolution to globally model the NDVI. For our experiments, we either initialize the model from pre-trained weights Pathak et al. (2023) or train a freshly-initialized architecture from scratch.

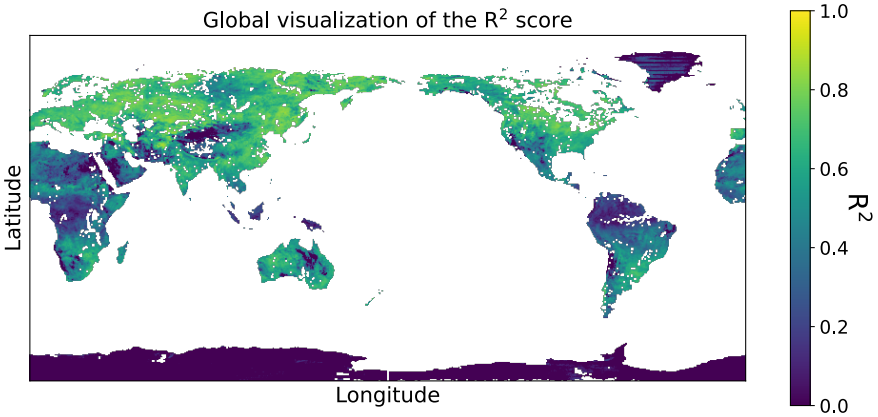

Figure 4: Global visualization of the R^2 score on the entire test set. R^2 scores below 0 are clipped to 0 for ease of visualization. Performance is the strongest for continental Europe and North America and decreases towards higher latitude regions. This might caused by the unevenly distributed data, as the majority of training samples are concentrated in the northern hemisphere, while the southern hemisphere contains large ocean masses. In preliminary experiments, we studied the ability of FCN to generalize by excluding entire continents during training. We found that for the excluded area, performance was much worse than for the unmasked areas. This observation indicates that further investigation is necessary to improve the generalization abilities of our global ecosystem model, possibly by including variables describing the topology.

Table 2: Extended results for local evaluation. Unweighted average of RMSE and R^2 score per biome of the 100 location from Higgins et al. (2023). The finetuned model improves the averaged R^2 score across all biomes compared to the SSMs except for boreal forests and tundra. Our global modeling approach thus captures biome-specific NDVI dynamics.

	RMSE		R2		Samples
Biome	FCN	SSM	FCN	SSM	
Boreal forest	0.0656	0.0834	0.7244	0.8321	16
Grassland	0.0416	0.0449	0.4716	0.4484	14
Mediterranean-type	0.0364	0.0393	0.1921	−0.6008	5
Tropical forest	0.0789	0.0508	0.1405	−0.0515	16
Savanna	0.0512	0.0516	0.7151	0.6628	18
Shrubland	0.0478	0.0445	0.3021	0.1966	16
Temperate forest	0.0598	0.0693	0.7726	0.5247	12
Tundra	0.0423	0.0600	0.8312	0.9149	9

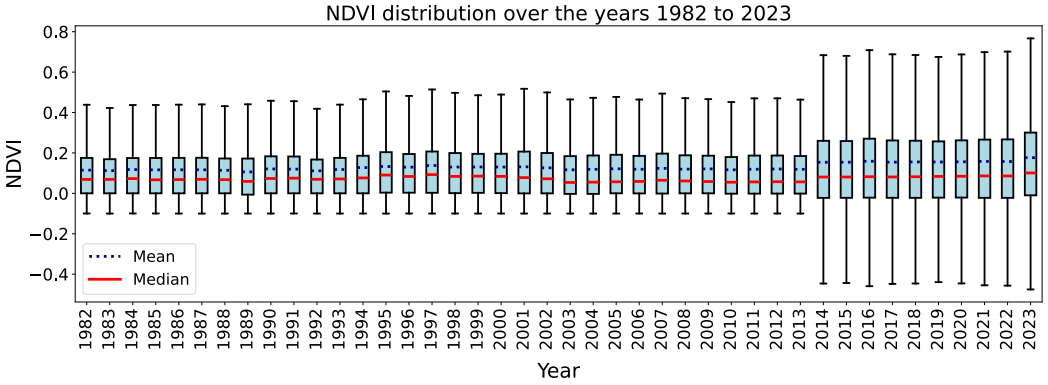

Figure 5: Distribution of the normalized difference vegetation index (NDVI) data from 1982 to 2023. NDVI data was only used until 2013 due to the noticeable data shift afterwards.

B FourCastNet training details

We train/finetune the FCN with the Adam optimizer (Kingma & Ba, 2014) and a learning rate of 0.0001 with cosine annealing (Loshchilov & Hutter, 2016) for 80 epochs (except when varying the number of training epochs in ablation study I, see Section 2) using an l_2 loss. Using a binary mask, the loss is only computed for locations with valid NDVI observations. The weights of the epoch with the lowest validation loss are kept. We trained our models using a single node equipped with 8 NVIDIA A100 40GB GPUs.

C Baseline and comparison models details

Table 3: Hyperparameter search ranges for the baseline CNN network.

Hyperparameter	Search space	Step size
n_layers	3 to 8	1
learning rate	1e−5 to 1e−5	log uniform
epochs	50 to 100	10
n_filters	16 to 512	16
kernel_size	3 to 7	2

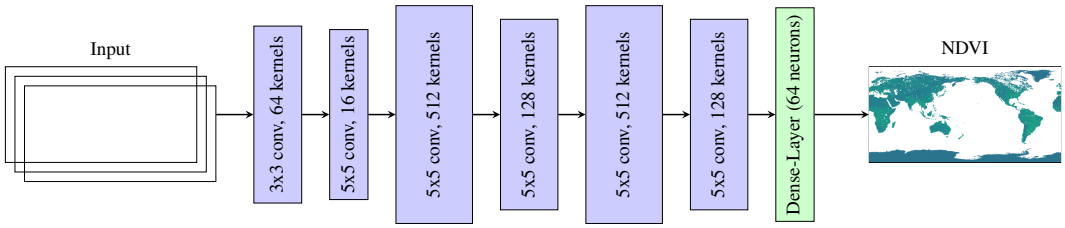

Figure 6: Overview of the hyperparameter-optimized architecture of our baseline CNN model. Its hyperparameters were optimized on the validation set over 60 trials. All convolution layers use a stride of 1 with the kernel sizes and number of kernels as indicated. Similar to FCN, a linear layer is used to reshape the data into a 720 x 1440 output.

We use a convolutional neural network (CNN) as a baseline model. The CNN's hyperparameters were optimized over 60 trials on the validation set with the search ranges given in Table 3. The hyperparameter-optimized CNN consists of six convolution layers with 64, 16, 512, 128, 512, 128 kernels, respectively. The used kernel sizes are 3, 5, 5, 5, 3, 5, each with a stride of 1. The CNN's last layer is a fully connected dense layer with 64 neurons, whose output is reshaped to the target resolution of 720 x 1440. See Fig. 6 for a visualization. The CNN was trained with a learning rate of 0.00006 for 80 epochs.

D EVALUATION SETTING

Global evaluation For global evaluation in comparison to LSTM models by Kraft et al. (2019), we compute 15-days averages of our model output and target values to match their temporal resolution. To provide a fair comparison to the reported results of Kraft et al. (2019), we replicated their evaluation setting to the best of our knowledge, since their code is not publicly available. To remove noisy pixels as defined by Kraft et al. (2019), we remove pixels with $50\,\%$ missing data in the time dimension, pixels with less than $20\,\%$ land mass, and barren-land pixels, which together removes coastal, high-latitude and desert regions. Further, to account for the varying size of pixels across different latitudes as Kraft et al. (2019), we use latitude-weighted RMSE and R^2 scores, with the latitude weighting factors $w_1, ..., w_I$ for I evaluated pixels given by

$$w_i = \frac{\cos(\text{lat}(i))}{\frac{1}{N_{\text{lat}}} \sum_{j=1}^{N_{\text{lat}}} \cos(\text{lat}(j))} \quad \forall i \in \{1, ..., I\} \quad .$$

We assume that our latitude-weighting is identical to their employed area weighting scheme, *i.e.* given the latitude weights $w_1, ..., w_I$, the corresponding pixel areas $\mathcal{A}_1, ..., \mathcal{A}_I$ and their total area A

$$\frac{\mathcal{A}_i}{A} = \frac{w_i}{\sum_{j=1}^{I} w_j} \quad \forall i \in \{1, ..., I\} \quad .$$

With this assumption, we can show that the reported biome-weighted $\text{RMSE}_{\text{global}}$ for the biomes $A^1, ..., A^B$ in Kraft et al. (2019) is identical to our latitude-weighted RMSE:

$$\begin{aligned}
\text{RMSE} &= \frac{1}{\sum_{j=1}^{I} w_j} \sum_{i=1}^{I} w_i \cdot \text{RMSE}_i \\
&= \frac{1}{A} \sum_{i=1}^{I} \mathcal{A}_i \cdot \text{RMSE}_i \\
&= \frac{1}{A} \sum_{b=1}^{B} \sum_{k \in I^b} \mathcal{A}_k^b \cdot \text{RMSE}_k^b \\
&= \frac{1}{A} \sum_{b=1}^{B} A^b \cdot (\sum_{k \in I^b} \frac{\mathcal{A}_k^b}{A^b} \cdot \text{RMSE}_k^b) \\
&= \frac{1}{A} \sum_{b=1}^{B} A^b \cdot \text{RMSE}^b \\
&= \text{RMSE}_{\text{global}}
\end{aligned}$$

The above equations also hold for the biome-weighted R^2_{global} in Kraft et al. (2019) and our latitude-weighted R^2 score. However, we do not perform the same 10-fold spatio-temporal cross-validation due to the long training time of our model. Additionally, note that the LSTM model is evaluated at a coarser resolution of $0.5°$than our FCN model ($0.25°$) and was trained on a different set of variables. Those are six dynamic meteorological variables and 21 static variables including water capacity, water table depth and land cover fractions.

Local evaluation For local evaluation in comparison to the state space models (SSM) results provided by Higgins et al. (2023), we average our model output and prediction at the same 100 locations to weekly resolution. We then compute unweighted average RMSE and R^2 scores across these locations. Note that an individual SSM is trained per location, using location-specific air temperature $2\,\text{m}$ above the surface, soil temperature, soil moisture, surface solar radiation, and atmospheric carbon dioxide at $0.083°$ spatial resolution as climate-forcing data. The target data is the NDVI at a weekly temporal resolution.

E Data scaling in ecosystem modeling

The results for ablation study II, visualized in Fig. 1b, show that over the evaluated range of number of finetuning data, performance increases as more data is used. A similar behaviour was observed for earth observation data by Smith et al. (2023), who trained a pixel-wise autoregressive Transformer model (Radford et al., 2019) on satellite-derived earth observation data. They noted that scaling their training data also scales model performance. In lieu of a scaling "law" that applies to earth observation data, they assume their model performance to follow the scaling law observed for large language models:

$$N \sim 20\ D,$$

where N is the number of model parameters and D is the number of training tokens (Hoffmann et al., 2022). In the following, we assume that this tendency also applies to our vegetation-modeling approach.

Our training dataset contains 29 training years (1982 to 2010) with 365 days per years, leading to $29 * 365 = 10585$ training samples. One sample has dimensionality 720 x 1440, and the patch size used to tokenize this image-like input is 8. For one training sample, this leads to $720/8 = 90$ tokens covering the latitudinal direction, and $1440/8 = 180$ tokens for the longitudinal direction, or $90 * 180 = 16200$ total tokens for one sample. Over the entire training data, we thus have $16200 * 10585 = 171477000$ tokens in total.

The FourCastNet we use has 73 million parameters. Assuming that the mentioned scaling law applies, we would thus need to train FCN on $20 * 73\,\text{million} = 1.46\,\text{billion}$ tokens for optimal per-

formance[1] However, this calculation assumes the validity of the large language model scaling law. For autoregressive, image-like data-generating models, different data-scaling might be better suited, such as the ones proposed by Henighan et al. (2020). Applying these laws to ecosystem models is thus an open research direction.

[1]Or, alternatively, we should scale down the model to 171477000 total tokens $/\ 20 = 8.573.850$ parameters, which would roughly be a two-Transformer block model. The observed results in Fig. 2 currently indicate that this configuration does not show the best performance, at least for a setting where these blocks are the only ones being trained, and further blocks – and thus parameters – exist but are frozen.