

Exploring Methods of Explainable AI - Data-driven Attribution of Climate Events

M.Sc. Marco Landt-Hayen
aus Kiel

Dissertation
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
(Dr.-Ing.)
der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel
eingereicht im Jahr 2023

Kiel Computer Science Series (KCSS) 2023/5 dated 2023-10-04

ISSN 2193-6781 (print version)

ISSN 2194-6639 (electronic version)

Electronic version, updates, errata available via <https://www.informatik.uni-kiel.de/kcss>

The author can be contacted via malaha@gmx.net

Published by the Department of Computer Science, Kiel University

Information Systems and Data Mining

Please cite as:

- ▷ Marco Landt-Hayen. *Exploring Methods of Explainable AI - Data-driven Attribution of Climate Events* Number 2023/5 in Kiel Computer Science Series. Department of Computer Science, 2023. Dissertation, Faculty of Engineering, Kiel University.

```
@book{LandtHayen2023,  
  author    = {Marco Landt-Hayen},  
  title     = {Exploring Methods of Explainable AI -  
              Data-driven Attribution of Climate Events},  
  publisher = {Department of Computer Science, Kiel University},  
  year      = {2023},  
  number    = {2023/5},  
  doi       = {10.21941/kcss/2023/5},  
  series    = {Kiel Computer Science Series},  
  note      = {Dissertation, Faculty of Engineering,  
              Kiel University.}  
}
```

© 2023 by Marco Landt-Hayen

The author has obtained the right to include the published and accepted articles in the thesis, with a condition that they are cited and/or copyright/credits are placed prominently in the references.

About this Series

The Kiel Computer Science Series (KCSS) covers dissertations, habilitation theses, lecture notes, textbooks, surveys, collections, handbooks, etc. written at the Department of Computer Science at Kiel University. It was initiated in 2011 to support authors in the dissemination of their work in electronic and printed form, without restricting their rights to their work. The series provides a unified appearance and aims at high-quality typography. The KCSS is an open access series; all series titles are electronically available free of charge at the department's website. In addition, authors are encouraged to make printed copies available at a reasonable price, typically with a print-on-demand service.

Please visit <http://www.informatik.uni-kiel.de/kcss> for more information, for instructions how to publish in the KCSS, and for access to all existing publications.

1. Gutachter: Prof. Dr. Peer Kröger
Christian-Albrechts-Universität
Kiel
2. Gutachter: Prof. Dr. Martin Claus
Christian-Albrechts-Universität
Kiel

Datum der mündlichen Prüfung: 29.06.2023

Zusammenfassung

Die vorliegende Arbeit widmet sich zum einen der Erkundung von Methoden, um den Lernprozess bei künstlichen neuronalen Netzen (KNNen) besser zu verstehen. Und zum anderen geht es darum, die gewonnenen Erkenntnisse anzuwenden, um neue Zusammenhänge im Klimasystem aufzuspüren.

Wenn man mit geophysikalischen Beobachtungsdaten arbeiten möchte, stehen konsistente Messdaten oft nur aus der jüngeren Vergangenheit zur Verfügung. Die meisten Methoden maschinellen Lernens sind jedoch datenhungrig. Es gibt einige Ausnahmen. Dazu zählen Echo State Netzwerke (ESNe). Wir versuchen zunächst, ein von anderen Netztypen bereits bekanntes Verfahren namens „layer-wise relevance propagation“ auf ESNe zu übertragen. Allerdings stoßen wir dabei auf Artefakte in den Darstellungen der Ergebnisse, die wir genauer untersuchen.

Anschließend versuchen wir, das Problem unzureichender Trainingsdaten abzumildern. Dafür rufen wir einen neuen Standard-Datensatz ins Leben. Es handelt sich dabei um eine Sammlung von Klimaindizes, die die wesentlichen Zusammenhänge im Klimasystem beschreiben. Damit wir konsistente Daten über einen möglichst langen Zeitraum bekommen, greifen wir auf Daten aus modernen Erdsystem-Modellen zurück.

Der neue Datensatz ermöglicht uns, ein weiteres Kernproblem im Zusammenhang mit geophysikalischen Daten anzugehen: Die Messdaten sind nämlich meistens lückenhaft. Um die Lücken zu füllen, benutzen wir sogenannte „U-Net“ Modelle aus dem Bereich der maschinellen Bildverarbeitung. Wir zeigen welche Eingabewerte besonders wichtig sind, um die fehlenden Informationen wiederherzustellen. Das hilft die Frage zu beantworten, an welchen Orten man eine brenzte Anzahl von Messstationen platzieren sollte mit größtmöglichem Nutzen. Und zuletzt prognostizieren wir zukünftige geophysikalische Felder und versuchen, Klimaereignisse, wie zum Beispiel El Niño oder Regenfälle in der Sahelzone Afrikas, in naher Zukunft vorherzusagen.

Abstract

In this work, we explore methods of explainable artificial intelligence (xAI) to better understand how artificial neural networks (ANNs) come to their conclusion and to visualize the relationship between input and output. Moreover, we aim to use our insights to improve our understanding of the climate system.

Working with observational data in the context of geophysics can be challenging since we have consistent data only from the recent past. But most machine learning (ML) methods are data hungry and need sufficient training data. In practice, there exist some exceptions. Among these, we find Echo State Networks (ESNs) as a certain type of recurrent neural networks. We show that layer-wise relevance propagation (LRP) can be used for ESNs. However, there are pitfalls in terms of model-inherent artifacts included in the obtained relevance maps. We revise LRP on various ANN architectures and give guidance on how to control model focus.

Furthermore, we address the problem of insufficient training data and introduce a new benchmark data set. In particular, we create a collection of climate indices used to describe the dynamics of the Earth system. In order to have consistent data over a sufficiently long time span, we favor to work with data from modern Earth System Models.

The new data set opens the door to tackle another problem regarding geospatial data. In the context of geophysics, we often have missing values. We use U-Net models from the domain of computer vision to reconstruct missing data in two-dimensional geospatial fields. Moreover, we introduce a technique to identify grid points that are most relevant for successful reconstruction. This helps to answer the question where to place a limited number of survey stations to get most information out of it. Eventually, we extend our approach to predict future geospatial fields from sparse inputs. We try to infer climate events like e.g., El Niño or rainfall in the African Sahel region from reconstructed data.

Acknowledgements

The story began back in 2001, when I started to study computer science at Kiel university, because I was interested in artificial intelligence (AI). It turned out that I was quite ahead of time. Back then, there were no lectures on AI, yet. So I decided to switch to physics and forgot about my initial plan for a couple of years. Until I met John Beggs and attended his lecture on biological and artificial neural networks in 2005. That was even before the deep learning hype started, but he immediately caught my interest. After finishing my studies, I began to work and again, the AI volcano inside of me fell asleep. The next major eruption was in 2019, when I remembered my passion for AI. As I looked for lectures on machine learning and deep learning, I found a new Master's program at the University of Applied Sciences in Kiel and decided to go for it. That turned out to be an excellent decision and once I finished the Master's program, I was looking for a way to make a living out of it. As the universe seems to be on my side, I accidentally met Christopher Somes on a playground some Sunday afternoon and started talking to him, while our kids cruised around. He works as a Postdoc at GEOMAR and told me about the Helmholtz School for Marine Data Science (MarDATA) and that they were about to hire a next round of doctoral researchers. And now I am here, almost done with my doctorate and quite happy that I am finally able to do what I always wanted to do. I want to thank my PIs, Peer Kröger and Martin Claus, for supporting me whenever I needed assistance and mostly for letting me find my own way through this exciting journey. And most of all, I want to thank Willi Rath and Yannick Wölker for many fruitful discussions, inspiration and for showing me the secrets of working with Git and high performance computing.

Contents

I Thesis Summary	3
1 Introduction	5
1.1 Thesis Structure	6
1.2 ENSO and Sahel Rainfall	7
1.3 List of Publications	8
2 Research Contributions	11
2.1 Layer-wise Relevance Propagation and Echo State Networks	11
2.2 A Climate Index Collection based on Model data (CICMoD)	12
2.3 Reconstruct Missing Data with U-Nets	13
3 Conclusion and Outlook	15
Bibliography	19
II Papers	23
A Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability	25
B Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures	41
C A Climate Index Collection based on Model Data	57
D CICMoD - A Climate Index Collection Benchmark (Data and Resources Paper)	71
E A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs	77

Contents

F	Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events	93
----------	---	-----------

Part I

Thesis Summary

Introduction

This work is based on interdisciplinary research at the interface of climate science and computer science. In computer science, we find machine learning (ML) and deep learning (DL) methods to be state-of-the-art for many applications, e.g., time series forecasting or computer vision. Despite their success, these methods and models are seen as black boxes. Understanding the reasoning behind such models is therefore essential to gain trust in the results. The subfield of explainable artificial intelligence (xAI) aims to open the black box in order to reveal, what ML and DL models have learned and how they come to their conclusion [RSG16; GBY+18].

In climate science, we have modes of climate variability that are connected to each other over large distances, also referred to as teleconnections. These relationships are often found to be non-stationary [PPV+14; PKP+18; ZMG+19]. We need to identify such teleconnections to predict certain climate events more precisely and with longer lead times.

With this work, we try to build a bridge between both domains. Using xAI methods, we aim to create interpretable models to enhance our understanding of the climate system in general in a data-driven way. However, we face two problems when applying ML and DL methods to geospatial data.

- ▷ **P1:** We only have consistent observational data from the recent past.
- ▷ **P2:** In the context of geophysics, we have to deal with missing values in observational data.

1. Introduction

With these constraints in mind, we address three research questions in this work.

- ▷ **Q1:** How to design artificial neural networks (ANNs) that can handle these constraints and still allow for interpretability of the results in the context of climate data?
- ▷ **Q2:** How to open the black box and visualize the relationship between model input and output to understand the reasoning behind such ANNs?
- ▷ **Q3:** How to use the insights from Q1 and Q2 to enhance the understanding of the climate system in general?

To approach Q1, we focus on specific network architectures. In particular, we investigate Echo State Networks (ESNs) [SSH+20] and encoder-decoder-style convolutional neural networks (CNNs) also referred to as U-Nets [RFB15]. For Q2, we adopt layer-wise relevance propagation (LRP) [BBM+15] to be used on ESNs trained on geospatial data. Furthermore, we introduce a new method of visualizing the relationship between input and output for U-Nets. For the climate science part of this thesis, we intend to apply the methods developed for Q1 and Q2 to certain climate events for Q3.

1.1 Thesis Structure

This thesis is divided into two parts. Part I gives an overview of the research that has been done during this thesis. All papers and manuscripts written during this work are included in Part II.

In the remainder of this chapter, we briefly introduce the El Niño Southern Oscillation (ENSO) and rainfall in the African Sahel region (Sahel rainfall) in Section 1.2 for readers stemming from computer science with low background knowledge on climate science and physical oceanography. These phenomena are used to test and showcase our methods in later sections. In Section 1.3, we list all papers, manuscripts and presentations that have been assembled during this thesis.

Chapter 2 summarizes the contributions and further implications of the research in this work. Chapter 3 concludes our findings and gives an outlook for further work.

1.2 ENSO and Sahel Rainfall

Throughout this work, we will use ENSO and Sahel rainfall as toy problems to test and demonstrate new methodologies. Therefore, we briefly introduce these phenomena with their characteristics and implications in this section.

ENSO is the predominant variation of winds and sea surface temperature (SST) in the Tropical Pacific. The positive phase (El Niño) is characterized by unusual warm SST and high sea level pressure (SLP) in the Eastern Tropical Pacific, whereas the negative phase (La Niña) relates to unusual cold SST and low SLP in the same region and above-average SST in the Western Tropical Pacific. Both events last several months and occur with a period of 2-7 years with varying intensity per period. ENSO tremendously affects those countries bordering the Pacific Ocean. Strong El Niños e.g., correspond to warm weather conditions with heavy rainfalls from April through October causing major flooding along the West coast of South America near Ecuador and the Northern part of Peru [CMG+20]. Consequences of La Niña are e.g., heavy rainfalls over Malaysia, the Philippines and Indonesia. Therefore, knowing the ENSO phase several months in advance is of high interest for society since it allows to take measures to avoid damage and to protect people.

Summer precipitation in the African Sahel region has been observed to be highly variable with floods and droughts occurring on a regular basis and has a high impact on living conditions in the region [BZG14]. Predicting Sahel rainfall and understanding the underlying processes is essential, since it allows taking measures in advance to avoid damage and prevent hunger crises.

1. Introduction

1.3 List of Publications

The following papers and presentations were assembled during this thesis.

Papers and Manuscripts

Marco Landt-Hayen, Peer Kröger, Martin Claus, and Willi Rath: “Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability“, In: *Proceedings of the 3rd International Conference on Machine Learning Techniques (MLTEC 2022), Computer Science & Information Technology*, vol. 12(20), 2022, pp. 115–130, DOI: [10.5121/csit.2022.122008](https://doi.org/10.5121/csit.2022.122008).

Marco Landt-Hayen, Willi Rath, Martin Claus, and Peer Kröger: “Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures“, Accepted for Publication at: *The 4th International Conference on Machine Learning Techniques (MLTEC 2023)*.

Marco Landt-Hayen, Willi Rath, Sebastian Wahl, Nils Niebaum, Martin Claus, and Peer Kröger: “A Climate Index Collection based on Model Data“, In: *Environmental Data Science*, 2, E9, 2023, DOI: [10.1017/eds.2023.5](https://doi.org/10.1017/eds.2023.5).

Marco Landt-Hayen, Willi Rath, Sebastian Wahl, and Martin Claus: “CIC-MoD - A Climate Index Collection Benchmark (Data and Resources Paper)“, In: *The 31st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23)*, DOI: [10.1145/3589132.3625627](https://doi.org/10.1145/3589132.3625627).

Marco Landt-Hayen, Yannick Wölker, Willi Rath, and Martin Claus: “A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs“, Accepted for Publication at: *The 19th International Conference on Advanced Data Mining and Applications (ADMA 2023)*.

Marco Landt-Hayen, Peer Kröger, Willi Rath, and Martin Claus: “Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events“, Accepted for Publication at: *The 19th IEEE International Conference on eScience (IEEE eScience 2023)*.

1.3. List of Publications

Presentations

Marco Landt-Hayen, Peer Kröger, Martin Claus, and Willi Rath: “Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability”, Poster Presentation: *7th Data Science Symposium, Hereon*, 2022.

Marco Landt-Hayen, Peer Kröger, Martin Claus, and Willi Rath: “Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability”, Oral Presentation: *AGU Fall Meeting*, 2022.

Marco Landt-Hayen, Willi Rath, Sebastian Wahl, Nils Niebaum, Martin Claus, and Peer Kröger: “Data-driven Attributing of Climate Events with Climate Index Collection based on Model Data (CICMoD)”, Oral Presentation: *EGU General Assembly*, 2023, DOI: 10.5194/egusphere-egu23-984.

Marco Landt-Hayen, Willi Rath, Martin Claus, and Peer Kröger: “Reconstruct missing data from sparse inputs with CNNs to find an optimized sampling strategy”, Oral Presentation: *54th International Liège Colloquium on Ocean Dynamics*, 2023.

Research Contributions

In this chapter, we briefly summarize the insights obtained from the interdisciplinary research at the interface of computer science and climate science.

2.1 Layer-wise Relevance Propagation and Echo State Networks

A problem with using ANNs on data of the Earth system is that we often only have short time series to predict on or a small number of events to learn from (P1). With this constraint, ESNs as a certain type of recurrent neural networks appear to be promising, since they need less training data compared to other types of ANNs like e.g., multilayer perceptrons (MLPs), CNNs or long short term memory (LSTM) networks. In its basic form, an ESN consists of an input and an output layer. In between, we find a reservoir of sparsely connected units. Weights and biases connecting inputs to reservoir units and internal reservoir weights and biases are randomly initialized. The input length determines the number of recurrent time steps inside the reservoir. We record the final reservoir states and only the output weights and bias are trained. But opposed to other types of neural networks, this does not encounter gradient descent methods but is rather done in a closed-form manner by applying linear regression of final reservoir states onto desired target values to get the output weights and bias.

We used ENSO as a toy problem and showed that ESNs can be used for image classification on SST anomaly fields. To open the black box, we adopted layer-wise relevance propagation from computer vision to ESNs.

2. Research Contributions

The aim is to understand, which parts of the input sample have highest relevance and hence most influence on the model prediction. Relevance can be traced back through the network to attribute a certain score to each input pixel. Relevance scores are then combined and displayed as heat maps and give humans an intuitive visual understanding of classification models.

However, we found pitfalls in terms of model-inherent artifacts included in the obtained relevance maps, that can easily be missed. But for a valid interpretation, these artifacts must not be ignored. Therefore, we revised LRP on various ANN architectures trained as classifiers on geospatial and synthetic data. Depending on the network architecture, we showed techniques to control model focus and gave guidance to improve the quality of obtained relevance maps to separate facts from artifacts.

For more details, see Papers A and B.

2.2 A Climate Index Collection based on Model data (CICMoD)

Instead of overcoming the constraint of having limited observational data (P1) by choosing an appropriate network design, we switched to model data obtained from Earth System Models (ESMs) to have consistent data over a sufficiently long time span.

The data science community provides standard data sets for many applications. However, benchmark data sets in the field of climate science are rare. As a new data set, we introduced a climate index collection based on model data (CICMoD). Therefore, we used control simulations from ESMs, namely the Flexible Ocean and Climate Infrastructure (FOCI) [MBW+20] and the Whole Atmosphere Community Climate Model (WACCM) as extension of the Community Earth System Model (CESM) [HHG+13; MMK+13]. In particular, we worked with SST, SLP, geopotential height at pressure level 500 millibar (Z500), surface air temperature (SAT), sea surface salinity (SSS) and total precipitation (PREC) as two-dimensional fields. From these variables, we derived a set of 29 climate indices over 1,000 and 999 years for FOCI and CESM, respectively. The obtained collection of climate indices serves as a reduced description of the Earth system

2.3. Reconstruct Missing Data with U-Nets

in a consistent and comprehensive way.

The data set is provided as an open-source framework that can be extended and customized to individual needs. It allows users to develop new ML methodologies and to compare results to existing methods and models as benchmark. For example, we used the data set to predict ENSO and Sahel rainfall with various ML models.

For more details, see Papers C and D.

2.3 Reconstruct Missing Data with U-Nets

When we work with observational data in the context of geophysics, we often have to deal with missing values (P2). In reanalysis data, these gaps are usually filled using statistical methods. Thus, we don't even know the ground truth. To address this problem, we proposed U-Nets to reconstruct complete data from sparse inputs. This approach is suitable for two-dimensional geospatial data. To have consistent data over a sufficiently long time span and to know the ground truth, we used our CICMoD data set including the underlying raw data. Missing values have been simulated with different types of masks. Additionally, we have permanently missing data for SST and SSS in terms of land masses. All missing values have been set to zero (zero-inflated).

We showed that our networks can restore complete information from incomplete input samples with varying rates of missing data. Moreover, we presented a bottom-up sampling strategy as a technique to identify the most relevant grid points of our input samples. In particular, we added and fixed grid points that lead to largest drop in reconstruction loss and assigned the relative loss reduction to the corresponding grid points. Results vary for different samples. Therefore, we averaged over multiple training samples to obtain representative mean relative loss reduction maps (MRLRMs), individually for all features from both ESMs. MRLRMs have been visualized as heat maps and give an intuitive understanding of grid points' relevance for successful reconstruction.

Choosing the optimal subset of grid points allowed us to successfully reconstruct SLP and SST anomaly fields, respectively, from ultra sparse univariate inputs. The insights obtained from ESM data have been transferred

2. Research Contributions

to real world observations to improve reconstruction quality.

We then extended our approach of reconstructing *current* fields from univariate inputs and tried to predict *future* fields of SST and PREC anomalies from ultra sparse multivariate inputs. As a proof of concept, we tried to predict ENSO and Sahel rainfall from restored SST and PREC data, respectively. To quantify uncertainty, we compared corresponding climate indices derived from reconstructed versus complete fields.

For more details, see Papers E and F.

Conclusion and Outlook

In this thesis, we explored methods for enhancing the interpretability of ANNs. As we showed in this work, ESNs can be used for image classification. The fact, that ESNs need less training data compared to other network topologies makes them suitable for the use on geospatial data (Q1) where we often have the constraint of limited training data (P1). We successfully adopted LRP to ESNs to gain insights into the classification engine of ESNs by visualizing the obtained mean relevance maps as heat maps (Q2). These maps are supposed to give humans an intuitive understanding of which parts of the input samples are most relevant for a certain model decision, in our case the discrimination of El Niño and La Niña samples (Q3). However, the simplicity of the training process for ESNs has the disadvantage that ESNs highlight spots in mean relevance maps that are irrelevant for classification. We were unable to reach a focus on only the relevant parts of the input samples. Opposed to ESNs, the focus of MLP and CNN models can be controlled by weight regularization (Q1). While highlighting only the most relevant parts appears to be efficient for successfully discriminating input samples, we lose information on features that are of minor relevance but potentially reveal insights of existing teleconnections. Overall, neither model is found to be superior for image classification with LRP.

Beyond the application to geospatial data, ESN models could be used for time series forecasting. Instead of feeding two-dimensional fields into the model, we may pass a certain number of climate indices with specific input length to an ESN model and LRP could serve as an alternative for the temporal attention mechanism often used in the context of LSTM sequence-to-sequence models.

As another approach of addressing the problem of insufficient training

3. Conclusion and Outlook

data, we introduced CICMoD as a new benchmark data set. This consistent and comprehensive collection of climate indices in combination with the underlying geospatial data allows the user to develop new ML methods and to compare results to existing methods and models in an objective way. The index collection is not complete since we focus on processes within the atmosphere, in the upper ocean and at the interface of ocean and atmosphere. However, our CICMoD data set serves as basis. Since it is provided as an open-source framework, it can be extended and customized to individual needs including the application to further ESMs. This opens the door for collaboration in many ways. Our new data set allows researchers from the data science community to tackle problems from the domain of climate science and get a deeper understanding of the Earth system. This requires involving scientists and practitioners from the domain of climate science.

As future work, we plan to combine CICMoD with an extensive toolbox of xAI methods. Our new data set in combination with this xAI toolbox can then be used e.g., for data science competitions or to educate students of both domains, computer science and climate science.

Additionally, our new data set opened the door to tackle the problem of having missing values in observational data (P2). We used U-Net models from the domain of computer vision to reconstruct missing data in two-dimensional geospatial fields (Q1). Moreover, we introduced a bottom-up sampling strategy to identify grid points that are most relevant for successful reconstruction. The resulting MRLRMs can be visualized as heat maps as a new xAI method to present the relationship between model input and output (Q2). Our insights obtained from ESM data can be transferred to real world data. This helps to answer the question where to place a limited number of survey stations to get most information out of it (Q3).

In a first step, we used univariate input data to reconstruct missing values in *current* fields. As an extension, we tried to look ahead and predict *future* SST and PREC fields with lead times of one and three months, respectively, from ultra sparse inputs. This required additional information in form of multivariate or time-lagged inputs to infer ENSO and Sahel rainfall from reconstructed data (Q3). While we succeeded to predict SST fields, the reconstruction of missing PREC data failed, at least

in the ultra sparse regime.

As future work, we plan to revise the way we prepare the input data. So far, we worked with two-dimensional geospatial data on a global scale and used a rectangular projection as prerequisite for our U-Net models with two-dimensional convolutions. Like this, we pretend to have equal distance of grid points in latitude and longitude directions. However, distances of neighboring grid points in longitude direction depend on latitude. One idea is to use a latitude-weighted loss function for training our U-Net models. Another idea is to try alternative projections for our geospatial input data in combination with spherical convolution or graph neural networks, as alternative network topologies.

Bibliography

- [BBM+15] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. “On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation”. In: *PLoS ONE*. Vol. 10. 2015.
- [BZG14] Hamada S. Badr, Benjamin F. Zaitchik, and Seth D. Guikema. “Application of Statistical Models to the Prediction of Seasonal Rainfall Anomalies over the Sahel”. In: *Journal of Applied Meteorology and Climatology*. Vol. 53 (3). 2014, pp. 614–636. DOI: 10.1175/JAMC-D-13-0181.1.
- [CMG+20] Wenju Cai et al. “Climate Impacts of the El Niño-Southern Oscillation on South America”. In: *Nature Reviews Earth & Environment*. Vol. 1. 2020, pp. 215–231.
- [GBY+18] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. “Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning”. In: *Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018)*. 2018, pp. 80–89. DOI: 10.1109/DSAA.2018.00018.
- [HHG+13] James W. Hurrell et al. “The Community Earth System Model: A Framework for Collaborative Research”. In: *Bulletin of the American Meteorological Society*. Vol. 94. 2013, pp. 1339–1360. DOI: 10.1175/BAMS-D-12-00121.1.
- [MBW+20] Katja Matthes et al. “The Flexible Ocean and Climate Infrastructure Version 1 (FOCI1): Mean State and Variability”. In: *Geoscientific Model Development*. Vol. 13 (6). 2020, pp. 2533–2568. DOI: 10.5194/gmd-13-2533-2020.

Bibliography

- [MMK+13] Daniel R. Marsh, Michael J. Mills, Douglas E. Kinnison, Jean-Francois Lamarque, Natalia Calvo, and Lorenzo M. Polvani. "Climate Change from 1850 to 2005 Simulated in CESM1(WACCM)". In: *Journal of Climate*. Vol. 26 (19). 2013, pp. 7372–7391. DOI: 10.1175/JCLI-D-12-00558.1.
- [PKP+18] Young-Hyang Park, Baek-Min Kim, Gyundo Pak, Masaru Yamamoto, Frédéric Vivier, and Isabelle Durand. "A Key Process of the Nonstationary Relationship between ENSO and the Western Pacific Teleconnection Pattern". In: *Scientific Reports*. Vol. 8 (9512). 2018.
- [PPV+14] Gyundo Pak, Young-Hyang Park, Frederic Vivier, Young-Oh Kwon, and Kyung-Il Chang. "Regime-Dependent Nonstationary Relationship between the East Asian Winter Monsoon and North Pacific Oscillation". In: *Journal of Climate*. Vol. 27 (21). 2014, pp. 8185–8204. DOI: 10.1175/JCLI-D-13-00500.1.
- [RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 9351. 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.
- [RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?: Explaining the Predictions of any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.
- [SSH+20] Chenxi Sun, Moxian Song, Shenda Hong, and Hongyan Li. "A Review of Designs and Applications of Echo State Networks". In: *arXiv*. 2020. DOI: 10.48550/arXiv.2012.02974.
- [ZMG+19] Wenjun Zhang, Xuebin Mei, Xin Geng, Andrew G. Turner, and Fei-Fei Jin. "A Nonstationary ENSO–NAO Relationship Due to AMO Modulation". In: *Journal of Climate*. Vol. 32 (1). 2019, pp. 33–43. DOI: 10.1175/JCLI-D-18-0365.1.

Part II

Papers

Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

Marco Landt-Hayen, Peer Kröger, Martin Claus, Willi Rath

Published (MLTEC 2022)

A. Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

LAYER-WISE RELEVANCE PROPAGATION FOR ECHO STATE NETWORKS APPLIED TO EARTH SYSTEM VARIABILITY

Marco Landt-Hayen¹, Peer Kröger², Martin Claus² and Willi Rath¹

¹GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

²Christian-Albrechts-Universität, Kiel, Germany

ABSTRACT

Artificial neural networks (ANNs) are powerful methods for many hard problems (e.g. image classification or time series prediction). However, these models are often difficult to interpret. Layer-wise relevance propagation (LRP) is a widely used technique to understand how ANN models come to their conclusion and to understand what a model has learned. Here, we focus on Echo State Networks (ESNs) as a certain type of recurrent neural networks. ESNs are easy to train and only require a small number of trainable parameters. We show how LRP can be applied to ESNs to open the black-box. We also show an efficient way of how ESNs can be used for image classification: Our ESN model serves as a detector for El Niño Southern Oscillation (ENSO) from sea surface temperature anomalies. ENSO is a well-known problem. Here, we use this problem to demonstrate how LRP can significantly enhance the explainability of ESNs.

KEYWORDS

Reservoir Computing, Echo State Networks, Layer-wise Relevance Propagation, Explainable AI.

1. INTRODUCTION

Machine learning (ML) provides powerful techniques in the field of artificial intelligence (AI) to discover meaningful relationships in all kinds of data. Within machine learning, artificial neural networks (ANNs) in shallow and deep architectures are found to be promising and very versatile. While these models considerably push the state-of-the-art solutions of many hard problems, they tend to produce black-box results that are difficult to interpret even by ML experts. Consequently, the question of enhancing the explainability of complex models ("explainable AI" or "xAI") has gained a lot of attention in the AI/ML community and stimulated a large amount of fundamental research [1], [2].

In its basic form layers of perceptrons [3] are stacked on top of each other to create a multilayer perceptron (MLP) [4]. These models are usually trained using some form of stochastic gradient descent (SGD) [5]. The aim is to minimize some objective or loss function. More sophisticated architectures e.g. make use of convolutional neural networks (CNNs) [6] or long short term memory (LSTM) [7] units to have recurrence in time in so-called recurrent neural networks (RNNs).

In this paper, we focus on geospatial data, which typically feature non-linear relationships among observations. In this szenario, ANNs are good candidate models, since ANNs are capable of handling complex and non-linear relations by learning from data and training some adjustable

David C. Wyld et al. (Eds): SIGEM, MLTEC, SEAPP, ITCON, NATL, FUZZY, CSEA - 2022

pp. 115-130, 2022. CS & IT - CSCP 2022

DOI: 10.5121/csit.2022.122008

weights and biases [8]. In recent years these methods have been used in various ways on geospatial data [9], [10], [11].

The problem with using ANNs on data of the Earth system is that we often only have relatively short time series to predict on or a small number of events to learn from. Using sophisticated neural networks encounters a large number of trainable parameters and these models are prone to overfitting. This requires a lot of expertise and effort to train these models and prevent them from getting stuck in local minima of the objective function. Famous techniques are dropout, early stopping and regularization [12], [13], [14].

In this work we overcome these problems by using Echo State Networks (ESNs) [15]. ESNs are a certain type of RNNs and have been widely used for time series forecasting [16], [17]. In its basic form an ESN consists of an input and an output layer. In between we find a reservoir of sparsely connected units. Weights and biases connecting inputs to reservoir units and internal reservoir weights and biases are randomly initialized. The input length determines the number of recurrent time steps inside the reservoir. We record the final reservoir states and only the output weights and bias are trained. But opposed to other types of neural networks, this does not encounter some gradient descent methods but is rather done in a closed-form manner by applying linear regression of final reservoir states onto desired target values to get the output weights and bias. This makes ESN models extremely powerful since they require only a very small number of trainable parameters (the output weights and bias). In addition to that, training an ESN is easy, fast and leads to stable and reproducible results. This makes them especially suitable for applications in the domain of climate and ocean research.

But as long as ESNs remain black-boxes, there is only a low level of trust in the obtained results and using these kinds of models is likely to be rejected by domain experts. This can be overcome by adopting techniques from computer vision developed for image data to climate data. Layer-wise relevance propagation (LRP) is a technique to trace the final prediction of a multilayered neural network back through its layers until reaching the input space [18], [19]. When applied to image classification, this reveals valuable insights in which input pixels have the highest relevance for the model to come to its conclusion.

Toms, Barnes and Ebert-Uphoff have shown in their work [20] that LRP can be successfully applied to MLP used for classification of events related to some well-known Earth system variability: El Niño Southern Oscillation (ENSO).

This work is inspired by [20] and goes beyond their studies: We also pick the well-known ENSO problem [21]. ENSO is found to have some strong zonal structure: It comes with anomalies in the sea surface temperature (SST) in Tropical Pacific. This phenomenon is limited to a quite narrow range of latitude and some extended region in terms of longitude. We use ESN models for image classification on SST anomaly fields. We then open the black-box and apply LRP to ESN models, which has not been done before - to the best of our knowledge.

SST anomaly fields used in this work are found to be noisy. For this reason, we focus on a special flavour of ESNs, that uses a leaky reservoir because they have been considered to be more powerful on noisy input data, compared to standard ESNs [22]. With the help of our LRP application to ESNs, we find the leak rate used in reservoir state transition to be a crucial parameter determining the memory of the reservoir. Leak rate needs to be chosen appropriately to enable ESN models to reach the desired high level of accuracy.

Our models yield competitive results compared to linear regression and MLP used as baselines. However, ESN models require significantly less parameters and hence prevent our model from

A. Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

overfitting. We even find our reservoirs to be robust against random permutation of input fields, destroying the zonal structure in the underlying ENSO anomalies.

This opens the door to use ESNs on unsolved problems from the domain of climate and ocean science and apply further techniques of the toolbox of xAI [23].

The rest of this work is structured as follows: In Section 2 we briefly introduce basic ESNs and focus on reservoir state transition for leaky reservoirs. We then sketch an efficient way to use ESN models for image classification. Section 3 outlines the concept of LRP in general before we customize LRP for our base ESN models by unfolding the reservoir recurrence. The classification of ENSO patterns and the application of LRP to MLP and ESN models is presented in Section 4. Our models are not only found to be competitive classifiers but also reveal valuable insights in what the models have learned. We show robustness of our ESN model on randomly permuted input samples and visualize how the leak rate determines the reservoir memory. Discussion and conclusion are found in Section 5, followed by technical details on the used ESN and baseline models in the Appendix.

2. ECHO STATE NETWORKS

An ESN is a special type of RNNs and comes with a strong theoretical background [15], [24], [25]. ESN models have shown outstanding advantages over other types of RNNs that use gradient descent methods for training. We use in this work a shallow ESN architecture consisting of an input and output layer. In between we find a single reservoir of sparsely connected units. The weights connecting input layer and reservoir plus the input bias terms are randomly initialized and kept fixed afterwards. We find some recurrence within the reservoir and reservoir weights and biases are also randomly set and not trainable. Reservoir units are sparsely connected with sparsity usually in the range of 20-30%. Further constraints are put to the largest Eigenvalue of the reservoir weight matrix W_{res} . This is required for the reservoir to be stable and show the so-called Echo State Property [26].

Only the output weights and bias are trained by solving a linear regression problem of final reservoir states onto desired target outputs. A sketch of a base ESN model is shown in Figure 1.

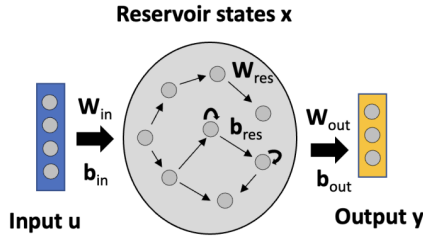


Figure 1. Sketch of base ESN: An input and an output layer, in between we find the reservoir.

In our ESN model, $u(t) \in \mathbb{R}^{D \times 1}$ denotes input values at time t with D input features. Inputs are fed into the model for T time steps, hence $t = 1..T$. Reservoir states at time $t = 1..T$ are denoted by $x(t) \in \mathbb{R}^{N \times 1}$, final reservoir states are obtained as $x(T)$. The final model output $y(T) \in \mathbb{R}^{M \times 1}$ at time T has M output values.

We then find input weights $W_{in} \in \mathbb{R}^{N \times D}$, connecting D input units to N reservoir units. Reservoir weights are given by $W_{res} \in \mathbb{R}^{N \times N}$ and output weights connecting N reservoir units to M output units read $W_{out} \in \mathbb{R}^{M \times N}$. In addition to weight matrices, we have bias vectors $b_{in} \in \mathbb{R}^{N \times 1}$, $b_{res} \in \mathbb{R}^{N \times 1}$ and $b_{out} \in \mathbb{R}^{M \times 1}$ for input, reservoir and output units, respectively.

We use a leaky reservoir with leak rate $\alpha \in [0, 1]$, as discussed in [22]. Leak rate serves as smoothing constant. The larger the leak rate, the faster reservoir states react to new inputs. In other words, the leak rate can be understood as the inverse of the memory time scale of the ESN: The larger the leak rate, the faster the reservoir forgets previous time steps' inputs. The reservoir state transition is defined by Equation 1.

$$x(t) = (1 - \alpha) x(t - 1) + \alpha \text{act}[W_{in}u(t) + b_{in} + W_{res}x(t - 1) + b_{res}] \quad (1)$$

Here $\text{act}(\cdot)$ is some activation function, e.g. *sigmoid* or *tanh*. From the initial reservoir states $x(t = 1)$ we can then obtain further states $x(t)$ for $t = 2..T$ by keeping a fraction $(1 - \alpha)$ of the previous reservoir state $x(t - 1)$. Current time step's input $W_{in}u(t) + b_{in}$ as well as recurrence inside the reservoir $W_{res}x(t - 1) + b_{res}$ are added after applying some activation and multiplying with leak rate α . Reservoir states $x(t)$ are only defined for $t = 1..T$. This requires special treatment of $x(t = 1)$ as outlined in Equation 2.

$$x(t = 1) = \alpha \text{act}[W_{in}u(t) + b_{in}] \quad (2)$$

The model output $y(T)$ is derived as linear combination of output weights W_{out} and biases b_{out} with final reservoir states $x(T)$, as shown in Equation 3.

$$y(T) = W_{out} x(T) + b_{out} \quad (3)$$

This is a linear problem that can be solved in a closed-form manner with multi-linear regression minimizing mean squared error to obtain trained output weights and biases.

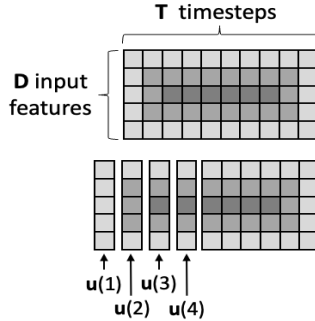


Figure 2. In the upper part we show a synthetic 2D input sample consisting of D input features and T time steps. Feeding the sample column by column into the base ESN model requires breaking the sample into columns. In the lower part we show inputs for the first four time steps.

ESN models have been widely used for time series forecasting [16], [17]. The idea is to feed a single signal or multiple time series of a specific length T into the model. In our work we want to

A. Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

use 2D image data as input samples. This can be done in various ways. One possibility is to flatten 2D image data to obtain a one-dimensional vector and then couple each input to one reservoir unit in the first time step. Without adding additional inputs, the reservoir then swings for some time steps to unfold its dynamics [27]. In this approach the number of reservoir units is directly linked to the number of input units. For high dimensional input data reservoirs can hence become quite large. As mentioned above, we need to put some constraint on the largest Eigenvalue of the reservoir weight matrix W_{res} for stability reasons. Getting the largest Eigenvalue becomes computationally intensive for huge reservoirs and we therefore chose a different approach:

Here we transform images into a temporal signal. This is done by transforming one of the spatial dimensions (longitude) to a temporal one and passing 2D images column-wise into a base ESN model [28]. This is sketched in Figure 2. Feeding an image with dimensions $D \times T$ into a base ESN model is equivalent to having D input time series with length T . This allows using ESN models for image classification.

3. LAYER-WISE RELEVANCE PROPAGATION

LRP was first introduced by Bach et al. in 2015 [18]. LRP aims at understanding decisions of non-linear classifiers like ANNs. It can be used on classification and regression problems. This technique opens the black-box by visualizing the contributions of single input units to model predictions. Resulting relevance scores for an individual input sample can be presented as a heat map and give an intuitive understanding of which parts of the input sample have the highest relevance.

LRP has been successfully applied to various network architectures including MLP, CNN or LSTM models [20], [29]. But to the best of our knowledge, LRP has not been used for ESN models. In this section we will briefly repeat the general idea behind LRP before we customize this technique for using it on base ESN models.

3.1. General idea of LRP

LRP, as presented in [18], does not provide some closed-form solution but rather comes as a set of constraints. Used on image data it serves as a concept for achieving a pixel-wise decomposition of the final model output $y(T)$, as stated in Equation 4.

$$y(T) = \sum_n R_n^{(1)} \quad (4)$$

The model output $y(T)$ is taken as the final or total relevance. The ultimate goal is to decompose the final relevance and find the contributions $R_n^{(1)}$, also referred to as relevance score of each of the n input pixels. Here superscript (1) refers to the first layer, which is the input layer.

To achieve that goal the relevance is traced back from the output layer all the way through lower layers until we finally reach the input layer. In addition to Equation 4 the second constraint is stated in Equation 5.

$$y(T) = \dots = \sum_j R_j^{(l+1)} = \sum_i R_i^{(l)} = \dots = \sum_n R_n^{(1)} \quad (5)$$

This framework guarantees total relevance to be preserved in each layer. For calculating the relevance map for an individual input sample, the trained model weights and biases are fixed. We

then start with the model output as final relevance. A common approach for tracing relevance back through lower layers is by taking only positive contributions of pre-activations into account. This clearly satisfies constraints in Equations 4 and 5. An example is sketched in Figure 3.

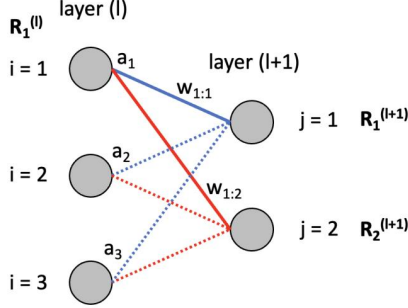


Figure 3. Illustrating the general idea behind LRP: Relevance is traced back from higher to lower layers.

Assume units $j = 1$ and $j = 2$ of layer $(l + 1)$ have known relevance scores $R_1^{(l+1)}$ and $R_2^{(l+1)}$, respectively. This relevance is now distributed on units $i = 1, 2, 3$ of layer (l) . Unit $i = 1$ ends up having relevance $R_1^{(l)}$ from two contributions, as stated in Equation 6: One from unit $j = 1$ and one from unit $j = 2$ of layer $(l + 1)$, indicated by solid blue and red lines in Figure 3, respectively.

$$R_1^{(l)} = \left(\frac{a_1 w_{1:1}}{a_1 w_{1:1} + a_2 w_{2:1} + a_3 w_{3:1}} \right) R_1^{(l+1)} + \left(\frac{a_1 w_{1:2}}{a_1 w_{1:2} + a_2 w_{2:2} + a_3 w_{3:2}} \right) R_2^{(l+1)} \quad (6)$$

Here a_1 , a_2 and a_3 denote activations of units $i = 1, 2, 3$ of layer (l) , respectively and $w_{i:j}$ denotes the weight connecting some unit i from layer (l) with some unit j from subsequent layer $(l + 1)$. This can be simplified using $z_{ij}^+ = \max(a_i w_{i:j}, 0)$, where $+$ denotes that we only consider positive contributions. Relevance $R_{i=i_0}^{(l)}$ for a unit i_0 of layer (l) is stated in Equation 7.

$$R_{i=i_0}^{(l)} = \sum_j \frac{z_{i_0 j}^+}{\sum_i z_{i j}^+} R_j^{(l+1)} \quad (7)$$

3.2. LRP customized for ESN models

Applying LRP to ESNs requires extending the basic methodology described in Section 3.1. Our ESN model consists of an input and an output layer. In between we have the reservoir with recurrence in time. Before we can apply LRP, we need to unfold the reservoir dynamics. Feeding an image consisting of T columns into a base ESN model leads to T time steps to be treated as individual layers. Accordingly, we have inputs $u(t) \in \mathbb{R}^{D \times 1}$ for time steps $t = 1..T$ with D input features.

In Section 2 we introduced our base ESN model including a leaky reservoir with leak rate α . Reservoir state transitions for time steps $t = 2..T$ have been stated in Equation 1. The initial reservoir states $x(1)$ are somewhat special, since there are no previous time step's reservoir states

A. Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

$x(0)$, as we have seen in Equation 2. Figure 4 shows unfolded reservoir dynamics. In addition to that we decomposed reservoir state transitions to visualize distinct contributions separately. As soon as we have trained our base ESN model, the sketch in Figure 4 can be used to understand how the model output is calculated by forward passing an input sample through the network: In this example we have $D = 5$ input features in every time step (shaded yellow), denoted as $u(t)$ for $t = 1..T$. For simplicity we further assume to only have six reservoir units providing reservoir states $x(t)$ for $t = 1..T$ (shaded red). Reservoir states multiplied with $(1 - \alpha)$ contribute to subsequent time step's reservoir states, as sketched in the lower track of Figure 4. The second contribution is given by $\alpha \text{act}(\cdot)$. Here $\text{act}(\cdot)$ (shaded blue) is some appropriate activation function (e.g. *sigmoid* or *tanh*) and takes as argument the current time step's input $W_{in}u(t) + b_{in}$ plus incorporates the recurrence inside the reservoir $W_{res}x(t-1) + b_{res}$. Once we calculated final reservoir states $x(T)$ we obtain model output $y(T)$ as seen in Equation 3 using trained output weights and bias.

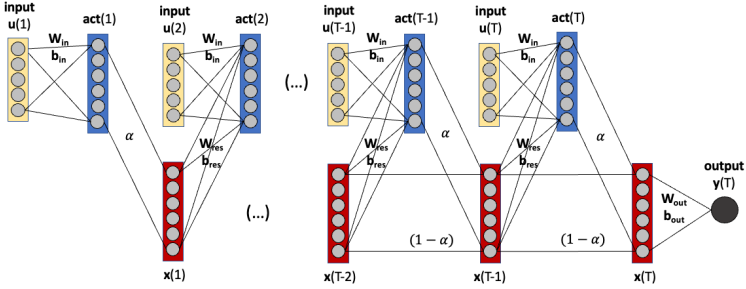


Figure 4. Unfolding our base ESN model in time.

But Figure 4 also illustrates how LRP works for our base ESN model. As usual, we pick an individual input sample and take the model output as final relevance. We then move backwards through all time steps. Opposed to the general concept of LRP, total relevance is not constant from time step to time step. Instead, a part of the total relevance is attributed to each time step's input $u(t)$ and only the remaining relevance is passed on until we reach the initial input $u(1)$. The initial input is special in a way, that it absorbs all residual relevance. For D input features we have $u(t) = (u_1(t), u_2(t), \dots, u_D(t))^T \in \mathbb{R}^{D \times 1}$ for each time step $t = 1..T$. And accordingly, we obtain relevance scores $R^{(t)} = (R_1^{(t)}, R_2^{(t)}, \dots, R_D^{(t)})^T \in \mathbb{R}^{D \times 1}$. These column vectors of relevance scores $R^{(t)}$ need to be combined to get the final relevance map $R \in \mathbb{R}^{D \times T}$, which can be visualized as heat map having the same dimensions as the input samples. Thus, total relevance is still preserved if we customize LRP to ESN models. However, Equations 4 and 5 need to be modified and can be combined to Equation 8.

$$y(T) = \sum_t R^{(t)} = \sum_t \sum_d R_d^{(t)} \quad (8)$$

The final model output $y(T)$ is taken as total relevance and equals the sum of relevance scores $R_d^{(t)}$ with $t = 1..T$ and $d = 1..D$. But as mentioned above, the initial input $u(1)$ absorbs all residual relevance. The residual relevance itself depends on the amount of relevance, that has already been attributed to all other time steps' inputs $u(2), \dots, u(T)$. The speed of decay for total relevance, as it is passed through the layers in a descending order, is controlled by leak rate α . The role of α as memory parameter has been discussed in Section 2. If α is chosen too low, we

find an unreasonably high amount of residual relevance to be assigned to the initial input $u(1)$. To overcome this problem, we add a dummy column of ones as initial column to all input samples. This does not affect model performance, since the additional column is identical for all samples. The relevance $R^{(1)}$ attributed to the dummy column is meaningless in the overall relevance map R and can be omitted.

4. APPLICATION TO ENSO

In this section we will briefly recap the main characteristics of ENSO. Additional details on ENSO can be found e.g. in [20], [21]. We then show results from using MLP and our base ESN model for classifying 2D input samples and open the black-box by applying LRP as described in Sections 3.1. and 3.2. We intentionally choose ENSO as well-known problem to gain confidence in our model and methodology to open the door for applying LRP and further xAI techniques with ESN models on unsolved problems in the context of Earth system and climate research.

4.1. ENSO Patterns

For our studies we use measured monthly mean SST for the years 1880 through 2021, provided by US National Oceanic and Atmospheric Administration. Raw data comes in a 2° by 2° latitude-longitude grid. Each sample consists of 89×180 grid points.

There are several indices used to monitor the sea surface temperature in the Tropical Pacific. All of these indices are based on SST anomalies averaged across a given region. Usually, the anomalies are computed relative to a seasonal cycle estimated from some reference period (climatology) of 30 years (here 1980 through 2009). For our purpose we use SST anomalies averaged over the most commonly used Niño 3.4 region (5°N – 5°S , 120 – 170°W), normalized by its standard deviation over the reference period to obtain a SST anomaly index used to define El Niño and La Niña events, which are associated with anomalous warm and cold SST, respectively. The index is shown in Figure 5.

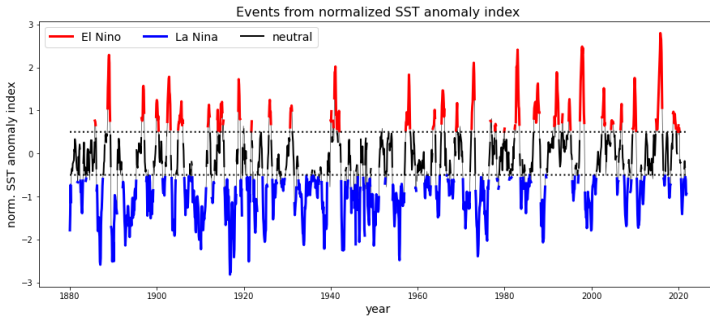


Figure 5. SST anomaly index used to define El Niño and La Niña events.

El Niño is referred to index values ≥ 0.5 (red), whereas La Niña events are referred to index values ≤ -0.5 (blue). In between we find *neutral* states, which are not considered here for classification. The SST anomaly index is used for labelling input samples and also as a single continuous target. In the time span from 1880 through 2021 we have a total number of 1,041

A. Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

samples and split data into train and validation samples, using the first 80% for training (832 samples) and remaining 20% for validation (209 samples). Composite average SST anomaly patterns for El Niño and La Niña are shown in Figure 6.

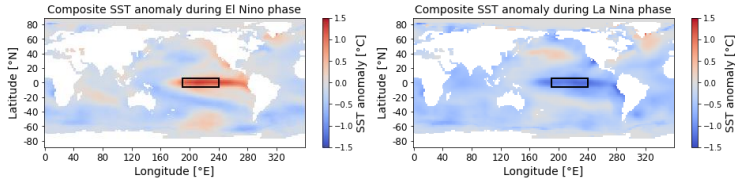


Figure 6. Composite average SST anomaly patterns for El Niño (left-hand side) and La Niña (right-hand side) events. Niño 3.4 region is highlighted by a black rectangle.

4.2. Classification and LRP

As described in Section 4.1 we train our models on 832 SST anomaly fields, where each input sample has dimensions 89×180 (latitude x longitude). SST is not defined over land masses. This reduces the number of valid grid points. Raw data shows some unreasonably high or low values: Here we limit SST anomalies to the range of $[-5^\circ\text{C}, 5^\circ\text{C}]$. Values exceeding these limits are set to upper and lower bound, respectively.

For our baseline models (linear regression and MLP) we vectorize valid grid points as inputs. SST anomalies are scaled to $[-1, 1]$. In any case we use the normalized SST anomaly index shown in Figure 5 as single continuous target. We then transform this regression problem to fit our classification problem by creating binary predictions from model output: Positive predictions refer to El Niño, whereas negative predictions refer to La Niña events.

With this setup we easily reach 100% classification accuracy on both, El Niño and La Niña samples from train and validation data. This perfection was expected, as already shown in [20] and is due to the simplicity of the underlying problem.

For the base ESN model, we do not flatten input samples, as done for the linear regression and MLP approach. Instead, we feed 2D SST anomaly fields into our model and use longitude as time dimension. In other words, we have 89 input features, each consisting of 180 time steps. We deal with invalid grid points by setting SST anomalies to zero after scaling to inputs to $[-1, 1]$. Again, we use normalized SST anomaly index as single continuous target and create binary predictions from model output. Reservoir's leak rate is set to $\alpha = 0.01$.

This also leads to perfect accuracy on El Niño and La Niña, at least on train data. Validation accuracy is found to be 99% for both, El Niño and La Niña. We then focus on El Niño, for which we show the mean relevance maps obtained from MLP and our base ESN model, averaged over all train samples in Figure 7.

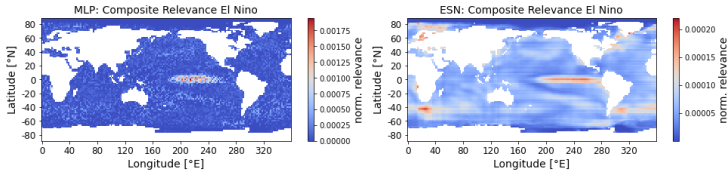


Figure 7. Mean relevance (normalized, unitless) obtained from LRP with MLP (left-hand side) and our base ESN model (right-hand side) on El Niño train samples.

We find the MLP to put its focus only on some narrow, elliptical region inside the Niño region in the Tropical Pacific. This was also found in [20] and appears to be reasonable and efficient to discriminate El Niño from La Niña samples. Compared to that, the mean relevance map obtained from our base ESN model also emphasizes the same spot to come to its conclusion. But in addition to that, we find significantly more structure in mean relevance highlighting other spots outside Niño region to be relevant. High relevance scores are attributed to the area between South Africa and Antarctica.

4.3. Random Permutation

ENSO patterns show some strong *zonal* structure: SST anomalies for both, El Niño and La Niña, are concentrated on some narrow range in latitude and some extended region in longitude. If we want to use our base ESN model to unknown problems, we need to make sure that this approach is also working without having such characteristic zonal structure present. To proof this, we apply some random (but reversible!) permutation on the columns of *all* input samples before training our model, to *shuffle* the order in time. The result is shown in Figure 8.

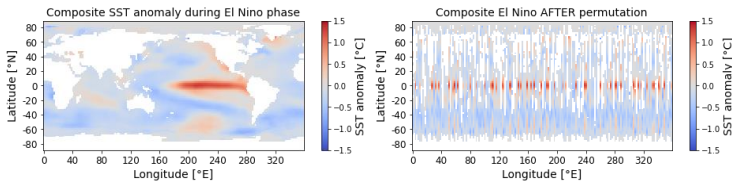


Figure 8. Composite average SST anomaly patterns for El Niño (left-hand side) and the same average SST anomaly AFTER some random permutation of columns (right-hand side).

We then train our base ESN model with unchanged parameters and apply LRP on permuted inputs. The obtained mean relevance map calculated on all El Niño train samples is shown in Figure 9. To restore some more familiar mean relevance map, the permutation needs to be reversed. The result is also shown in in Figure 9.

A. Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

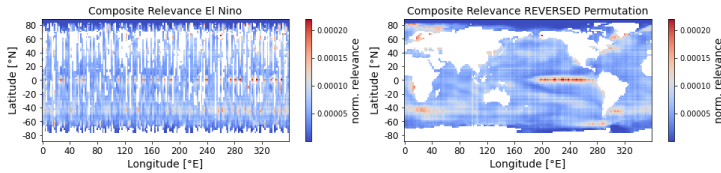


Figure 9. Mean relevance (normalized, unitless) obtained from LRP with base ESN model on PERMUTED El Niño train samples (left-hand side). And restored mean relevance after REVERSED permutation (right-hand side).

We find the *restored* mean relevance map to resemble the *original* mean relevance map, shown in Figure 7. This clearly proves that our approach to pass 2D image data into base ESN models does not rely on the underlying structure in the input data. We also find the same accuracy for base ESN models trained *with* or *without* permuting input columns. This empowers Echo State Networks to be used on unknown problems in the context of climate and ocean science in combination with xAI techniques.

4.4. Fading Memory

In Section 2 we introduced the reservoir state transition as defined by Equation 1. Leak rate α is found to be a crucial parameter. It determines the memory of the reservoir and can be seen as the inverse of the memory time scale of the ESN: The larger the leak rate, the faster the reservoir forgets previous time steps' inputs. Here we use 2D input samples with $T = 180$ time steps for our base ESN model. In other words, we feed a 2D input sample column by column into the model, starting on the left-hand side. This procedure requires α to be chosen low enough to enable the reservoir to remember inputs from all time steps. This is especially important if we apply our method to unknown problems, since we do not know in advance which time steps are most relevant for achieving optimal performance.

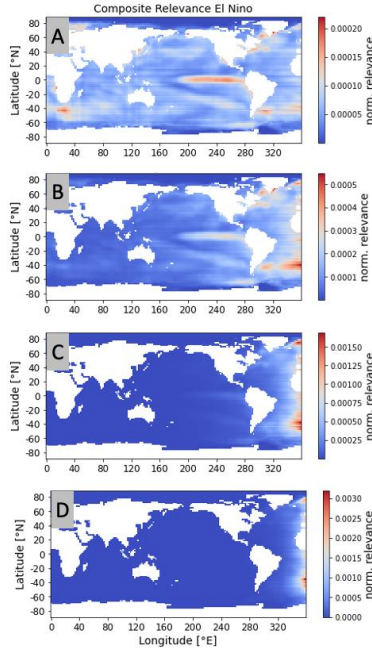


Figure 10. Fading memory effect: Mean relevance (normalized, unitless) obtained from LRP with base ESN model on El Niño train samples for $\alpha = 0.01$ (A), 0.05 (B), 0.2 (C) and 0.4 (D), respectively.

With increasing leak rate, the reservoir memory fades. This effect is visualized in Figure 10. Here we show mean relevance maps for El Niño obtained from ESN models trained with four different leak rates $\alpha = 0.01, 0.05, 0.2$ and 0.4 , respectively. For $\alpha = 0.01$ and 0.05 we find classification accuracy on train samples to be 100%, while validation accuracy reaches 99%. Accordingly, we observe high relevance in the Tropical Pacific region, as seen in relevance maps (A) and (B) in Figure 10. This appears to be reasonable for discriminating ENSO patterns. With further increasing $\alpha = 0.2$ and 0.4 the validation accuracy drops to 95% and 58%, respectively. Mean relevance maps (C) and (D) in Figure 10 explain this decline in model performance: The reservoir simply loses its memory of former input time steps and we find nonzero relevance concentrated on the right-hand side of the relevance maps, representing later time steps. For $\alpha = 0.4$ the model fails to distinguish between El Niño and La Niña samples. An accuracy of only 58% is close to random guessing.

5. DISCUSSION AND CONCLUSION

In this work we successfully used ESNs for image classification and applied LRP to this special type of RNNs, which has not been done before. This enabled us to look inside the model and understand, what the model has learned. LRP is a well-known approach and belongs to the xAI

A. Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

toolbox. Using this technique on a reservoir with $T = 180$ time steps is challenging, but possible. Our proposed LRP customized for ESNs also empowers to study the effect of leak rate α . We found out that α needs to be chosen appropriately to allow the model to take inputs from all time steps into account.

Compared to MLP, mean relevance maps obtained from our ESN model reveal additional structure in terms of high relevance scores outside the Niño region. This needs to be further investigated and could point out existing teleconnections and help to find, where else ENSO leaves its footprint.

We find accuracy to be competitive compared to baseline models (linear regression and MLP). The advantage of ESNs is the low number of trainable parameters, which makes them fast and, thus, easy to train. In addition, our permutation experiments show, that ESN models yield reproducible and stable results. This even holds true if we only have limited train data, as often in the domain of Earth system and climate research. So, we can combine the advantages of ESN models with the power of the broad xAI toolbox. Further techniques to be applied to ESN models on similar problems may be backward optimization, sensitivity analysis or salience maps.

Beyond application to geospatial data, similar ESN models could be used for time series prediction: Instead of feeding 2D images into the model, we may pass a certain number of climate indices with specific input length to an ESN model and LRP could serve as an alternative for the temporal attention mechanism often used in the context of LSTM sequence-to-sequence models. In this way ESN models have good prospects to help understanding known teleconnections in atmospheric science or to find new relationships.

APPENDIX: MODEL DETAILS

In this section we briefly present some technical details on the multilayer perceptron used as baseline model and on our ESN model. The MLP was trained on vectorized SST anomaly fields, where we only considered valid grid points. In this case we worked with 10,988 input values for each sample. The input layer of the MLP consists of the same number of input units. We then have two hidden layers of 8 units each and finally one output unit. For a fully connected MLP we end up with 87,993 trainable weights and biases. We used a linear activation function (identity) for all layers and the Adam optimizer [30] with constant learning rate $lr = 0.0005$. The model was trained over 30 epochs with a batch size of 10. Since we have a regression problem using continuous SST anomaly index as single target, we took the mean squared difference of model output and ground truth as loss function, also referred to as mean squared error loss.

For our base ESN model, the number of reservoir units is set to $n_{res} = 300$. Input and reservoir weights and biases are drawn from a random uniform distribution in $[-0.1, 0.1]$. Reservoir units are only sparsely connected with $sparsity = 0.3$. After initialization the reservoir weights are normalized: The largest Eigenvalue of the reservoir weight matrix is set to 0.8. Leak rate is set to $\alpha = 0.01$. As activation in the reservoir state transition, we use tanh. With this setup our base ESN model only requires 300 trainable output weights plus one output bias, which is significantly less compared to 87,993 trainable parameters for the MLP model.

Raw data used in this work has been uploaded to Zenodo [31]. Annotated Python code can be found in our GitHub repository [32].

ACKNOWLEDGEMENTS

This work was supported by the Helmholtz School for Marine Data Science (MarDATA) funded by the Helmholtz Association (Grant HIDSS-0005).

REFERENCES

- [1] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning," CoRR abs/1806.00069, 2018. <http://arxiv.org/abs/1806.00069>
- [2] Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," CoRR abs/1602.04938, 2016. <http://arxiv.org/abs/1602.04938>
- [3] Frank Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review* vol. 65, pp. 386–408, 1958.
- [4] Hassan Ramchoun, Mohammed Amine, Janati Idrissi, Youssef Ghanou, and Mohamed Ettaouil. "Multilayer Perceptron: Architecture Optimization and Training," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, pp. 26–30, 2016.
- [5] Sebastian Ruder. "An overview of gradient descent optimization algorithms," CoRR abs/1609.04747, 2016. <http://arxiv.org/abs/1609.04747>
- [6] Keiron O'Shea and Ryan Nash. "An Introduction to Convolutional Neural Networks," CoRR abs/1511.08458, 2015. <http://arxiv.org/abs/1511.08458>
- [7] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory," *Neural Computation* vol. 9, no. 8, pp. 1735–1780, 1997.
- [8] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. "Deep Learning," MIT Press, Cambridge, MA, USA, 2016. <http://www.deeplearningbook.org>
- [9] Ankur Mahesh, Maximilian Evans, Garima Jain, Climateai Mattias, Castillo Climateai, Aranildo Lima, Climateai Brent, Lughino Climateai, Himanshu Gupta, Climateai Carlos, Gaitan Climateai, Jarrett Climateai, Omeed Tavasoli, Patrick Brown, and V Balaji. "Forecasting El Niño with Convolutional and Recurrent Neural Networks," In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, pp. 8–14, 2019.
- [10] Arvind W. Kiwelekar, Geetanjali S. Mahamunkar, Laxman D. Netak, and Valmik B.Nikam. "Deep Learning Techniques for Geospatial Data Analysis," CoRR abs/2008.13146, 2020. <https://arxiv.org/abs/2008.13146>
- [11] Jinah Kim, Minho Kwon, Sung-Dae Kim, Jong-Seong Kug, Joon-Gyu Ryu, and Jaeh Kim. "Spatiotemporal neural network with attention mechanism for El Niño forecasts," *Scientific Reports*, vol. 12, no. 7204, 2022.
- [12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [13] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. "Understanding and Improving Early Stopping for Learning with Noisy Labels," CoRR abs/2106.15853, 2021. <https://arxiv.org/abs/2106.15853>
- [14] Anders Krogh and John A. Hertz. "A Simple Weight Decay Can Improve Generalization," In Proceedings of the 4th International Conference on Neural Information Processing Systems (NIPS 1991), San Francisco, CA, USA, pp. 950–957, 1991.
- [15] Chenxi Sun, Moxian Song, Shenda Hong, and Hongyan Li. "A Review of Designs and Applications of Echo State Networks," CoRR abs/2012.02974, 2020. <https://arxiv.org/abs/2012.02974>
- [16] Tachwan Kim and Brian King. "Time series prediction using deep echo state networks," *Neural Computing and Applications*, vol. 32, 2020
- [17] Claudio Gallicchio, Alessio Micheli, and Luca Pedrelli. "Deep reservoir computing: A critical experimental analysis," *Neurocomputing*, vol. 268, pp. 87–99, 2017.
- [18] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS ONE*, vol. 10, 2015.

A. Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability

- [19] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognition*, vol. 65, pp.211-222, 2017.
- [20] Benjamin A. Toms, Elizabeth A. Barnes, and Imme Ebert-Uphoff, "Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 9, 2020.
- [21] C. F. Ropelewski and M. S. Halpert, "North American Precipitation and Temperature Patterns Associated with the El Niño Southern Oscillation (ENSO)," *Monthly Weather Review*, vol. 114, no. 12, p. 2352, 1986.
- [22] Herbert Jaeger, Mantas Lukosevicius, Dan Popovici, and Udo Siewert, "Optimization and applications of echo state networks with leaky-integrator neurons," *Neural Networks: The official journal of the International Neural Network Society*, vol. 20, p. 335, 2007
- [23] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," arXiv:1312.6034, 2014.
- [24] Herbert Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, vol. 148, p. 34, 2001.
- [25] Herbert Jaeger, "Adaptive Nonlinear System Identification with Echo State Networks," In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS 2002)*, Vancouver, Canada, pp. 609–616, 2002.
- [26] Herbert Jaeger, "Echo State Network," *Scholarpedia*, vol. 2, no. 9, p. 2330, 2007.
- [27] Alexander Woodward and Takashi Ikegami, "A Reservoir Computing approach to Image Classification using Coupled Echo State and Backpropagation Neural Networks," In *Proceedings of the 26th International Conference on Image and Vision Computing*, Auckland, New Zealand, p. 543, 2011.
- [28] Nils Schaetti, Michel Salomon, and Raphaël Couturier, "Echo State Networks-Based Reservoir Computing for MNIST Handwritten Digits Recognition," In *Proceedings of the IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) and 15th International Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, pp. 484–491, 2016.
- [29] Yeon-Jee Jung, Seung-Ho Han, and Ho-Jin Choi, "Explaining CNN and RNN Using Selective Layer-Wise Relevance Propagation," *IEEE Access* vol. 9, pp. 18670–18681, 2021.
- [30] Kingma, Diederik P. and Ba, Jimmy, "Adam: A Method for Stochastic Optimization," In *Proceedings of the 3rd International Conference for Learning Representations*, San Diego, CA, USA, 2014. <https://arxiv.org/abs/1412.6980>
- [31] Marco Landt-Hayen, "NOAA Extended Reconstructed SST V5 as Supplementary Data for Publication," 2022. <https://doi.org/10.5281/zenodo.6517186>
- [32] https://github.com/MarcoLandtHayen/ESN_Classification_LRP_ENSO

Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

Marco Landt-Hayen, Willi Rath, Martin Claus, Peer Kröger

Accepted for Publication at MLTEC 2023

B. Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures*

Marco Landt-Hayen^{1,2}, Willi Rath², Martin Claus^{1,2}, and Peer Kröger¹

¹ Christian-Albrechts-Universität zu Kiel, Germany

² GEOMAR Helmholtz Centre for Ocean Research, Germany

Abstract. Layer-wise relevance propagation (LRP) is a widely used and powerful technique to reveal insights into various artificial neural network (ANN) architectures. LRP is often used in the context of image classification. The aim is to understand, which parts of the input sample have highest relevance and hence most influence on the model prediction. Relevance can be traced back through the network to attribute a certain score to each input pixel. Relevance scores are then combined and displayed as heat maps and give humans an intuitive visual understanding of classification models. Opening the black box to understand the classification engine in great detail is essential for domain experts to gain trust in ANN models. However, there are pitfalls in terms of model-inherent artifacts included in the obtained relevance maps, that can easily be missed. But for a valid interpretation, these artifacts must not be ignored. Here, we apply and revise LRP on various ANN architectures trained as classifiers on geospatial and synthetic data. Depending on the network architecture, we show techniques to control model focus and give guidance to improve the quality of obtained relevance maps to separate facts from artifacts.

Keywords: Artificial Neural Networks, Image Classification, Layer-wise Relevance Propagation, Geospatial Data, Explainable AI.

1 Introduction

Image classification refers to assigning one or more class labels to a two-dimensional image. Here, we focus on binary classification problems and have only two distinct classes. Instead of having a discrete class label, one often assigns some continuous target value to each sample. The class label is then derived from the target value by defining certain thresholds. Like this, the initial classification problem becomes a regression problem. The task is then to predict some continuous target value from all input pixels. Generally, the relationship between inputs and targets is highly nonlinear. Artificial neural networks (ANNs) are state-of-the-art for this task. In this work, we will revise the use of various ANN architectures for image classification. In particular, we work with multilayer perceptrons (MLPs), convolutional neural networks (CNNs) and Echo State Networks (ESNs). MLP models have frequently been used for image classification [1–3]. CNN architectures aim to detect objects and structures in the underlying samples and are powerful tools for image classification [4, 5], object detection [6] or semantic segmentation [7]. With their inherent shared weights philosophy, CNNs need less trainable parameters, compared to MLPs. ESNs differ from MLPs and CNNs, as they are a special type of recurrent neural networks (RNNs), originally designed for time series forecasting [8, 9]. To use two-dimensional images as inputs, samples may be sliced column- or row-wise, respectively. Like this, either the x - or the y -dimension is turned into a time dimension, as done in [10].

As input data, we use real world geospatial data. In particular, we choose a well-understood Earth System Variability, the El Niño Southern Oscillation (ENSO). The current ENSO phase can be detected from spatially averaged sea surface temperature (SST)

* This work was supported by the Helmholtz School for Marine Data Science (MarDATA) funded by the Helmholtz Association (Grant HIDSS-0005).

anomalies in the so-called Niño 3.4 region in the Tropical Pacific. But ENSO is a complex phenomenon and is related to climate anomalies over large distances, also referred to as teleconnections. ENSO leaves its footprint outside the Tropical Pacific, e.g. at the African West coast near Angola [11]. Here, we use ANN models as ENSO detectors on two-dimensional samples of SST anomalies. The two classes are either "El Niño" or "La Niña", which can be associated with unusual warm or cold sea surface temperature in the Niño region, respectively.

While all models perfectly perform on this simple classification task, our aim is *not* to find a superior classifier. Here, we are interested in getting a deeper understanding of how the models come to their conclusions and explain the reasoning of the models' classification engine by using ENSO as a toy problem. A variety of techniques from the domain of explainable artificial intelligence (xAI) exists to open the black box of ANNs. In backward optimization [12], the inputs for a trained model are modified to maximize the network's confidence in the obtained output. The goal is to generate some optimal input. A saliency map [13] aims to highlight areas in a given input sample that show strongest support towards a given class. Other techniques compute the gradient of a prediction with respect to the input pixels in a sensitivity analysis [14], to reveal insights in how sensitive the model output depends on slight modifications of input values. Our purpose is more general. We want to find each input pixel's contribution to the model output. This can be achieved by layer-wise relevance propagation (LRP). The concept of LRP has been introduced by Bach et al. [15]. The model prediction is taken as final relevance and is then traced back through all layers until reaching the input space to assign individual relevance scores to each input pixel. LRP has been extended from a practical point of view by Montavon et al. [16]. In their work, they provide a set of concrete propagation rules for tracing relevance back through the layers.

Toms et al. [1] recently applied LRP to MLP models trained as ENSO detector. They show mean relevance maps obtained from all El Niño samples. Significant relevance is only found in the Niño region, as one might expect on the first sight. There are no other spots of high relevance outside this area. In our earlier work [10], we proposed to use ESN models on the same task and also performed LRP on ENSO. We found mean relevance maps for El Niño samples with a far more subtle structure, also highlighting Niño region but highest relevance is found on the passage between South Africa and Antarctica.

Moreover, the leak rate in the reservoir transition of an ESN determines the reservoir memory and hence is a crucial parameter [10]. Here, we give a heuristic for setting the leak rate in relation to the number of time steps.

In addition to MLP and ESN models, we also apply LRP to CNN models in this work. For MLPs and CNNs used for image classification, regularization of weights is found to be essential. We can force MLP and CNN models to focus only on a very narrow spot in the Tropical Pacific to discriminate El Niño from La Niña, by driving small weights to zero to favor sparse weight matrices in the training process [17].

To get a deeper understanding, we step back and first apply MLP, CNN and ESN models to synthetic samples, where we exactly know the relationship between inputs and outputs by design. Only with a sound understanding of pitfalls related to LRP, we can use this powerful technique on unknown problems to help domain experts to explain their models, gain trust in models' predictions and avoid misinterpretations. Like this, LRP bears good prospects to help finding new relationships and understanding unknown teleconnections [18–20].

B. Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

Our main contributions are as follows:

- We revise LRP on various ANN architectures used for image classification.
- We show how to control the model focus of MLP and CNN networks by weight regularization.
- For ESNs, we find the leak rate to be the crucial parameter and give a heuristic to choose it appropriately.
- Some of the additional spots of high relevance found in mean relevance maps from ESN models are identified as artifacts. Furthermore, we propose a technique to erase these artifacts.

The rest of this work is structured as follows: In Section 2 we briefly introduce synthetic and real world data used for our experiments. Section 3 outlines the concept of LRP customized to MLP, CNN and ESN models. The application of LRP to synthetic and real world ENSO patterns is presented in Section 4. A detailed discussion and conclusion is found in Section 5, followed by technical details, remarks on availability of data and annotated code in the Appendix.

2 Data

In this section, we introduce SST anomaly fields related to ENSO used as real world data. After that, we describe the design of our synthetic samples. In this work, we use two-dimensional monthly mean SST anomaly fields for the years 1880 through 2021 as real world inputs. Raw SST data is provided by the NOAA Physical Sciences Laboratory [21]. Input samples consist of 89×180 grid points on a 2° by 2° latitude-longitude grid. Several indices are used to monitor ENSO activity based on SST anomalies in the Tropical Pacific [22]. Morrow et al. [23] define in their work an index from the Niño 3.4 region (5°N - 5°S , 120 - 170°W), which we use as target in this work. El Niño and La Niña events

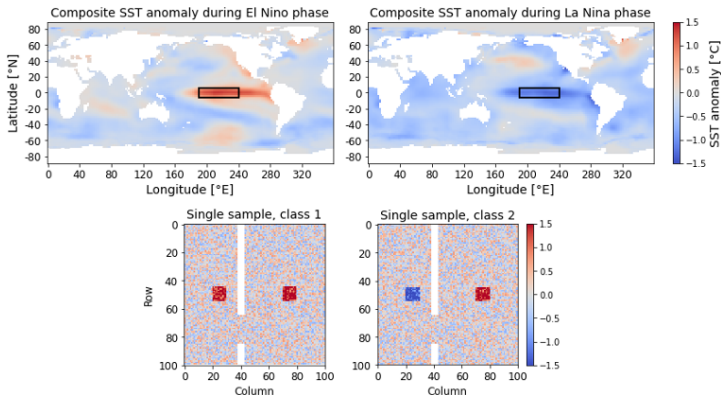


Fig. 1. Upper part: Composite average SST anomaly patterns for El Niño (left-hand side) and La Niña (right-hand side) events. A black rectangle highlights Niño 3.4 region. Lower part: Exemplary samples for classes 1 (left-hand side) and 2 (right-hand side) of synthetic data.

are characterized by positive and negative SST anomalies in the Niño 3.4 region, referred to index values ≥ 0.5 and ≤ -0.5 , respectively. Neutral states in between, are not considered for classification, here. The total number of 1,041 samples is split, using the first 80% as training data (832 samples) and the remaining 20% for validation (209 samples). The upper part of Figure 1 shows composite average SST anomaly fields for El Niño and La Niña.

As synthetic data, we create quadratic samples of dimension 100×100 pixels and define two distinct classes, class 1 and class 2, respectively. On the right-hand side of each sample, we have a vertically centered square of value +1, identical for both classes. The left-hand side also contains a vertically centered square of values +1 and -1 for classes 1 and 2, respectively, as equivalent for SST anomalies in the Niño region in real world samples. Classes can be uniquely discriminated by the square on the left-hand side, while the other only serves as an additional but irrelevant feature. We add random noise to all samples, drawn from a uniform distribution in the interval $[-0.5, 0.5]$. In addition to that, we include some vertical regions of invalid grid points with a small gateway to all samples, similar to land masses in real world samples, where SST anomalies are not defined. Target values are +1 and -1 for classes 1 and 2, respectively. Training data consists of 400 samples, 200 for each class, while validation data consists of 100 samples, 50 for each class. The lower part of Figure 1 shows single samples for both classes as example.

3 Methods and Models

In this section, we briefly recap the ANN architectures and sketch the concept of LRP customized to MLP, CNN and ESN models. MLPs are feedforward networks and consist of a specified number of layers. The input layer is connected to the input samples. This requires the number of input units to equal the number of input pixels. Two-dimensional samples are therefore transformed into a one-dimensional vector. CNNs are also feedforward networks. The convolution is done by specified kernels scanning the input image and producing so-called feature maps. After a desired number of convolutions and optionally pooling operations, we obtain the final feature maps. Here, we flatten these final feature maps and stack fully connected dense layers on top of the underlying CNN part. The basic form of an ESN model consists of an input and an output layer and a reservoir of sparsely connected units in between. Once randomly initialized, the input and reservoir weights and biases are kept fixed. We then feed a certain number of input features for a specified number of time steps into the model and record the final reservoir states. Only the output weights and bias are trained by regressing final reservoir states onto targets. There is no backpropagation of errors in the training process, as for MLPs or CNNs. Equation 1 states the initial reservoir states $x(t = 1)$ for our ESN model. We have a leaky reservoir with leak rate $\alpha \in [0, 1]$, as discussed in [24]. For larger leak rates, the reservoir states react faster to new inputs. $u(t)$ denote current time step's inputs, W_{in} and b_{in} are input weights and biases, respectively.

$$x(t = 1) = \alpha \text{act}[W_{in}u(t = 1) + b_{in}] \quad (1)$$

Reservoir state transition for subsequent time steps is outlined in Equation 2. A fraction $(1 - \alpha)$ of the previous reservoir states $x(t - 1)$ is kept and combined with the term inside the activation $\text{act}(\cdot)$, here \tanh . $W_{res}x(t - 1) + b_{res}$ denotes the recurrence inside the reservoir, with reservoir weights and biases W_{res} and b_{res} , respectively.

$$x(t) = (1 - \alpha)x(t - 1) + \alpha \text{act}[W_{in}u(t) + b_{in} + W_{res}x(t - 1) + b_{res}] \quad (2)$$

B. Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

Toms et al. [1] describe in their work, how LRP can be applied to MLP and CNN models, as both are feedforward networks. The model output Y is taken as final relevance and is then traced back through all layers until reaching the input space to assign relevance scores $R_n^{(1)}$ to each of the n input pixels, as stated in Equation 3.

$$Y = \sum_n R_n^{(1)} \quad (3)$$

Starting with model output Y , relevance is distributed back from the output layer through lower layers until we reach the input layer. To preserve total relevance in each layer, we have a second constraint, as stated in Equation 4.

$$Y = \dots = \sum_j R_j^{(l+1)} = \sum_i R_i^{(l)} = \dots = \sum_n R_n^{(1)} \quad (4)$$

Samples of class 1 have positive target values by design. For this reason, we consider only positive contributions of pre-activations as propagation rule. Assume we have j units in layer $(l+1)$ with known relevance scores $R_j^{(l+1)}$, the relevance $R_{i=i_0}^{(l)}$ for unit i_0 of layer (l) is obtained by Equation 5, using $z_{ij}^+ = \max(a_i w_{i,j}, 0)$. Here, a_i denotes activation of units i of layer (l) and $w_{i,j}$ denotes the weight connecting unit i from layer (l) with unit j from layer $(l+1)$. For class 2 samples, we have negative targets. We start with absolute prediction values as final relevance, since relevance is defined to be positive, and only consider negative contributions of pre-activations.

$$R_{i=i_0}^{(l)} = \sum_j \frac{z_{i_0 j}^+}{\sum_i z_{ij}^+} R_j^{(l+1)} \quad (5)$$

This LRP concept needs to be modified for our ESN models, as described in great detail in [10], since we have a recurrence in time. Reservoir transition needs to be unfolded. Each time step is equivalent to one layer. A fraction of the total relevance is attributed to each time step’s inputs. Remaining relevance is passed on until we reach the initial input. Finally, all relevance scores attributed to all time step’s inputs can be composed to obtain a relevance map with same dimensions as the input sample.

4 Results

In this section, we present results from LRP experiments with MLP, CNN and ESN models on synthetic and real world data. By design, we know that our synthetic samples can uniquely be discriminated by looking at the left square. Figure 2 shows mean relevance maps for class 1 obtained from MLP (upper part) and CNN models (lower part) with and without regularization of hidden layer’s weights. Without regularization, we find blurry mean relevance maps for both models with highest relevance on both squares. By adding regularization terms, we successfully force both models to focus on the left square. For the CNN we find artifacts in form of quadratic patches in mean relevance maps.

For ESN models, only the output weights are trained by regressing final reservoir states onto targets. This limits the use of weight regularization. However, for ESN models the leak rate α determines reservoir dynamics and memory. Figure 3 shows mean relevance maps for ESN models, feeding class 1 samples column-wise into the model, starting with the leftmost column. For $\alpha = 0.005$, we find an equal amount of summed relevance over both squares. But highest relevance is found on the gateway in the barrier of invalid grid points. With increasing leak rate ($\alpha = 0.05$) the reservoir’s memory starts to fade, summed

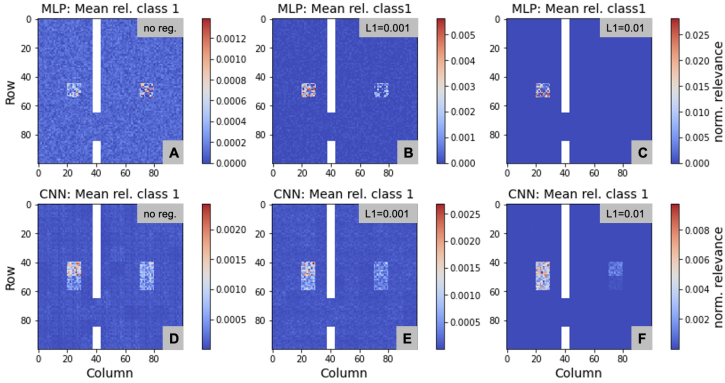


Fig. 2. Mean relevance maps for MLP (upper part) and CNN (lower part) on class 1 for different weight regularization. A, B, C: MLP without regularization, $L1=0.001$ and $L1=0.01$, respectively. D, E, F: CNN without regularization, $L1=0.001$ and $L1=0.01$, respectively.

relevance on the left square is decreased. Still, the model succeeds to classify all samples correctly. At some point ($\alpha = 0.2$) the reservoir loses its memory of the left half. All relevance is bundled rightmost and classification accuracy for validation samples drops to random guessing. In this case, $\alpha = 0.005$ is a suitable choice for the leak rate, since it guarantees all time steps' inputs to be equally considered.

With that choice of the leak rate, we compare various techniques to use two-dimensional input samples for ESN models. Figure 4 shows mean relevance maps for ESN models, feeding class 1 samples column- or row-wise into the model. We find horizontal and vertical stripes, respectively. We don't find increased relevance on the gateway when feeding samples row-wise into the model. As third approach, we propose to split input samples into equal-sized pieces, considering only valid grid points. The resulting mean relevance map for class 1 samples is also shown in Figure 4. We find a perfectly equal amount of summed relevance over both squares. Besides that, there are no artifacts in form of stripes and we don't find any increased relevance on the gateway.

Switching to real world samples, the upper part of Figure 5 shows mean relevance maps

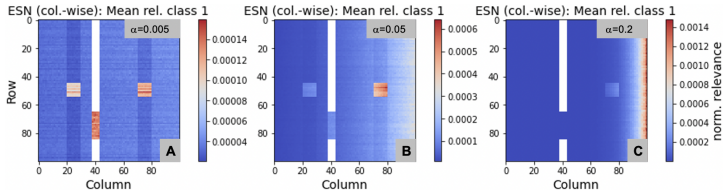


Fig. 3. Mean relevance maps for ESN on class 1, feeding samples column-wise into the model. Fading memory experiment with various leak rates. A: $\alpha = 0.005$, B: $\alpha = 0.05$, C: $\alpha = 0.2$.

B. Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

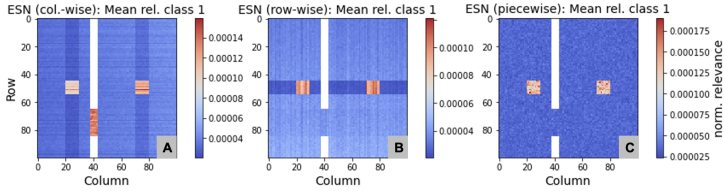


Fig. 4. Mean relevance maps for class 1 samples from ESN models using different techniques to feed two-dimensional inputs into the models. A: column-wise, B: row-wise, C: piecewise.

for El Niño, obtained from the MLP and CNN model with regularization ($L1=0.01$). We find strong and only focus inside Niño region. The lower part of Figure 5 shows MLP results without regularization. We find two complementary modes on the exact same parameter setting.

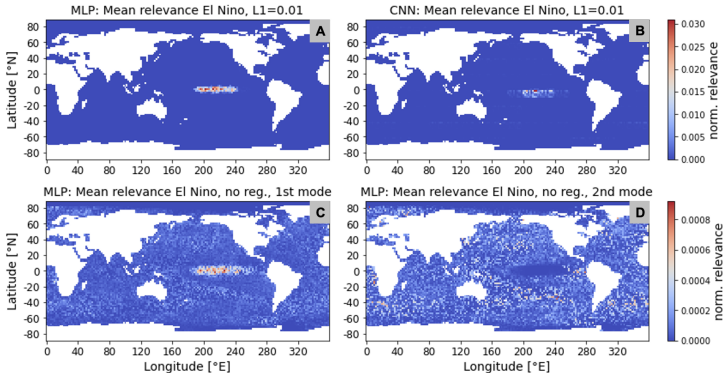


Fig. 5. Mean relevance maps for El Niño using MLP and CNN models with different weight regularization. A: MLP, $L1=0.01$, B: CNN, $L1=0.01$, C and D: MLP 1st and 2nd mode, respectively, no regularization.

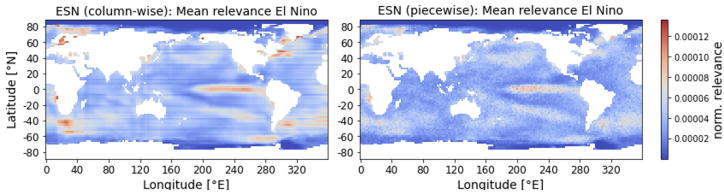


Fig. 6. Mean relevance maps for El Niño from ESN models using different techniques to feed inputs into the model. A: column-wise, B: piecewise.

Figure 6 compares mean relevance maps for El Niño, obtained from ESN models, feeding samples column-wise and as equal-sized pieces, respectively. In both cases, we find high relevance inside the Tropical Pacific. However, feeding samples column-wise, highest relevance is found on the passage between South Africa and Antarctica and we find horizontal stripes.

Table 1. Evaluation metrics for selected models trained on synthetic samples: mse on training and validation data, the ratio of summed mean relevance over both squares in relation to total relevance, the relative amount of summed mean relevance over the left square in relation to both squares, and the number of trainable parameters.

	MLP no reg.	MLP L1=0.01	CNN L1=0.01	ESN piecewise
mse_{train}	0.01	0.001	0.005	0.001
mse_{val}	0.053	0.002	0.008	0.180
$\frac{\sum_{both\ sq} rel}{\sum_{total} rel}$	7%	92%	40%	7%
$\frac{\sum_{left\ sq} rel}{\sum_{both\ sq} rel}$	48%	100%	98%	50%
trainable parameters	19,205	19,205	2,223	301

5 Discussion and Conclusion

In this section, we will discuss all results from the previous section in detail and give an outlook on future research. We have shown, that our MLP and CNN models can be successfully forced to focus mainly on the left square of our synthetic samples, which is sufficient to fulfill the classification task. For our CNN model we find artifacts in form of quadratic patches in mean relevance maps, due to shared kernel weights. When it's L1 weight regularization found to be the key to force MLP and CNN models to focus, we find the leak rate to be crucial for ESN models to determine reservoir memory. Leak rate needs to be chosen appropriately, to equally take all time steps' inputs into account, since we don't know in advance, which time steps contain most relevant information. Using two-dimensional inputs column- or row-wise, we find horizontal and vertical stripes, respectively, which slightly differ for individual training runs. Input samples don't show any stripes. This, and the fact, that we find stripes to depend on how we feed sample into the model, clearly unmasks these stripes to be artifacts, stemming from random initialized input weights. Highest relevance for ESN models is found on the small gateway only, when we feed samples column by column. This effect of *squeezed* relevance is caused by the varying number of valid grid points for different columns. The number of valid grid points per row also varies, but the effect is much smaller and can hardly be recognized. As a solution, we propose a new technique of considering only valid grid points. These grid points are transformed to a one-dimensional vector, permuted and split into equal-sized pieces. Like this, stripes disappear and we no longer find elevated relevance on the gateway. However, both squares are found to be equally relevant, since we cannot apply weight regularization on ESN models.

From a practical point of view, the model choice in combination with parameter setting is a tradeoff between explainability and performance. Table 1 shows evaluation metrics for selected models trained on synthetic samples. While all models perfectly perform the classification task with 100% accuracy, the mean squared error (mse) of model predictions compared to true targets and the model focus varies. MLP with weight regularization shows lowest mse and focuses completely on the left square. This leads to high performance

B. Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

in terms of efficiently discriminating classes but low explainability, since all features with minor relevance are suppressed. The ESN model with feeding samples as equal-sized pieces comes with a higher mse on validation samples compared to MLP and CNN models and thus lower performance on the classification task. However, mean relevance maps from ESN models highlight all features involved in the underlying input samples, that could reveal valuable insights into existing teleconnections and thus offer higher explainability.

Trained on real world data, our MLP and CNN models with weight regularization act as a very efficient classifier and only focus on a narrow spot in the Niño region. Without regularization, the MLP model reveals more structure in mean relevance maps. Moreover, we find two complementary modes, due to random initialization of weights and biases and the stochastic nature of the learning algorithm. The first mode shows strong focus on the Niño region, while the second mode completely ignores the very same region. With this second mode, the MLP still successfully discriminates ENSO patterns by encountering information from outside Tropical Pacific.

Using real world samples column-wise for training our ESN model, we again find horizontal stripes in the obtained mean relevance map, that have already been identified as artifacts in experiments on synthetic data. Highest relevance is found in the region between South Africa and Antarctica. Since SST input data is only defined over the ocean, the number of valid grid points is smallest for these columns, forcing relevance to be compressed onto the remaining valid grid points. Again, splitting input samples into equal-sized pieces erases both artifacts in mean relevance maps obtained from ESN models. Like this, our proposed method enhances the explainability for ESN models.

Overall, neither model is found to be superior for image classification with LRP. While MLP and CNN models appear to be *efficient* in focusing on most relevant features, mean relevance maps of ESNs reveal valuable information on existing teleconnections. It is also worth to mention that ESN models come with a significantly lower number of trainable parameters, allowing these models to be trained on less training data without overfitting. With this advantage, ESNs have good prospects for further use on geospatial data and could also be used in combination with LRP in the context of time series prediction, when we replace two-dimensional inputs by a number of e.g. climate indices. This approach could then serve as alternative for long short-term memory (LSTM) models with attention mechanisms.

Appendix

In this work we only show reproducible results. All models have been implemented in Python (version 3.9.9) using Tensorflow (version 2.4.1). To keep it as transparent as possible, we provide data and annotated Python code in Jupyter notebooks containing all experiments and details on models and methods [25].

References

1. Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I.: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability, *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 9 (2020)
2. Coskun, N. and Yildirim, T.: The effects of training algorithms in MLP network on image classification, *Proceedings of the International Joint Conference on Neural Networks*, vol. 2, pp. 1223-1226 (2003), DOI: 10.1109/IJCNN.2003.1223867
3. Shubathra, S., Kalaivaani, P., and Santhoshkumar, S.: Clothing Image Recognition Based on Multiple Features Using Deep Neural Networks, *International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 166-172 (2020), DOI: 10.1109/ICESC48915.2020.9155959.

4. Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Communications of the ACM*, vol. 60, issue 6, pp. 84–90 (2017), DOI: 10.1145/3065386
5. Kadam S. S., Adamuthe A. C., and Patil, A.: CNN Model for Image Classification on MNIST and Fashion-MNIST Dataset, vol. 64, issue 2, pp. 374-384 (2020), DOI:10.37398/jsr.2020.640251
6. Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., and Yuille, A.: The role of context for object detection and semantic segmentation in the wild, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 891–898 (2014)
7. Redmon, J., Divvala, S., Girshick, R., and Farhadi, A.: You only look once: Unified, real-time object detection, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 779–788 (2016)
8. Kim, T. and King, B.: Time series prediction using deep echo state networks, *Neural Computing and Applications*, vol. 32 (2020)
9. Gallicchio, C., Micheli, A., and Pedrelli, L.: Deep reservoir computing: A critical experimental analysis, *Neurocomputing*, vol. 268, pp. 87–99 (2017)
10. Landt-Hayen, M., Kröger, P., Claus, M., and Rath, W.: Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability, In *Proceedings of the 3rd International Conference on Machine Learning Techniques (MLTEC 2022)*, Zurich, Switzerland, vol. 12, no. 20, pp. 115-130 (2022), DOI: 10.5121/csit.2022.122008
11. Shannon, L. V., Agenbag J. J., and Buys M. E. L.: Large- and mesoscale features of the Angola-Benguela front, *South African Journal of Marine Science*, vol. 5, no. 1, pp. 11-34 (1987), DOI: 10.2989/025776187784522261
12. Olah, C., Schubert, L., and Mordvintsev, A.: Feature Visualization, *Distill* (2017), <https://distill.pub/2017/feature-visualization>, DOI: 10.23915/distill.00007
13. Simonyan, K., Vedaldi, A., and Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, *arXiv* (2013), <https://arxiv.org/abs/1312.6034>, DOI: 10.48550/ARXIV.1312.6034
14. Nourani, V. and Fard, M. S.: Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes, *Advances in Engineering Software*, vol. 47, pp. 127-146 (2012)
15. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller K.-R., and Samek W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, *PLoS ONE*, vol. 10 (2015)
16. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R.: Explaining nonlinear classification decisions with deep Taylor decomposition, *Pattern Recognition*, vol. 65, pp. 211-222 (2017)
17. Krogh, A. and Hertz, J. A.: A Simple Weight Decay Can Improve Generalization, In *Proceedings of the 4th International Conference on Neural Information Processing Systems (NIPS 1991)*, San Francisco, USA, pp. 950–957 (1991)
18. Pak, G., Park, Y.-H., Vivier, F., Kwon, Y.-O., and Chang, K.-I.: Regime-Dependent Nonstationary Relationship between the East Asian Winter Monsoon and North Pacific Oscillation, *Journal of Climate*, vol. 27, no. 21, pp. 8185–8204 (2014)
19. Park, Y.-H., Kim, B.-M., Pak, G., Yamamoto, M., Vivier, F., and Durand, I.: A key process of the nonstationary relationship between ENSO and the Western Pacific teleconnection pattern, *Scientific Reports*, vol. 8, no. 9512 (2018)
20. Zhang, W., Mei, X., Geng, X., Turner, A. G., and Jin, F.-F.: A Nonstationary ENSO–NAO Relationship Due to AMO Modulation, *Journal of Climate*, vol. 32, no. 1, pp. 33-43 (2019)
21. <https://downloads.psl.noaa.gov/Datasets/noaa.ersst.v5/sst.mmmean.nc>
22. <https://climatedataguide.ucar.edu/climate-data/>
23. Morrow, R., Ward, M. L., Hogg, A. McC., and Pasquet, S.: Eddy response to Southern Ocean climate modes, *Journal of Geophysical Research*, vol. 115, DOI: 10.1029/2009JC005894 (2010)
24. Jaeger, H., Lukosevicius, M., Popovici, D., and Siewert, U.: Optimization and applications of echo state networks with leaky-integrator neurons, *Neural Networks: The official journal of the International Neural Network Society*, vol. 20, p. 335 (2007)
25. <https://github.com/MarcoLandtHayen/fact-or-artifact>

B. Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

Supplementary Material

LRP for ESNs

In Section 4 we give a heuristic on how to choose the leak rate α from a practical point of view. Feeding a sample with T time steps into an ESN model, the aim is to equally take all time steps' inputs into account, if possible. Here, we add a mathematical validation to support the empirical approach and explain side effects related to the choice of α . As described in Section 3, the general concept of LRP needs to be customized for ESN models. In particular, we work with leaky reservoirs. Reservoir state transition for the first time step $t = 1$ and subsequent time steps $t = 2..T$ is outlined in Equations 1 and 2, respectively. The leak rate α determines the reservoir dynamics. To compute reservoir states $x(t)$ for a given time step in the forward pass, a fraction $(1 - \alpha)$ of the previous reservoir states $x(t - 1)$ is kept. The current time step's inputs $u(t)$ are included in some activation term $\text{act}[W_{in}u(t) + b_{in} + W_{res}x(t - 1) + b_{res}]$. This term is multiplied with α . The larger the leak rate, the faster the reservoir reacts to new inputs.

For obtaining mean relevance maps, we unfold reservoir dynamics in time. Each time step is treated as an individual layer. We use the backward pass and start with model prediction Y as final or total relevance. A portion of the total relevance is attributed to every time step's inputs. Only the remaining relevance is then traced back through the layers or time steps. Multiplication with leak rate α in reservoir state transition leads to an exponential decay, when tracing relevance backwards. $R(t)$ denotes residual relevance for time step t . Once we reach the initial time step $t = 1$, all residual relevance is attributed to the initial inputs $u(t = 1)$.

$$R(t) = \exp(-\alpha(T - t)) \quad (6)$$

Equation 6 states the the exponential decay for a given leak rate α . The left part of Figure 7 shows the result for different leak rates. We clearly see the fading memory effect that has been described in Section 4. For leak rates $\alpha = 0.1$ and 0.2 , only the inputs for time steps $t \geq 50$ are considered.

Table 2 shows residual relevance $R(t = 1)$ approximated as exponential decay according to Equation 6 for various leak rates. Here, we assume to have $T = 100$ time steps. In addition to that, Table 2 shows empirical residual relevance $\hat{R}(t = 1)$ obtained from an ESN model trained on synthetic samples feeding samples column-wise into the model.

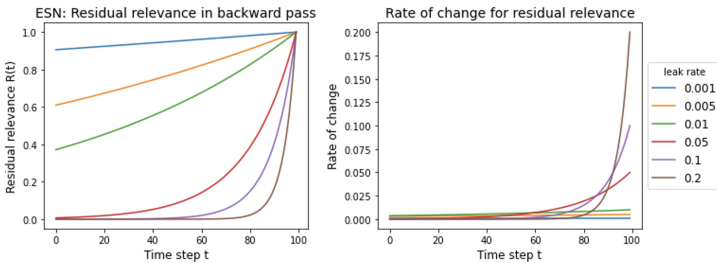


Fig. 7. Residual relevance approximated by an exponential decay over $T = 100$ time steps for various leak rates α (left-hand side) and rate of change for residual relevance (right-hand side).

Table 2. Residual relevance $R(t = 1)$ from approximation compared to empirical residual relevance $\hat{R}(t = 1)$ for first time step for ESN model trained on synthetic samples with different leak rates α .

Leak rate α	$R(t = 1)$	$\hat{R}(t = 1)$
0.001	90.5%	90.7%
0.005	60.7%	63.1%
0.01	36.8%	41.9%
0.05	0.7%	2.9%
0.1	0.0%	0.1%
0.2	0.0%	0.0%

Synthetic samples also consist of $T = 100$ time steps. For leak rates $\alpha = 0.001, 0.005$ and 0.01 , we find $\hat{R}(t = 1) = 91\%$, 63% and 42% of total relevance, respectively, as residual relevance. This amount is attributed to initial inputs $u(t = 1)$, which gives the initial inputs an unreasonable high relevance compared to all other time steps' inputs. For leak rates $\alpha = 0.05, 0.1$ and 0.2 , we find the remaining relevance attributed to the initial inputs to be close to zero. The exponential decay is found to be a good approximation.

Using ESN models, the aim is to equally consider all time steps' inputs, since we don't know in advance, which time steps contain most relevant information. Therefore, we need to look at the rate of change for residual relevance traced backwards through all time steps. The rate of change can be approximated as first derivative $R'(t)$ of residual relevance $R(t)$ and is stated in Equation 7.

$$R'(t) = \frac{\partial R}{\partial t} = \alpha \exp(-\alpha(T - t)) \quad (7)$$

The right-hand side of Figure 7 shows the resulting rate of change for different leak rates. We find the rate of change to be almost constant for small leak rates $\alpha = 0.001, 0.005$ and 0.01 , reflecting a similar amount of total relevance attributed to each time step's inputs $u(t)$, as desired. This is violated for higher leak rates $\alpha = 0.05, 0.1$ and 0.2 , respectively. For these leak rates the ESN model puts significantly higher weights on later input steps.

Taking residual relevance and rate of change into account, we need to find a compromise. We want all time steps to be about equally considered. This encourages us to choose small leak rates. However, we need to avoid an unreasonable high amount of residual relevance to be put on the initial time step's inputs $u(t = 1)$. One possible solution is to add a dummy column of constant value one as first input time step for all input samples. Residual relevance is then absorbed by this dummy column, which doesn't affect classification in general and doesn't distort mean relevance maps, since it is identical for all samples of both classes. For showing the obtained mean relevance maps, the dummy column is then omitted.

Data Preprocessing

While the main body of this work focuses on models, methods and results, we provide details on data preprocessing in this section. Different ANN models require input samples to be customized in dimensionality and we need to deal with missing data at grid points located on land. For our MLP models, we only consider valid grid points of raw two-dimensional input samples. These grid points are then transformed into a one-dimensional vector as sketched in Figure 8. The number of input units then equals the number of valid grid points.

B. Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

Our CNN models require two-dimensional input samples. Missing values are replaced by zero. Input samples are then scanned by quadratic kernels. Kernel sizes and step sizes, also referred to as strides, need to be specified. Figure 9 sketches on the left-hand side an exemplary sample of dimensions 5×6 grid points. Here, we assume to work with a kernel of size 3×3 and step size is also chosen to be 3 in both directions, x and y. This requires adding a row of zeros to allow the sample to be completely scanned by our specified kernel and step size. The modified input sample is shown on the right-hand side of Figure 9. Adding a row of zeros doesn't affect mean relevance maps. Zero relevance is attributed to these additional grid points when we only consider positive pre-activations z^+ in the propagation rule, as stated in Equation 5.

For our ESN models, we need to add a dummy column of ones as first input step to absorb residual relevance in any case as described above. The way we deal with missing data varies for different techniques of feeding samples into ESN models. For feeding samples column- or row-wise, missing values are set to zero, as for our CNN models. Samples are then sliced column- or row-wise, as sketched in the upper and lower part of Figure

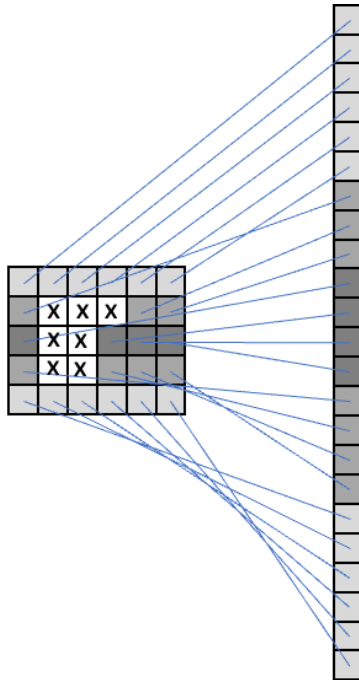


Fig. 8. Transformation of an exemplary two-dimensional input sample consisting of 5 rows and 6 columns into a one-dimensional vector for our MLP models. Consider only valid grid points. Invalid grid points are marked as X.

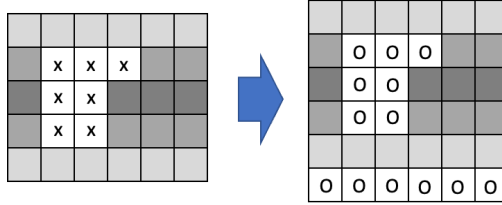


Fig. 9. Modification of an exemplary two-dimensional input sample consisting of 5 rows and 6 columns by replacing missing values and adding an additional row of zeros for our CNN models. Invalid grid points and zeros are marked as X and O, respectively.

10, respectively. When feeding samples as equal-sized pieces, we only consider valid grid points. These grid points are then transformed into a one-dimensional vector, randomly permuted and split. The size or the number of grid points per piece needs to be specified. The number of valid grid points needs to be dividable by the piece size. Therefore, we optionally add zeros on the trailing edge to allow splitting into equal-sized pieces. Again, this doesn't affect mean relevance maps, since zero relevance is attributed to these grid points. For showing resulting mean relevance maps, additional grid points are discarded. Preprocessing is demonstrated in Figure 11 for an exemplary sample consisting of 5 rows and 6 columns. Desired piece size is 5 in this example.

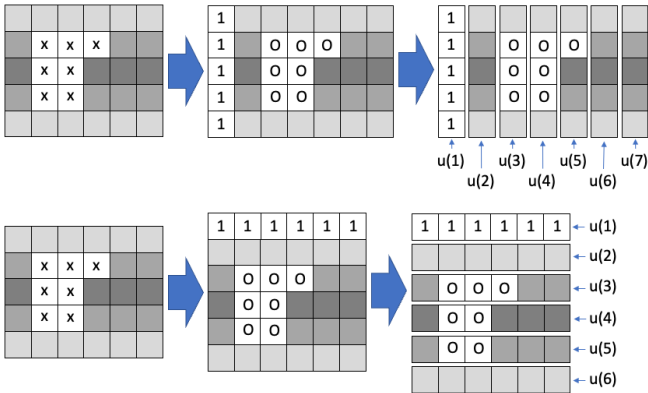


Fig. 10. Preprocessing of an exemplary two-dimensional input sample for feeding samples column-wise (upper part) or row-wise (lower part) into our ESN models requires adding an additional column or row of ones, respectively, as first time step $u(1)$ and replacing missing values by zero in any case. Invalid grid points, zeros and ones are marked as X, O and 1, respectively.

B. Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures

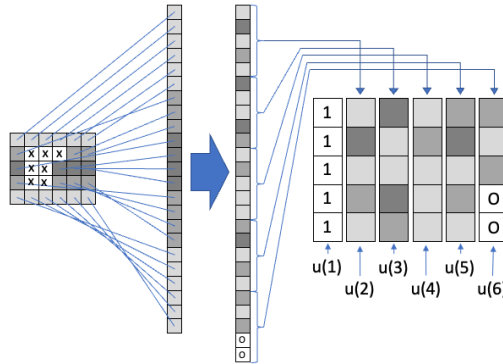


Fig. 11. Preprocessing of an exemplary two-dimensional input sample for feeding samples as equal-sized pieces of size 5 into our ESN models. Only valid grid points are transformed into a one-dimensional vector and randomly permuted. Need to add two zeros at the trailing edge before splitting. First time step's inputs consist of ones. Invalid grid points, zeros and ones are marked as X, O and 1, respectively.

Authors

Marco Landt-Hayen Doctoral researcher at Helmholtz School for Marine Data Science (MarDATA) in Kiel, Germany, with a project on attribution of climate events and focus on exploring methods of explainable AI.

Dr. Willi Rath Data Science technician at GEOMAR Helmholtz Centre for Ocean Research in Kiel, Germany.

Prof. Dr. Martin Claus Junior Professor for Ocean Dynamics at GEOMAR Helmholtz Centre for Ocean Research in Kiel, Germany and Christian-Albrechts-Universität zu Kiel, Germany.

Prof. Dr. Peer Kröger Head of Information Systems and Data Mining group at Christian-Albrechts-Universität zu Kiel, Germany.

A Climate Index Collection based on Model Data

**Marco Landt-Hayen, Willi Rath, Sebastian Wahl, Nils Niebaum,
Martin Claus, Peer Kröger**

Published in Environmental Data Science, special edition on "Tackling Climate Change with Machine Learning", in collaboration with "Climate Change AI" workshop at NeurIPS 2022



C. A Climate Index Collection based on Model Data

Environmental Data Science (2023), 2: e9, 1–13
doi:10.1017/eds.2023.5



DATA PAPER  

A climate index collection based on model data

Marco Landt-Hayen^{1,2} , Willi Rath², Sebastian Wahl², Nils Niebaum² , Martin Claus^{2,3} and Peer Kröger¹

¹Information Systems and Data Mining, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

²Ocean Circulation and Climate Dynamics, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

³Faculty of Mathematics and Natural Sciences, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

Corresponding author: Marco Landt-Hayen; Email: mlandt-hayen@geomar.de

Received: 11 January 2023; **Revised:** 30 March 2023; **Accepted:** 05 April 2023

Keywords: Climate events; data mining; deep learning; machine learning; time series forecasting

Abstract



Machine learning (ML) and in particular deep learning (DL) methods push state-of-the-art solutions for many hard problems, for example, image classification, speech recognition, or time series forecasting. In the domain of climate science, ML and DL are known to be effective for identifying causally linked modes of climate variability as key to understand the climate system and to improve the predictive skills of forecast systems. To attribute climate events in a data-driven way, we need sufficient training data, which is often limited for real-world measurements. The data science community provides standard data sets for many applications. As a new data set, we introduce a consistent and comprehensive collection of climate indices typically used to describe Earth System dynamics. Therefore, we use 1000-year control simulations from Earth System Models. The data set is provided as an open-source framework that can be extended and customized to individual needs. It allows users to develop new ML methodologies and to compare results to existing methods and models as benchmark. For example, we use the data set to predict rainfall in the African Sahel region and El Niño Southern Oscillation with various ML models. Our aim is to build a bridge between the data science community and researchers and practitioners from the domain of climate science to jointly improve our understanding of the climate system.

Impact Statement

Machine learning (ML) models learn from data. To compare and improve ML methods and models, data scientists need standard data sets as benchmark. There exist many standard data sets, like a collection of handwritten digits or images. Our contribution adds a consistent and comprehensive collection of climate indices as new benchmark data set. This collection can be used to train ML models to understand the complex short-term and long-term variability of the climate system and to predict climate events.

1. Introduction

To develop and compare machine learning (ML) methods and models in an objective way, there exist standard data sets as benchmark. Among these data sets, we find, for example, a collection of handwritten digits provided by the National Institute of Standards and Technology, referred to as MNIST data set

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

<https://doi.org/10.1017/eds.2023.5> Published online by Cambridge University Press

(Lecun et al., 1998). Other data sets contain images, like the CIFAR-10 data set from Canadian Institute for Advanced Research, introduced by Krizhevsky (2009), or CelebSET, as a collection of photos of celebrities with different ethnicity (Raji et al., 2020). These data sets are mostly suitable for classification algorithms. Famous data sets for pattern recognition and clustering applications are, for example, Palmer Penguins or the Wine data set, provided by Horst et al. (2020) and the UCI Machine Learning Repository (Murphy and Aha, 1994), containing attributes for various species of penguins and results of a chemical analysis of wines, respectively. Furthermore, the data science community also provides standard time series collections, for example, the Rainforest Automation Energy data set that contains energy consumption time series for various household appliances (Makonin et al., 2018). However, benchmark data sets in the field of climate science are rare. To name a few, Mamalakis et al. (2022) provide a framework to create synthetic data sets designed for problems in geosciences. And Watson-Parris et al. (2022) introduced ClimateBench, as a benchmark for data-driven climate projections.

Here, we are interested in describing the underlying dynamics of the Earth System. Real-world data in this context are limited to observable features that can be measured in a comprehensive way or that can be reconstructed from sparse measurements. Examples are sea surface temperature (SST), sea level pressure (SLP), surface air temperature (SAT), sea surface salinity (SSS), geopotential height at various pressure levels, for example, at 500 millibar (Z500), or total precipitation (PREC). These variables reflect some of the main dynamics of the Earth system in form of known modes of climate variability, patterns, and oscillations, for example, Atlantic Multidecadal Oscillation (AMO) (Schlesinger and Ramankutty, 1994), the Southern Annular Mode (SAM) (Gong and Wang, 1999), or the El Niño Southern Oscillation (ENSO) (Philander, 1989). To describe the Earth System dynamics in a compressed way, multi-dimensional physical fields can be reduced to specific climate indices that capture the main processes. For instance, ENSO is a complex phenomenon that can be detected as periodic SST fluctuations in the Tropical Pacific. Several indices are defined to compute the current ENSO phase from area-averaged SST anomalies (SSTA) in certain regions. For instance, the Niño 3.4 index defined from the Niño 3.4 region (5°N–5°S, 120°W–170°W) by the National Oceanic and Atmospheric Administration (NOAA) is often used in the context of ENSO (Climate Diagnostics Bulletin, n.d.). While ENSO is a large-scale driver of the climate system, other indices aim to capture regional variability in specific features, like the Sahel precipitation index (SPI) (Badr et al., 2014). This index measures anomalies of summer rainfall in the African Sahel region (10°N–20°N, 20°W–10°E). Real-world climate data are, for example, provided by the Joint Institute for the Study of the Atmosphere and Ocean (n.d.) or National Oceanic and Atmospheric Administration (NOAA) (n.d.). However, climate indices are limited in their temporal extent, since consistent real-world measurements started only in recent history or measurements are subject of specific research projects that run over a certain period in time.

Our aim is to better understand existing modes of climate variability and to find new relationships. Therefore, we require a consistent and comprehensive collection of climate indices over a sufficiently long time span, which favors the use of model data over real-world data. Earth System Models (ESMs) aim to simulate processes of the Earth system in specified temporal and spatial resolution. The Flexible Ocean and Climate Infrastructure (FOCI) (Matthes et al., 2020) and the Whole Atmosphere Community Climate Model (WACCM) as extension of the Community Earth System Model (CESM) (Hurrell et al., 2013; Marsh et al., 2013) are both coupled, global climate models that provide state-of-the-art computer simulations of the past, present and future states of the Earth system. The quality of model outputs is evaluated on certain control runs. This can, for example, be done by starting an ESM with pre-industrial conditions from the year 1850 and letting the model unfold its dynamics without external forcing over a desired time span. Here, we use the output of FOCI and CESM control runs. In particular, we work with SST, SAT, SLP, Z500, SSS, and PREC as two-dimensional fields. From these variables, we derive a set of climate indices over 1000 and 999 years, respectively. The obtained collection of climate indices based on model data (CICMoD) serves as a reduced description of the Earth system in a consistent and comprehensive way. Our main contributions are as follows:

C. A Climate Index Collection based on Model Data

- We introduce CICMoD as a new benchmark data set to the data science community, describing the climate system.
- CICMoD allows the user to develop new ML methods and to compare results to existing methods and models in an objective way.
- We provide an open-source framework that can be extended and customized to individual needs, for example, by including further ESMs.
- Additionally, we briefly sketch two examples of how our CICMoD data set can be used.

Relationships in the climate systems are often characterized as nonlinear and nonstationary (Pak et al., 2014, 2018; Zhang et al., 2019). ML and deep learning (DL) models have been shown to be useful for this kind of problems, for example, by Mayer and Barnes (2021) or Pegion et al. (2022). However, working with benchmark data sets bears the risk of having undetected errors and biases or of being unrepresentative. For instance, Liao et al. (2021) and Luccioni and Rolnick (2022) argue that the ubiquity of benchmarks in computer science has led to efforts that chase benchmark performance at the expense of real-world applications. Thus, once we find new relationships in the Earth System, these findings need to be confirmed on real-world data to identify artifacts in ESM simulations. A better understanding of causally linked modes within our climate system is essential to tackle climate change and to attenuate its impacts.

The rest of this work is structured as follows. In Section 2, we provide a short description of FOCI and CESM. An overview of all indices included in the CICMoD data set and details on how the indices are derived from raw ESM outputs are given in Section 3. In Section 4, we show two exemplary applications and use climate indices from CICMoD data set to predict Sahel rainfall and ENSO, respectively. A detailed discussion of all results and a conclusion is found in Section 5.

2. Model Data

The climate indices included in our CICMoD data set are derived from monthly averaged output of climate model simulations with FOCI and CESM, respectively. The CESM simulation is based on version 1.0.6 with WACCM version 4 (Drews et al., 2022). The FOCI simulation used in this manuscript is the control simulation referred to as “FOCI-piCtI” based on FOCI version 1.3.0 (Matthes et al., 2020). Both simulations were run using pre-industrial external forcing that is representative for the year 1850. The FOCI pre-industrial control simulation has been initialized from an ocean at rest with a salinity and temperature distribution based on observations approximately from the last 30 years and then ran for 1500 years. Here, we only use the latter 1000 years and skip the first 500 years to allow the model to find its equilibrium. The CESM control simulation has been initialized from another multi-centennial pre-industrial control run provided by the core development team of the National Center for Atmospheric Research (NCAR) (National Center for Atmospheric Research, n.d.) and is therefore already in equilibrium. FOCI ($1.8^\circ \times 1.8^\circ$, 95 vertical levels) and CESM ($1.8^\circ \times 2.5^\circ$, 106 vertical levels) were run at similar horizontal and vertical resolution, although the vertical distribution of the model layers differs significantly between FOCI and CESM. Both models have been extensively evaluated and used in various climate studies. FOCI and CESM are based on very different component models (see Hurrell et al., 2013; Marsh et al., 2013; Matthes et al., 2020, for details) with different strengths and weaknesses in simulating various aspects of the global climate.

From both control simulations’ output we use Z500, SLP, SST, SSS, SAT, and PREC. All features except SSS are provided on a two-dimensional atmospheric latitude–longitude grid. As SSS was originally stored on the curvilinear ocean grids, it was sampled to the grid of the other atmospheric fields of the respective model by aggregation with xhistogram (Abernathey et al., 2022). Note, that SAT refers to the temperature in 2 m height for both, CESM and FOCI data.

3. Climate Index Collection

In this section, we give an overview of all indices included in the CICMoD data set and reveal details on how the indices are derived. In total, our CICMoD data set consists of 29 climate indices. The indices can

be grouped by the underlying feature. Each feature is discussed separately in the following subsections. We conclude this section with remarks on statistics and pairwise correlation of all indices.

3.1. Geopotential height

Geopotential height is a vertical coordinate with reference to Earth's mean sea level. Its contours are used to calculate the geostrophic wind which is of interest for climate dynamics. Here, we choose geopotential height at constant pressure of 500 millibar, referred to as Z500. According to NOAA, Z500 relates to winds in the range between 5000 and 6000 meters above mean sea level (National Oceanic and Atmospheric Administration's National Weather Service, n.d.). We use Z500 to compute the SAM index which relates to the principal mode of variability in the Southern Hemisphere (SH) extratropics. The SAM index can be obtained as the Principle Component (PC) time series of the leading Empirical Orthogonal Function (EOF) of monthly geopotential height anomalies over parts of the SH (20°S–90°S) (Thompson and Wallace, 2000). SAM has large impact on climate dynamics of the SH, including Australian rainfall and Antarctic surface temperatures (Marshall, 2007).

3.2. Sea level pressure

SLP refers to the air pressure at sea level. Several indices are derived from SLP and its anomalies. Opposed to the PC-based version described in Section 3.1, the SAM index was originally defined by Gong and Wang (1999) as the difference of normalized monthly zonal mean SLP at 40°S and 65°S, respectively. Both versions are included in our CICMoD data set.

The Southern Oscillation Index (SOI) is defined as normalized SLP differences between Tahiti (17°41'S, 149°27'W) and Darwin, Australia (12°27'S, 130°50'E) (Walker and Bliss, 1932). It is used as a measure of the large-scale fluctuations in the air pressure between the Western and Eastern Tropical Pacific and is closely related to ENSO. Similar to SOI, the North Atlantic Oscillation (NAO) index can be computed from SLP as the normalized difference between Reykjavik (64°9'N, 21°56'W) and Ponta Delgada (37°45'N, 25°40'W) (Hurrell, 1995). Additionally, the NAO index can be obtained as the PC time series of the leading EOF of monthly SLP anomalies over the Atlantic sector (20°N–80°N, 90°W–40°E) (National Center for Atmospheric Research, n.d.). The NAO refers to swings in the atmospheric SLP between the Arctic and the subtropical Atlantic that are associated with changes in the mean wind speed and direction. Such changes are reflected in the seasonal mean heat and moisture transport between the Atlantic and the neighboring continents and have an impact on the intensity and number of storms, their paths, and their weather (Hurrell et al., 2003).

The North Pacific (NP) index measures interannual to decadal variations in the atmospheric circulation. It is derived from area-weighted SLP anomalies in a box bordered by 30°N to 65°N and 160°E to 140°W (Trenberth and Hurrell, 1994). Each grid point's SLP anomaly value represents the mean value over the corresponding grid box. Since the area of the grid boxes depends on the latitude, we need to use area-weighted SLP anomalies to avoid overestimating values in high latitudes. Usually, the index focuses on anomalies during November and March. Here we keep full information and provide monthly anomalies for all months of a year.

3.3. Sea surface temperature

SST is the ocean temperature close to the surface. By removing the seasonal cycle, we obtain SST anomalies (SSTA). In particular, we subtract the mean over time separately for each month. SSTA impact the energy transfer at the interface between ocean and atmosphere and are of high interest for describing processes in the climate system. Several modes of variability are known to exist on different time scales in the range of years, decades, or even longer. AMO refers to a natural variability occurring in the SST of the North Atlantic with a multidecadal period of 60–80 years. AMO is computed from area-weighted SSTA of the North Atlantic (Trenberth and Shea, 2006).

The Pacific Decadal Oscillation (PDO) index is obtained as the PC time series of the leading EOF of monthly SSTA in the North Pacific basin (20°N–60°N, 120°E–260°E). PDO resembles ENSO in its

C. A Climate Index Collection based on Model Data

spatial pattern. However, ENSO is referred to as an interannual phenomenon while PDO is decadal in scale (Newman et al., 2016).

ENSO is characterized by periodic fluctuations in SST in the Tropical Pacific. Its positive and negative phases relate to unusual warm (El Niño) or cold (La Niña) SST, respectively. Tropical Pacific is divided into specific regions, so-called Niño regions. ENSO indices are then derived from SSTA in the corresponding region by spatial averaging. Indices are divided by the standard deviation of area-weighted SST over time in the same region, as normalization. Here, we include Niño 1 + 2 region as the smallest and eastern-most Niño region where the phenomenon was first recognized by the local coastal population. Additionally, we present ENSO indices on Niño 3, 3.4, and 4 regions, respectively (National Oceanic and Atmospheric Administration, n.d.). ENSO is a large-scale driver of the climate system (Philander, 1989). Usually, ENSO indices are smoothed by taking the rolling mean over several months to erase noise. Here, we omit the rolling mean and provide pure SSTA indices instead, to preserve full information.

Other regions of interest in the context of climate dynamics related to SSTA are Tropical North Atlantic, Tropical South Atlantic, Eastern Subtropical Indian Ocean, Western Subtropical Indian Ocean, Mediterranean Sea, and hurricane main development region, respectively. For instance, the African summer monsoon is found to be highly sensitive to SST variability in all tropical basins (Giannini et al., 2003). Corresponding indices are included in our CICMoD data set.

3.4. Sea surface salinity

SSS measures the amount of salt dissolved in the ocean surface water and plays an important role in ocean circulation processes. Furthermore, rainfall on land is largely supplied by evaporation over the ocean and that evaporation leaves an imprint in SSS. Here, we include several indices derived from SSS anomalies (SSSA) in specific regions of the Atlantic Ocean introduced by Li et al. (2016).

3.5. Surface air temperature

SAT is the air temperature close to the surface and relates to the ability of evaporation, since warmer air has a higher storage capacity for water vapor. Like this, SAT anomalies (SATA) influence the energy transfer at the interface between Earth's surface and atmosphere. Here, we track area-averaged monthly SATA on large scales with indices covering complete NH and SH (Jones et al., 1999). Additionally, we split NH and SH into land-only and ocean-only regions, respectively, and include corresponding indices in our CICMoD data set. The ocean masks are taken from the native model grids.

3.6. Precipitation

Precipitation has a high impact on society in form of extreme events like flooding caused by heavy rainfall or droughts due to missing or lower as normal rainfall. As an example, we include the SPI as measure for rainfall in the African Sahel region (Badr et al., 2014). The rainy season in this area is centered on June through October (Joint Institute for the Study of the Atmosphere and Ocean, n.d.). In its original form, the SPI gives a measure of the year to year variability of Sahel rainfall as mean over the rainy season. Moreover, we provide the SPI as monthly anomalies of rainfall in the Sahel zone (10°N–20°N, 20°W–10°E) to preserve full information.

3.7. CICMoD data set

Table 1 gives an overview of all 29 indices included in CICMoD data set ordered by the underlying feature. By definition, all indices have zero mean over time, whereas only NAO_PC, PDO_PC, SAM_PC, SAM_ZM, and SOI are normalized by design to have unit variance. If required, normalization of the remaining indices can be done in pre-processing. Exemplary, pairwise correlation coefficients for all indices included in CICMoD data set derived from FOCI data are shown in Figure 1. Indices derived from CESM data show similar characteristics. ENSO indices are found to be highly correlated, as expected,

Table 1. All indices are included in CICMoD data set with their acronyms and spatial domains, ordered by the underlying feature.

	Index	Acronym	Spatial domain	
			Lat in °N	Lon in °E
Z500	Southern Annular Mode (PC-based)	SAM_PC	−90 to −20	
SLP	Southern Annular Mode (zonal mean)	SAM_ZM	−65 to −40	
	Southern Oscillation	SOI	Tahiti	Darwin
			(−18°N, 211°E)	(−12°N, 131°E)
	North Atlantic Oscillation (station)	NAO_ST	Reykjavik	Ponta Delgada
			(64°N, 338°E)	(38°N, 334°E)
	North Atlantic Oscillation (PC-based)	NAO_PC	20 to 80	−90 to 40
	North Pacific Pattern	NP	30 to 65	−160 to 220
SST	Atlantic Multidecadal Oscillation	AMO	0 to 70	Atlantic basin
	Pacific Decadal Oscillation	PDO_PC	20 to 60	120 to 260
	El Niño Southern Oscillation (1 + 2)	ENSO_12	−10 to 0	270 to 280
	El Niño Southern Oscillation (3)	ENSO_3	−5 to 5	210 to 270
	El Niño Southern Oscillation (3,4)	ENSO_34	−5 to 5	190 to 240
	El Niño Southern Oscillation (4)	ENSO_4	−5 to 5	160 to 210
	Tropical North Atlantic SSTA	SST_TNA	5 to 25	−55 to −15
	Tropical South Atlantic SSTA	SST_TSA	−20 to 0	−30 to 10
	Eastern Subtrop. Indian Ocean SSTA	SST_ESIO	−28 to −18	90 to 100
	Western Subtrop. Indian Ocean SSTA	SST_WSIO	−37 to −27	55 to 65
	Mediterranean Sea SSTA	SST_MED	30 to 45	0 to 25
Hurricane main dev. region SSTA	SST_HMDR	10 to 20	−85 to −20	
SSS	North Atlantic SSSA	SSS_NA	25 to 50	−50 to −15
	Western North Atlantic SSSA	SSS_WNA	25 to 38	−50 to −40
	Eastern North Atlantic SSSA	SSS_ENA	25 to 50	−40 to −15
	South Atlantic SSSA	SSS_SA	−22.5 to −10	−42 to −10
SAT	Northern Hemisphere SATA	SAT_N_ALL	0 to 90	
	Northern Hemisphere SATA (ocean)	SAT_N_OCEAN	0 to 90	Ocean
	Northern Hemisphere SATA (land)	SAT_N_LAND	0 to 90	Land
	Southern Hemisphere SATA	SAT_S_ALL	−90 to 0	
	Southern Hemisphere SATA (ocean)	SAT_S_OCEAN	−90 to 0	Ocean
	Southern Hemisphere SATA (land)	SAT_S_LAND	−90 to 0	Land
PREC	Sahel Precipitation	PREC_SAHEL	10 to 20	−20 to 10

C. A Climate Index Collection based on Model Data

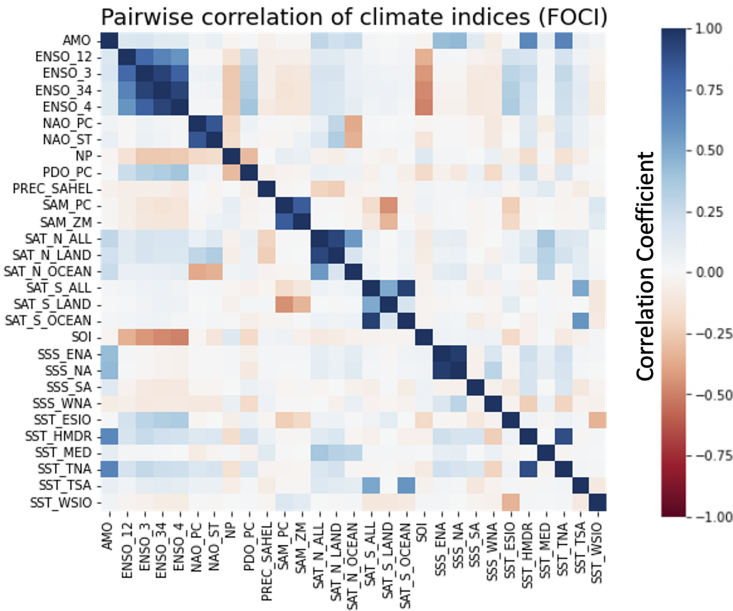


Figure 1. Pairwise correlation coefficients of all CICMoD indices derived from FOCI data.

since these indices are all computed from SSTA in some narrow region of the Tropical Pacific. Additionally, we find station- and PC-based NAO indices to be highly correlated, as well as PC-based SAM and SAM from zonal mean, as these indices are designed to describe the same processes. Indices regarding SATA in the NH and SH, respectively, and indices regarding SSS anomalies in the North Atlantic also reveal similarities in terms of high correlation, since by design spatially related features are involved in the computation. Besides that, SOI is found to be negatively correlated to all ENSO indices. Periods of negative (positive) SOI values coincide with warmer (colder) than normal ocean water across the Eastern Tropical Pacific, which is typical for El Niño (La Niña) episodes (Power and Kociuba, 2011).

4. Application and Results

In this section, we briefly sketch two applications of our CICMoD data set to predict Sahel rainfall and ENSO, respectively.

4.1. Sahel rainfall

Sahel summer precipitation has been observed to be highly variable with floods and droughts occurring on a regular basis and has a high impact on living conditions in the region. Predicting Sahel rainfall and understanding the underlying processes is essential, since it allows taking measures in advance to avoid damage and prevent hunger crises. As a first application, we use ML models on our CICMoD data set to

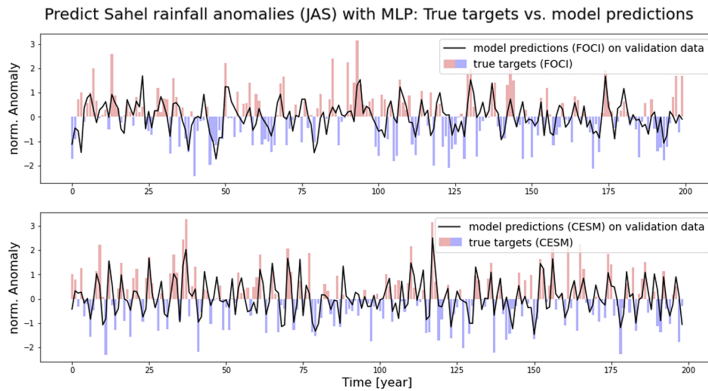


Figure 2. Fidelity check on validation data: Sahel rainfall predictions (black line) from MLP models on FOCI data (upper part) and CESM data (lower part), respectively, compared to true targets shown as bar plot.

predict rainfall in the Sahel region. In particular, we apply a linear regression model as a baseline and, additionally, train a multilayer perceptron (MLP). Following the approach of Badr et al. (2014), we use April to June mean index values for all indices included in our CICMoD data set except SPI as predictors to infer July to September seasonal sum of SPI as target. The input layer of the MLP thus consists of 28 input units. Additionally, we have two hidden layers with 20 and 10 units, respectively, and a single output unit. We use a linear activation function and train the MLP over 10 epochs with a batch size of 10, set the learning rate to 0.0005 and use the Adam optimizer (Kingma and Ba, 2014). For FOCI and CESM data, the first 800 years are used as training data, while the remaining 200 and 199 years, respectively, are used for validation. Figure 2 shows results from MLP models. Results from linear regression are similar and therefore not shown, here. To evaluate model performance, we look at mean squared error (MSE) of predictions compared to true targets used as objective or loss function. Additionally, the correlation of predicted values and true targets is computed as metric. Corresponding MSE and correlation for linear regression and MLP models on FOCI and CESM data, respectively, are shown in Table 2.

4.2. El Niño Southern Oscillation

ENSO is the predominant variation of winds and SST in the Tropical Pacific. The positive phase (El Niño) is characterized by unusual warm SST and high SLP in the Eastern Tropical Pacific, whereas the negative phase (La Niña) relates to unusual cold SST and low SLP in the same region and above-average SST in the Western Tropical Pacific. Both events last several months and occur with a period of 2–7 years with varying intensity per period. ENSO tremendously affects those countries bordering the Pacific Ocean. Strong El Niños, for example, correspond to warm weather conditions with heavy rainfalls from April through October causing major flooding along the West coast of South America near Ecuador and the Northern part of Peru (Cai et al., 2020). Consequences of La Niña are, for example, heavy rainfalls over Malaysia, the Philippines, and Indonesia. Therefore, knowing the ENSO phase several months in advance is of high interest for society since it allows to take measures to avoid

C. A Climate Index Collection based on Model Data

Table 2. Evaluating model performance for predicting Sahel rainfall with linear regression (lin. reg.) and MLP models trained on FOCI and CESM data, respectively.

	FOCI		CESM	
	Lin. reg.	MLP	Lin. reg.	MLP
MSE _{train}	0.86	0.88	0.49	0.51
MSE _{val}	0.83	0.78	0.62	0.60
Correl _{train}	0.55	0.53	0.70	0.69
Correl _{val}	0.50	0.52	0.68	0.69

Note. The MSE and correlation (Correl) of predicted values and true targets are shown separately for training and validation data.

damage and to protect people. As second example, we use different ANN models on our CICMoD data set to predict ENSO with various lead times. In particular, we train convolutional neural networks (CNNs) and long short-term memory (LSTM) models to predict current ENSO phase and ENSO phase 3 and 6 months into the future, respectively. Here, targets are derived from ENSO_34 time series included in CICMoD, which reflects Niño 3.4 index. As input features, we use all remaining indices from our CICMoD data set, excluding other Niño indices due to the high correlation to our targets. Input features are split into sequences of 24 months. We thus try to predict current and future ENSO phases from past 2 years' conditions.

In particular, the CNN models are based on two one-dimensional convolutions with 10 and 20 filters, respectively. Kernel size is set to 5 with a stride of 1. Each convolution is followed by batch normalization, a leaky rectified linear unit activation with negative slope coefficient $\alpha = 0.3$, and a maximum pooling operation with a pool size of 2. The output of the final pooling operation is then flattened and used as input for two fully connected layers with 20 and 10 units, respectively, and finally, we have a single output unit. The LSTM models are based on two LSTM layers with 10 and 20 units, respectively. The output of the final LSTM layer is used as input for two fully connected layers with 20 and 10 units, respectively, and finally, we have a single output unit, similar to the CNN models. Furthermore, we use a linear activation function for all fully connected layers and the output unit and train the models over 20 epochs with a batch size of 20, set the learning rate to 0.0001 and use the Adam optimizer.

Figure 3 shows pairwise correlation of targets and input features used in this experiment. Again, the first 800 years of FOCI and CESM data are used as training data, while the remaining 200 and 199 years, respectively, are used for validation. Figure 4 shows results from CNN models on the first 500 months of FOCI and CESM validation data, respectively, for various lead times. To evaluate model performance, we again look at MSE and correlation of predictions compared to true targets, as shown in Table 3.



Figure 3. Pairwise correlation coefficients of Niño 3.4 index with various lead times (current phase, 3 and 6 months into the future) used as targets and input features, both derived from FOCI data.

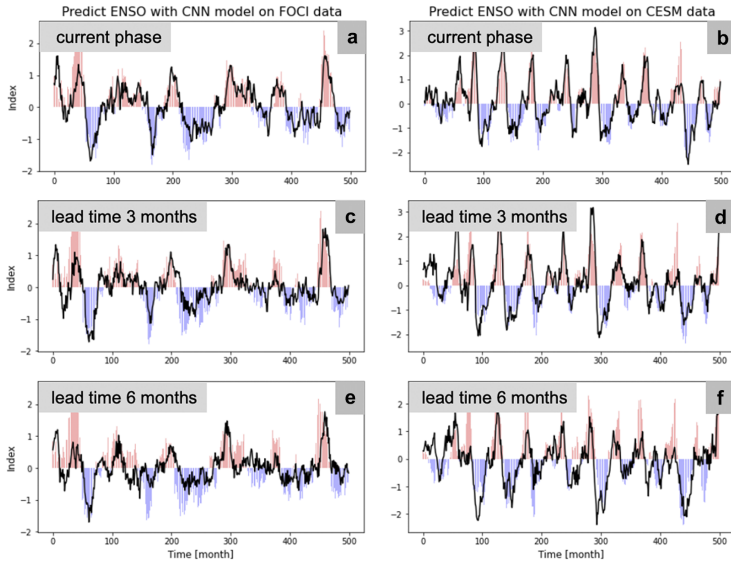


Figure 4. Fidelity check on the first 500 months of validation data: Compare predictions (black line) from CNN models on FOCI (left-hand side) and CESM data (right-hand side), respectively, compared to true targets shown as bar plot for various lead times. (a,b) Current phase. (c,d) Three months into the future. (e,f) Six months into the future.

Table 3. Evaluating model performance for predicting ENSO with CNN and LSTM models trained on FOCI and CESM data, respectively.

		CNN			LSTM		
		ENSO_34	Lead 3	Lead 6	ENSO_34	Lead 3	Lead 6
FOCI	MSE _{train}	0.16	0.21	0.27	0.16	0.23	0.28
	MSE _{val}	0.24	0.37	0.47	0.26	0.41	0.53
	Correl _{train}	0.84	0.79	0.72	0.84	0.77	0.71
	Correl _{val}	0.81	0.68	0.56	0.79	0.64	0.47
CESM	MSE _{train}	0.20	0.31	0.42	0.20	0.32	0.41
	MSE _{val}	0.29	0.48	0.65	0.30	0.50	0.73
	Correl _{train}	0.88	0.80	0.72	0.88	0.79	0.72
	Correl _{val}	0.84	0.72	0.59	0.83	0.70	0.56

Note. The MSE and correlation (Correl) of predicted values and true targets are shown separately for training and validation data. ENSO phases at 3 and 6 months into the future are denoted as lead 3 and lead 6, respectively.

Abbreviations: CESM, community earth system model; FOCI, flexible ocean and climate infrastructure.

C. A Climate Index Collection based on Model Data

5. Discussion and Conclusion

In this work, we introduce a consistent and comprehensive collection of climate indices as a new benchmark data set. The collection is consistent in a sense that we use the output of ESM control runs to derive all indices. For FOCI and CESM control runs, we have 1000 and 999 years of monthly data, respectively, as an advantage compared to real-world data, since ML models require sufficient training data. The collection is comprehensive as we include a broad selection of known patterns, oscillations, and variability of the Earth system. The index collection is not complete since we focus on processes within the atmosphere, in the upper ocean and at the interface of ocean and atmosphere. However, our CICMoD data set serves as basis. Additionally, we provide an open-source framework that can be extended and customized to individual needs including the application to further ESMs. This opens the door for collaboration in many ways. Our new data set allows researchers from the data science community to adapt existing ML models and develop new ML methods to tackle problems from the domain of climate science and get a deeper understanding of the Earth system. This requires involving scientists and practitioners from the domain of climate science.

To give an impression of how the new data set can be used, we apply several ML models on our CICMoD data set to predict Sahel rainfall and ENSO, respectively. In particular, we compare linear regression and MLP models to predict SPI. Results are shown in Section 4.1. Linear regression models perform slightly better on training data in terms of lower MSE combined with higher correlation of predictions and true targets, while MLP models show better performance on validation data, hence generalize better to unseen data. Comparing FOCI and CESM, we find lower MSE and higher correlation for linear regression and MLP models trained on indices derived from CESM data. As future work, these differences need to be further investigated.

As the second example, we predict current ENSO phase and ENSO phase 3 and 6 months into the future with CNN and LSTM models, respectively. Targets are derived from Niño 3.4 index and as predictors we use all remaining indices, excluding Niño indices. Input features and targets are found to be mostly uncorrelated with correlation coefficients in the range of -0.5 and 0.4 . Results are shown in Section 4.2. We find a higher frequency of ENSO events in time series derived from CESM data, compared to FOCI. Still, periodicity for El Niño events falls in the expected range of 2–7 years for both ESMs. Overall, our CNN models slightly outperform LSTM models for predicting ENSO. Again, we look at MSE and correlation for evaluating model performance. The longer the target horizon, the worse the model performance in terms of higher MSE and lower correlation, as expected, since ENSO is a complex phenomenon that hinders long-term prediction beyond several months. As for Sahel rainfall prediction, our ML models perform better on indices derived from CESM data, compared to FOCI, which needs to be further investigated in future work.

ESMs aim to simulate Earth system dynamics. Different ESMs have their individual strengths and weaknesses. For our CICMoD data set, we use two distinct ESMs to derive all indices. Whenever we find some relation in one model context, we may try to reproduce our findings on the other model's data to gain trust before repeating our experiments on real-world data. Like this, our CICMoD data set can help to reveal blind spots in ESMs and to find new causally linked modes within the real-world climate system. As future project, we plan to combine CICMoD with an extensive toolbox of explainable artificial intelligence (xAI) methods. Our new data set in combination with this xAI toolbox can then be used, for example, for data science competitions to tackle climate change and push the understanding of the climate system.

Author contribution. Conceptualization: M.L.-H., W.R., M.C., P.K.; Data curation: M.L.-H., S.W.; Data visualization: M.L.-H., N.N.; Methodology: M.L.-H., W.R., N.N.; Writing—original draft: M.L.-H., W.R., S.W.; Writing—review and editing: M.C., P.K. All authors approved the final submitted draft.

Competing interest. The authors declare no competing interests exist.

Data availability statement. The software project for our CICMoD data set is hosted on GitHub: https://github.com/MarcoLandtHayen/climate_index_collection. This work is based on release number v2023.03.29.1, which is published on Zenodo,

including the complete CICMoD data set as csv-file: <https://doi.org/10.5281/zenodo.7779883>. Furthermore, we reference a Docker container providing a Python environment with Jupyter notebooks, Tensorflow and climate_index_collection as pre-installed python package. Exemplary applications of our CICMoD data set to predict ENSO and Sahel rainfall are stored in a separate GitHub repository: https://github.com/MarcoLandtHayen/cicmod_application. Raw ESM data are stored on Zenodo: <https://doi.org/10.5281/zenodo.7774316>.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This work was supported by the Helmholtz School for Marine Data Science (MarDATA) funded by the Helmholtz Association (Grant HIDSS-0005).

References

- Abernathy R, Squire DNT, Nicholas T, Bourbeau JGG, Spring A, Bell R and Bailey S (2022) xhistogram (v.0.3.2) Zenodo. <https://doi.org/10.5281/zenodo.7095156>
- Badr HS, Zaitchik BF and Guikema SD (2014) Application of statistical models to the prediction of seasonal rainfall anomalies over the Sahel. *Journal of Applied Meteorology and Climatology* 53(3), 614–636. <https://doi.org/10.1175/JAMC-D-13-0181.1>
- Cai W, McPhaden MJ, Grimm AM, Rodrigues RR, Taschetto AS, Garreaud RD, Dewitte B, Poveda G, Ham Y-G, Santoso A, Ng B, Anderson W, Wang G, Geng T, Jo H-S, Marengo JA, Alves LM, Osman M, Li S, Wu L, Karamperidou C, Takahashi K and Vera C (2020) Climate impacts of the El Niño-southern oscillation on South America. *Nature Reviews Earth & Environment* 1, 215–231.
- Climate Diagnostics Bulletin (2016) National Oceanic and Atmospheric Administration (1996) 96(4).
- Drews A, Huo W, Matthes K, Kodera K and Kruschke T (2022) The Sun's role in decadal climate predictability in the North Atlantic. *Atmospheric Chemistry and Physics* 22(12), 7893–7904. <https://doi.org/10.5194/acp-22-7893-2022>
- Giannini A, Saravanan R and Chang P (2003) Oceanic forcing of Sahel rainfall on interannual to Interdecadal time scales. *Science* 302(5647), 1027–1030. <https://doi.org/10.1126/science.1089357>
- Gong D and Wang S (1999) Definition of Antarctic oscillation index. *Geophysical Research Letters* 26(4), 459–462.
- Horst AM, Hill AP and Gorman KB (2020) palmerpenguins: Palmer Archipelago (Antarctica) penguin data. *R package version 0.1.0*. Available at <https://allisonhorst.github.io/palmerpenguins/>; <https://doi.org/10.5281/zenodo.3960218> (accessed 24 March 2023).
- Hurrell JW (1995) Decadal trends in the North Atlantic oscillation: Regional temperatures and precipitation. *Science* 269(5224), 676–679.
- Hurrell JW, Holland MM, Gent PR, Ghan S, Kay JE, Kushner PJ, Lamarque J-F, Large WG, Lawrence D, Lindsay K, Lipscomb WH, Long MC, Mahowald N, Marsh DR, Neale RB, Rasch P, Vavrus S, Vertenstein M, Bader D, Collins WD, Hack JJ, Kiehl J and Marshall S (2013) The community earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society* 94, 1339–1360.
- Hurrell JW, Kushnir Y, Ottersen G and Visbeck M (2003) The North Atlantic oscillation: Climatic significance and environmental impact. *Geophysical Monograph Series* 134. <https://doi.org/10.1029/GM134>
- Joint Institute for the Study of the Atmosphere and Ocean (n.d.). Available at <http://research.jisao.washington.edu/data/> (accessed 24 March 2023).
- Jones PD, New M, Parker DE, Martin S and Rigor IG (1999) Surface air temperature and its changes over the past 150 years. *Reviews of Geophysics* 37(2), 173–199.
- Kingma DP and Ba JL (2014) Adam: A Method for Stochastic Optimization. In *Proceedings of the Third International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015. arXiv:1412.6980 (accessed 24 March 2023).
- Krizhevsky A (2009) Learning Multiple Layers of Features from Tiny Images. Available at <https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf> (accessed 24 March 2023).
- Lecun Y, Botou L, Bengio Y and Haffner P (1998) Gradient-based learning applied to document recognition. *IEEE* 86(11), 2278–2324.
- Li L, Schmitt RW, Ummenhofer CC and Karnauskas KB (2016) North Atlantic salinity as a predictor of Sahel rainfall. *Science Advances* 2(5), e1501588.
- Liao T, Taori R, Raji ID and Schmidt L (2021) Are we learning yet? A meta review of evaluation failures across machine learning. In *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. Red Hook, NY, USA: Curran Associates, Inc.
- Luccioni AS and Rolnick D (2022) Bugs in the Data: How Imagenet Misrepresents Biodiversity. arXiv Preprint <https://arxiv.org/abs/2208.11695> (accessed 24 March 2023).
- Makonin S, Wang ZJ and Tumpach C (2018) RAE: The rainforest automation energy dataset for smart grid meter data analysis. *Data* 3(1), 8.
- Mamalakis A, Ebert-Uphoff I and Barnes E (2022) Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science* 1, E8. <https://doi.org/10.1017/eds.2022.7>
- Marsh DR, Mills MJ, Kinnison DE, Lamarque JF, Calvo N and Polvani LM (2013) Climate change from 1850 to 2005 simulated in CESM1(WACCM). *Journal of Climate* 26(19), 7372–7391.

C. A Climate Index Collection based on Model Data

- Marshall GJ** (2007) Half-century seasonal relationships between the southern annular mode and Antarctic temperatures. *International Journal of Climatology* 27(3), 373–383.
- Matthes K, Biastoch A, Wahl S, Harlaß J, Martin T, Brücher T, Drews A, Ehlert D, Getzlaff K, Krüger F, Rath W, Scheinert M, Schwarzkopf FU, Bayr T, Schmidt H and Park W** (2020) The Flexible Ocean and climate infrastructure version 1 (FOCI1): Mean state and variability. *Geoscientific Model Development* 13(6), 2533–2568.
- Mayer KJ and Barnes EA** (2021) Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters* 48(10), e2020GL092092.
- Murphy P and Aha D** (1994) UCI repository of machine learning databases.
- National Center for Atmospheric Research** (n.d.). Available at <https://ncar.ucar.edu> (accessed 24 March 2023).
- National Oceanic and Atmospheric Administration** (n.d.). Available at <https://psl.noaa.gov/data/climateindices/> (accessed 24 March 2023).
- National Oceanic and Atmospheric Administration's National Weather Service** (n.d.). Glossary. Available at <https://w1.weather.gov/glossary/index.php?letter=h> (accessed 24 March 2023).
- Newman M, Alexander MA, Ault TR, Cobb KM, Deser C, Di Lorenzo E, Mantua NJ, Miller AJ, Minobe S, Nakamura H, Schneider N, Vimont DJ, Phillips AS, Scott JD and Smith CA** (2016) The Pacific decadal oscillation, revisited. *Journal of Climate* 29(12), 4399–4427. <https://doi.org/10.1175/JCLI-D-15-0508.1>
- Pak G, Park Y-H, Vivier F, Kwon Y-O and Chang K-I** (2014) Regime-dependent nonstationary relationship between the east Asian winter monsoon and North Pacific oscillation. *Journal of Climate* 27(21), 8185–8204.
- Park Y-H, Kim B-M, Pak G, Yamamoto M, Vivier F and Durand I** (2018) A key process of the nonstationary relationship between ENSO and the Western Pacific teleconnection pattern. *Scientific Reports* 8, 9512.
- Pegion K, Becker EJ and Kirtman BP** (2022) Understanding predictability of daily southeast US precipitation using explainable machine learning. *Artificial Intelligence for the Earth Systems* 1(4), e220011.
- Philander SG** (1989) *El Niño, La Niña, and the Southern Oscillation*, Vol. 46. San Diego, USA: Academic Press.
- Power SB and Kociuba G** (2011) The impact of global warming on the southern oscillation index. *Climate Dynamics* 37(9–10), 1745–1754. <https://doi.org/10.1007/s00382-010-0951-7>
- Raji ID, Gebru T, Mitchell M, Buolamwini J, Lee J and Denton E** (2020) Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 145–151. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3375627.3375820>
- Schlesinger ME and Ramankutty N** (1994) An oscillation in the global climate system of period 65–70 years. *Nature* 367, 723–726.
- Thompson DWJ and Wallace JM** (2000) Annular modes in the extratropical circulation. Part I: Month-to-month variability. *Journal of Climate* 13(5), 1000–1016. [https://doi.org/10.1175/1520-0442\(2000\)013](https://doi.org/10.1175/1520-0442(2000)013)
- Trenberth KE and Hurrell JW** (1994) Decadal atmosphere-ocean variations in the Pacific. *Climate Dynamics* 9(6), 303–319.
- Trenberth KE and Shea DJ** (2006) Atlantic hurricanes and natural variability in 2005. *Geophysical Research Letters* 33(12), L12704.
- Walker GT and Bliss EW** (1932) World weather V. *Memoirs of the Royal Meteorological Society* 4, 53–84.
- Watson-Parris D, Rao Y, Olivié D, Seland Ø, Nowack P, Camps-Valls G, Stier P, Bouabid S, Dewey M, Fons E, Gonzalez J, Harder P, Jeggle K, Lenhardt J, Manshausen P, Novitasari M, Ricard L and Roesch C** (2022) ClimateBench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems* 14(10), e2021MS002954. <https://doi.org/10.1029/2021MS002954>
- Zhang W, Mei X, Geng X, Turner AG and Jin F-F** (2019) A nonstationary ENSO–NAO relationship due to AMO modulation. *Journal of Climate* 32(1), 33–43.

Cite this article: Landt-Hayen M, Rath W, Wahl S, Niebaum N, Claus M, and Kröger P. (2023). A climate index collection based on model data. *Environmental Data Science*, 2: e9. doi:10.1017/eds.2023.5

Manuscript D

CICMoD - A Climate Index Collection Benchmark (Data and Resources Paper)

Marco Landt-Hayen, Willi Rath, Sebastian Wahl, Martin Claus

Published (SIGSPATIAL '23)

D. CICMoD - A Climate Index Collection Benchmark (Data and Resources Paper)

CICMoD - A Climate Index Collection Benchmark (Data and Resources Paper)

Marco Landt-Hayen
mlandt-hayen@geomar.de

GEOMAR Helmholtz Centre for Ocean Research
Kiel, Germany

Sebastian Wahl
swahl@geomar.de

GEOMAR Helmholtz Centre for Ocean Research
Kiel, Germany

Willi Rath
wrath@geomar.de

GEOMAR Helmholtz Centre for Ocean Research
Kiel, Germany

Martin Claus
mclaus@geomar.de

Christian-Albrechts-Universität zu Kiel
Kiel, Germany

ABSTRACT

In the domain of climate science, machine learning (ML) and in particular deep learning (DL) methods are known to be effective for identifying causally linked modes of climate variability as key to understand the climate system and to improve the predictive skills of forecast systems. To attribute climate events in a data-driven way, we need sufficient training data, which is often limited for real world measurements. The data science community provides standard data sets for many applications. As a new data set, we introduce a consistent and comprehensive collection of climate indices typically used to describe Earth System dynamics. Therefore, we use 1000-year control simulations from Earth System Models. The data set is provided as an open-source framework that can be extended and customized to individual needs. It allows users to develop new ML methodologies and to compare results to existing methods and models as benchmark.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Earth and atmospheric sciences**.

KEYWORDS

Benchmark Data Set, Time Series Forecasting, Data Mining

ACM Reference Format:

Marco Landt-Hayen, Willi Rath, Sebastian Wahl, and Martin Claus. 2023. CICMoD - A Climate Index Collection Benchmark (Data and Resources Paper). In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23)*, November 13–16, 2023, Hamburg, Germany. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3589132.3625627>

1 INTRODUCTION

To develop and compare machine learning (ML) methods in an objective way, there exist standard data sets as benchmark. Among

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0168-9/23/11...\$15.00

<https://doi.org/10.1145/3589132.3625627>

these data sets, we find e.g., a collection of handwritten digits also referred to as MNIST [1] or images, like the CIFAR-10 data set introduced by Krizhevsky [2]. Both data sets are mostly suitable for classification algorithms. Famous data sets for pattern recognition and clustering are e.g., Palmer Penguins [3] or the Wine data set [4]. Furthermore, the data science community also provides standard time series collections, like e.g., energy consumption time series for various household appliances [5]. However, benchmark data sets in the field of climate science are rare. To name a few, Mamelakis et al. provide a framework to create synthetic data sets designed for problems in geosciences [6]. And Watson-Parris et al. introduced ClimateBench, as a benchmark for data-driven climate projections [7].

Here, we are interested in describing the underlying dynamics of the Earth System. Real world data in this context are limited to observable features that can be measured in a comprehensive way or that can be reconstructed from sparse measurements. Examples are sea surface temperature (SST), sea level pressure (SLP), surface air temperature (SAT), sea surface salinity (SSS), geopotential height at various pressure levels, e.g., at 500 millibar (Z500), or total precipitation (PREC). These variables reflect some of the main dynamics of the Earth system in form of known modes of climate variability, patterns and oscillations, like e.g., Atlantic Multidecadal Oscillation (AMO) [8], the Southern Annular Mode (SAM) [9] or the El Niño Southern Oscillation (ENSO) [10]. To describe the Earth System dynamics in a compressed way, multi-dimensional geospatial data can be reduced to specific climate indices that capture the main processes. However, climate indices are limited in their temporal extent, since consistent real world measurements started only in recent history or measurements are subject of specific research projects that run over a certain period in time.

Our aim is to better understand existing modes of climate variability and to find new relationships. Therefore, we require a consistent and comprehensive collection of climate indices over a sufficiently long time span, which favors the use of model data over real world data. Earth System Models (ESMs) aim to simulate processes of the Earth system in specified temporal and spatial resolution. The Flexible Ocean and Climate Infrastructure (FOCI) [11] and the Whole Atmosphere Community Climate Model (WACCM) as extension of the Community Earth System Model (CESM) [12, 13] are both coupled, global climate models that provide state-of-the-art computer simulations of the past, present and future states of the

Table 1: All 29 indices included in CICMoD data set with their acronyms and spatial domains, ordered by the underlying feature

	Index	Acronym	Spatial Domain	
			Lat in °N	Lon in °E
Z500	Southern Annular Mode (PC-based) [15]	SAM_PC	-90 to -20	
SLP	Southern Annular Mode (zonal mean) [9]	SAM_ZM	-65 to -40	
	Southern Oscillation [16]	SOI	Tahiti (-18°N, 211°E)	Darwin (-12°N, 131°E)
	North Atlantic Oscillation (station) [17]	NAO_ST	Reykjavik (64°N, 338°E)	Ponta Delgada (38°N, 334°E)
	North Atlantic Oscillation (PC-based) [18]	NAO_PC	20 to 80 -90 to 40	
	North Pacific Pattern [19]	NP	30 to 65 -160 to 220	
SST	Atlantic Multidecadal Oscillation [20]	AMO	0 to 70 Atlantic basin	
	Pacific Decadal Oscillation [21]	PDO_PC	20 to 60	120 to 260
	El Niño Southern Oscillation (1+2)	ENSO_12	-10 to 0 270 to 280	
	El Niño Southern Oscillation (3)	ENSO_3	-5 to 5 210 to 270	
	El Niño Southern Oscillation (3.4) [22]	ENSO_34	-5 to 5 190 to 240	
	El Niño Southern Oscillation (4)	ENSO_4	-5 to 5 160 to 210	
	Tropical North Atlantic SSTA	SST_TNA	5 to 25 -55 to -15	
	Tropical South Atlantic SSTA	SST_TSA	-20 to 0 -30 to 10	
	Eastern Subtrop. Indian Ocean SSTA	SST_ESIO	-28 to -18 90 to 100	
	Western Subtrop. Indian Ocean SSTA [23]	SST_WSIO	-37 to -27 55 to 65	
	Mediterranean Sea SSTA	SST_MED	30 to 45 0 to 25	
Hurricane main dev. region SSTA	SST_HMDR	10 to 20 -85 to -20		
SSS	North Atlantic SSSA	SSS_NA	25 to 50 -50 to -15	
	Western North Atlantic SSSA	SSS_WNA	25 to 38 -50 to -40	
	Eastern North Atlantic SSSA	SSS_ENA	25 to 50 -40 to -15	
	South Atlantic SSSA [24]	SSS_SA	-22.5 to -10 -42 to -10	
SAT	Northern Hemisphere SATA	SAT_N_ALL	0 to 90	
	Northern Hemisphere SATA (ocean)	SAT_N_OCEAN	0 to 90	ocean
	Northern Hemisphere SATA (land) [25]	SAT_N_LAND	0 to 90	land
	Southern Hemisphere SATA	SAT_S_ALL	-90 to 0	
	Southern Hemisphere SATA (ocean)	SAT_S_OCEAN	-90 to 0	ocean
	Southern Hemisphere SATA (land)	SAT_S_LAND	-90 to 0	land
PREC	Sahel Precipitation [26]	PREC_SAHEL	10 to 20	-20 to 10

Earth system. Here, we use the output of FOCI and CESM control runs. In particular, we work with SST, SAT, SLP, Z500, SSS and PREC as two-dimensional fields. From these variables, we derive a set of climate indices over 1,000 and 999 years, respectively. The obtained collection of climate indices based on model data (CICMoD) serves as a reduced description of the Earth system in a consistent and comprehensive way. Our main contributions are as follows:

- We introduce CICMoD as a new benchmark data set describing the climate system.
- CICMoD allows the user to develop new ML methods and to compare results to existing methods.
- We provide an open-source framework that can be extended and customized to individual needs.

The rest of this work is structured as follows. In Section 2 we provide a short description of FOCI and CESM. An overview of all indices included in the CICMoD data set is given in Section 3. A discussion and conclusion is found in Section 4.

2 MODEL DATA

The climate indices included in our CICMoD data set are derived from monthly averaged output of climate model simulations with FOCI and CESM, respectively. The CESM simulation is based on version 1.0.6 with WACCM version 4 [14]. The FOCI simulation used in this manuscript is the control simulation referred to as “FOCI-piCl” based on FOCI version 1.3.0 [11]. Both simulations were run using pre-industrial external forcing that is representative for the year 1850. FOCI (1.8° × 1.8°, 95 vertical levels) and CESM (1.8° × 2.5°, 106 vertical levels) were run at similar horizontal and vertical resolution, although the vertical distribution of the model layers differs significantly between FOCI and CESM. Both models have been extensively evaluated and used in various climate studies. FOCI and CESM are based on very different component models (see [11–13] for details) with different strengths and weaknesses in simulating various aspects of the global climate. From both control simulations’ output we use Z500, SLP, SST, SSS, SAT and PREC.

D. CICMoD - A Climate Index Collection Benchmark (Data and Resources Paper)

CICMoD - A Climate Index Collection Benchmark
(Data and Resources Paper)

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

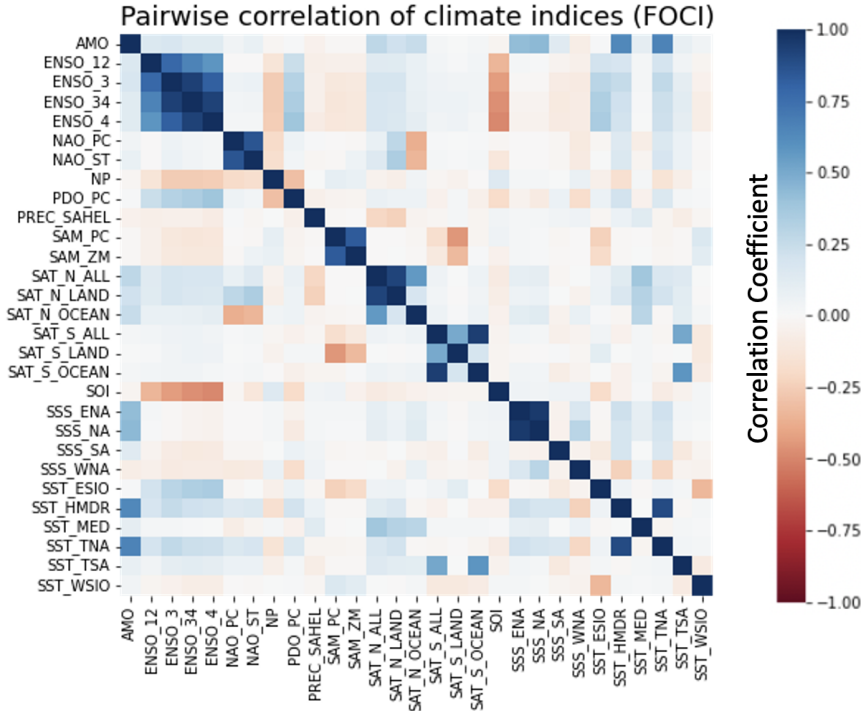


Figure 1: Pairwise correlation coefficients of all CICMoD indices derived from FOCI data

3 CLIMATE INDEX COLLECTION

In this section we give an overview of all 29 indices included in the CICMoD data set. The indices can be grouped by the underlying feature. Each feature is briefly put into context. Furthermore, we provide references on how the indices are derived in detail.

- **Geopotential height** is a vertical coordinate with reference to Earth’s mean sea level. Its contours are used to calculate the geostrophic wind which is of interest for climate dynamics.
- **Sea level pressure** refers to the air pressure at sea level. Several indices are derived from SLP and its anomalies.
- **Sea surface temperature** is the ocean temperature close to the surface. SST anomalies (SSTA) impact the energy transfer at the interface between ocean and atmosphere.
- **Sea surface salinity** measures the amount of salt dissolved in the ocean surface water and plays an important role in ocean circulation processes. Furthermore, rainfall on land is largely supplied by evaporation over the ocean and that

evaporation leaves an imprint in SSS. Several indices are derived from SSS anomalies (SSSA).

- **Surface air temperature** is the air temperature close to the surface and relates to the ability of evaporation, since warmer air has a higher storage capacity for water vapor. SAT anomalies (SATA) influence the energy transfer at the interface between Earth’s surface and atmosphere.
- **Precipitation** has a high impact on society in form of extreme events like flooding caused by heavy rainfall or droughts due to missing or lower as normal rainfall.

Table 1 gives an overview of all 29 indices included in CICMoD data set ordered by the underlying feature. For details on how each index is defined, we provide references. By definition, all indices have zero mean over time, whereas only NAO_PC, PDO_PC, SAM_PC, SAM_ZM and SOI are normalized by design to have unit variance. If required, normalization of the remaining indices can be done in pre-processing. Pairwise correlation coefficients for all indices included in CICMoD data set derived from FOCI data are

shown in Figure 1. Indices derived from CESM data show similar characteristics. ENSO indices are found to be highly correlated, as expected, since these indices are all computed from SSTA in some narrow region of the Tropical Pacific. Additionally, we find station- and PC-based NAO indices to be highly correlated, as well as PC-based SAM and SAM from zonal mean, as these indices are designed to describe the same processes. Indices regarding SATA in the NH and SH, respectively, and indices regarding SSS anomalies in the North Atlantic also reveal similarities in terms of high correlation, since by design spatially related features are involved in the computation. Besides that, SOI is found to be negatively correlated to all ENSO indices. Periods of negative (positive) SOI values coincide with warmer (colder) than normal ocean water across the Eastern Tropical Pacific, which is typical for El Niño (La Niña) episodes [27].

4 DISCUSSION AND CONCLUSION

In this work, we introduced a consistent and comprehensive collection of climate indices as a new benchmark data set. The collection is consistent in a sense that we use the output of ESM control runs to derive all indices. For FOCI and CESM control runs, we have 1,000 and 999 years of monthly data, respectively, as an advantage compared to real world data, since ML models require sufficient training data. The collection is comprehensive as we include a broad selection of known patterns, oscillations and variability of the Earth system. The index collection is not complete since we focus on processes within the atmosphere, in the upper ocean and at the interface of ocean and atmosphere. However, our CICMoD data set serves as basis. Additionally, we provide an open-source framework that can be extended and customized to individual needs including the application to further ESMs. This opens the door for collaboration in many ways. Our new data set allows researchers from the data science community to adapt existing ML models and develop new ML methods to tackle problems from the domain of climate science and get a deeper understanding of the Earth system. This requires involving scientists and practitioners from the domain of climate science.

To give an impression of how the new data set can be used, we applied several ML models to our CICMoD data set and tried to predict rainfall in the African Sahel region and ENSO, respectively. For the results, we refer to our GitHub repositories. The corresponding links are provided below.

DATA AVAILABILITY STATEMENT

The software project for our CICMoD data set is hosted on GitHub: https://github.com/MarcoLantHayen/climate_index_collection. This work is based on release number *v2023.03.29.1*, which is published on Zenodo, including the complete CICMoD data set as csv-file: <https://doi.org/10.5281/zenodo.7779883>. Furthermore, we reference a Docker container providing a Python environment with Jupyter notebooks, Tensorflow and climate_index_collection as pre-installed python package. Exemplary applications of our data set to predict ENSO and Sahel rainfall are stored in a separate GitHub repository: https://github.com/MarcoLantHayen/cicmod_application. Raw ESM data are stored on Zenodo: <https://doi.org/10.5281/zenodo.7774316>.

FUNDING STATEMENT

This work was supported by the Helmholtz School for Marine Data Science (MarDATA) funded by the Helmholtz Association (Grant HIDSS-0005).

REFERENCES

- [1] Lecun et al. (1998) Gradient-based learning applied to document recognition, *IEEE* 86(11), 2278–2324.
- [2] Krizhevsky A. (2009) Learning Multiple Layers of Features from Tiny Images <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- [3] Horst et al. (2020) palmerpenguins: Palmer Archipelago (Antarctica) penguin data, *R package version 0.1.0*.
- [4] Murphy P. and Aha D. (1994) UCI repository of machine learning databases.
- [5] Makonin et al. (2018) RAE: The Rainforest Automation Energy Dataset for Smart Grid Meter Data Analysis, *Data* 3(1):8.
- [6] Mamalakis et al. (2022) Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset, *Environmental Data Science*, 1(E8).
- [7] Watson-Parris et al. (2022) ClimateBench v1. 0: A Benchmark for Data-Driven Climate Projections, *Advances in Modeling Earth Systems*, 14(10).
- [8] Schlesinger M.E. and Ramankutty N. (1994) An oscillation in the global climate system of period 65-70 years, *Nature* 367, 723–726.
- [9] Gong D. and Wang S. (1999) Definition of Antarctic oscillation index, *Geophysical Research Letters* 26(4), 459–462.
- [10] Philander S.G. (1989) El Niño, La Niña, and the Southern Oscillation, vol. 46, *Academic Press*, San Diego, USA.
- [11] Matthes et al. (2020) The Flexible Ocean and Climate Infrastructure version 1 (FOCI1): mean state and variability, *Geoscientific Model Development* 13(6), 2533–2568.
- [12] Hurrell et al. (2013) The Community Earth System Model: A framework for collaborative research, *Bulletin of the American Meteorological Society*, 94, 1339–1360.
- [13] Marsh et al. (2013) Climate change from 1850 to 2005 simulated in CESM1(WACCM), *Journal of Climate*, 26(19), 7372–7391.
- [14] Drews et al. (2022) The Sun's role in decadal climate predictability in the North Atlantic, *Atmospheric Chemistry and Physics*, 22(12), 7893–7904.
- [15] Thompson et al. (2000) Annular Modes in the Extratropical Circulation. Part I: Month-to-Month Variability, *Journal of Climate*, 13(5), 1000–1016.
- [16] Walker G.T. and Bliss E.W. (1932) World Weather V., *Memoirs of the Royal Meteorological Society*, 4, 53–84.
- [17] Hurrell J.W. (1995) Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation, *Science*, 269(5224), 676–679.
- [18] National Center for Atmospheric Research, <https://ncar.ucar.edu>.
- [19] Trenberth K.E. and Hurrell J.W. (1994) Decadal atmosphere-ocean variations in the Pacific, *Climate Dynamics*, 9(6), 303–319.
- [20] Trenberth K.E. and Shea D.J. (2006) Atlantic hurricanes and natural variability in 2005, *Geophysical Research Letters*, 33(12).
- [21] Newman et al. (2016) The Pacific Decadal Oscillation, Revisited, *Journal of Climate*, 29(12), 4399–4427.
- [22] National Oceanic and Atmospheric Administration, <https://psl.noaa.gov/data/climateindices/>.
- [23] Giannini et al. (2003) Oceanic Forcing of Sahel Rainfall on Interannual to Interdecadal Time Scales, *Science*, 302(5647), 1027–1030.
- [24] Li et al. (2016) North Atlantic salinity as a predictor of Sahel rainfall, *Science Advances*, 2(5).
- [25] Jones et al. (1999) Surface air temperature and its changes over the past 150 years, *Reviews of Geophysics*, 37(2), 173–199.
- [26] Badr et al. (2014) Application of Statistical Models to the Prediction of Seasonal Rainfall Anomalies over the Sahel, *Applied Meteorology and Climatology* 53(3).
- [27] Power S.B. and Kociuba G. (2011) The impact of global warming on the Southern Oscillation Index, *Climate Dynamics*, 37(9–10), 1745–1754.

A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

Marco Landt-Hayen, Yannick Wölker, Willi Rath, Martin Claus

Accepted for Publication at ADMA 2023

E. A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs*

Marco Landt-Hayen^{1,2[0000-0003-3606-7760]}, Yannick Wölker^{1,2[0009-0006-6076-8996]}, Willi Rath^{2[0000-0003-1951-8494]}, and Martin Claus^{1,2[0000-0002-7525-5134]}

¹ Christian-Albrechts-Universität zu Kiel, Germany

² GEOMAR Helmholtz Centre for Ocean Research, Germany

Abstract. Working with observational data in the context of geophysics can be challenging, since we often have to deal with missing data. This requires imputation techniques in pre-processing to obtain data-mining-ready samples. Here, we present a convolutional neural network (CNN) approach from the domain of deep learning to reconstruct complete data from sparse inputs. CNN architectures are state-of-the-art for image processing. As data, we use two-dimensional fields of sea level pressure (SLP) and sea surface temperature (SST) anomalies. To have consistent data over a sufficiently long time span, we favor to work with output from control simulations of two Earth System Models (ESMs), namely the Flexible Ocean and Climate Infrastructure and the Community Earth System Model. Our networks can restore complete information from incomplete input samples with varying rates of missing data. Moreover, we present a technique to identify the most relevant grid points of our input samples. Choosing the optimal subset of grid points allows us to successfully reconstruct SLP and SST anomaly fields from ultra sparse inputs. As a proof of concept, the insights obtained from ESMs can be transferred to real world observations to improve reconstruction quality. As uncertainty measure, we compare several climate indices derived from reconstructed versus complete fields.

Keywords: Missing value imputation · Optimal sampling strategy · Convolutional neural networks · Explainable AI.

1 Introduction

Geospatial data in the context of ocean and atmosphere are often provided on a two-dimensional latitude-longitude grid. Examples are sea level pressure (SLP) or sea surface temperature (SST). Observational data are obtained e.g., from

* This work was supported by the Helmholtz School for Marine Data Science (MarDATA) funded by the Helmholtz Association (Grant HIDSS-0005).

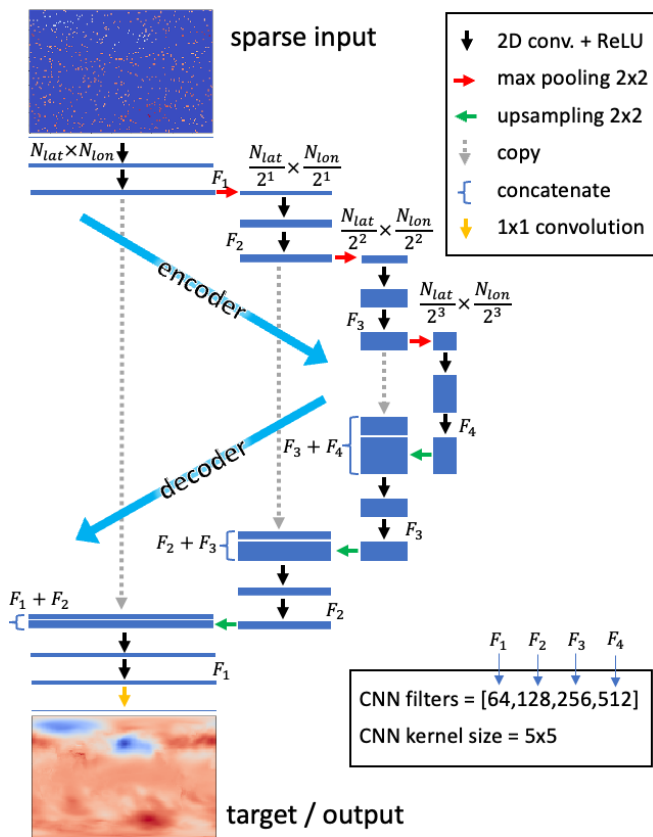


Fig. 1. Encoder-decoder-style U-Net architecture used for reconstructing complete two-dimensional geospatial fields from sparse inputs. In this sketch, we show an input sample for SLP CESM data with 95% missing values and the corresponding complete sample used as target. Blue rectangles symbolize resulting feature maps. Grid dimensions (N_{lat} and N_{lon}) in combination with pooling or upsampling operations determine height and width of feature maps. The depth is specified by the number of CNN filters $F_1..F_4$.

E. A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

remote sensing by satellites or from in situ measurements by survey stations, ships and planes. Observations are usually incomplete due to technical, physical or economical reasons. For instance, instrumental errors can lead to missing data. Moreover, infrared radiation is absorbed by clouds.

As prerequisite for advanced data mining, we need consistent and complete data. This requires imputation of missing values. Beckers and Rixen [1] introduced DINEOF as a method to infer missing data from oceanographic data series using empirical orthogonal functions. DINEOF has been and still is a widely used technique in the context of geospatial data. To name a few, Alvera-Azcárate et al. [2] applied the technique to SST data with rates of missing data typically in the range of 40 – 80%. DINEOF has also been used for multivariate reconstruction of missing data [3] with a focus on specific regions of interest [4]. The method is powerful, self-consistent and easy to use. But it has its limits in the allowed rates of missing values. For instance, SST samples with cloud coverage exceeding 95% are eliminated in pre-processing, since DINEOF fails in this regime. Chai et al. [5] proposed an encoder-decoder-style U-Net convolutional neural network (CNN) for reconstructing seismic data with regularly and irregularly missing values and found it to be superior compared to Fourier transform interpolation methods. In their work, rates of missing values were also limited to 95%. Similarly, Barth et al. [6] present a convolutional autoencoder to reconstruct missing SST observations and showed that deep neural networks are currently state-of-the art for this task.

In this work, we investigate methods for reconstructing full two-dimensional geospatial fields from incomplete input data on a *global* scale. We introduce a CNN that can handle input data with varying rates of missing values and go beyond existing approaches allowing to have ultra sparse inputs with up to 99.9% missing values. To overcome existing limits, we propose a bottom-up sampling strategy and set the benchmark for the ultra sparse regime. We show how to identify grid points that are essential to reconstruct dominant structures in SLP and SST anomaly fields from only 0.1% of the original data. Furthermore, we assign certain scores that can be visualized as a heat map to give an intuitive understanding of grid points' relevance for *reconstruction*. Thus, our approach differs from existing methods that aim to find feature relevance for *model predictions* like e.g., layer-wise relevance propagation (LRP) [7] or shapley additive explanations (SHAP) [8].

To have consistent data over a sufficiently long time span, we favor to work with output from control simulations of two Earth System Models (ESMs), namely the Flexible Ocean and Climate Infrastructure (FOCI) [9] and the Whole Atmosphere Community Climate Model as extension of the Community Earth System Model (CESM) [10,11]. In particular, we work with SLP and SST anomaly fields over 1,000 and 999 years obtained from FOCI and CESM, respectively. Eventually, we transfer results from ESMs to real world (RW) observations. Our main contributions are as follows:

- We present CNN models that can reconstruct missing data from inputs with varying rates of missing values and clearly outperform DINEOF.

- Our framework is provided as open-source project that can be extended and customized to individual needs, e.g., by including further methods or data.
- We propose a bottom-up sampling strategy to identify grid points in geospatial input fields that are most relevant for successful reconstruction.
- With this optimal subset of grid points, we restore complete information from ultra sparse inputs and set the benchmark in this regime.
- We introduce relative loss reduction maps to give a visual and intuitive understanding of grid points’ relevance for reconstruction.
- As a proof of concept, we transfer insights obtained from ESMs to RW observations to improve reconstruction quality.

The rest of this work is structured as follows. In Section 2 we provide a short description of ESM and observational data and outline data pre-processing. An overview of our U-Net models is given in Section 3. Moreover, we describe our sampling strategy and introduce relative loss reduction maps in Section 3. In Section 4 we apply our U-Net models and DINEOF to ESM and observational data. Additionally, we evaluate model performance and compute several climate indices on reconstructed versus complete SLP and SST fields. Discussion of all results and a conclusion are found in Section 5.

2 Data

In this section, we briefly describe ESM and observational data before we outline data pre-processing. We start with monthly mean SLP and SST fields from FOCI and CESM control runs, respectively. Both simulations were run using pre-industrial external forcing that is representative for the year 1850. FOCI and CESM are based on very different component models (see [9,10,11] for details) with different strengths and weaknesses in simulating various aspects of the

Table 1. Overview of CESM, FOCI and RW data used throughout this work. The grid resolution determines the grid dimensions in latitude and longitude. The number of valid grid points excludes permanently missing data, due to land masses. Finally, we show the total number of samples and the time span in years.

Source	Feature	Grid resolution	Grid dimensions $N_{lat} \times N_{lon}$	Valid grid points	Samples (years)
CESM	SLP	$1.8^\circ \times 2.5^\circ$	96×144	13,824	11,988 (999)
	SST			8,276	
FOCI	SLP	$1.8^\circ \times 1.8^\circ$	96×192	18,432	12,000 (1,000)
	SST			12,949	
RW	SLP	$2.5^\circ \times 2.5^\circ$	72×144	10,368	900 (75)
	SST	$2^\circ \times 2^\circ$	80×176	9,913	1,716 (143)

E. A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

global climate. Additionally, we use reanalysis data provided by the National Oceanic and Atmospheric Administration (NOAA) as RW observations for SLP and SST [12]. Table 1 gives an overview of all data.

SLP and SST anomaly fields are obtained from raw data by removing the seasonal cycle over a specified climatology period. For RW data, we choose years 1980 through 2009 as climatology period, whereas for ESM data, we use the whole time span. To simulate missing data, we use different settings:

- **Fixed mask:** Randomly choose a subset of grid points as missing data with a discrete rate of missing values. This subset is then identical for *all* samples.
- **Variable discrete mask:** Only fix the discrete rate of missing values. Then, randomly choose subsets of grid points as missing values. Thus, different samples have different subsets of missing values. As extension, we can use each sample multiple times. Here, we use factors 1, 2 and 3 for data augmentation.
- **Variable range mask:** The set of missing values still differs for different samples, as for the variable discrete mask. Additionally, the rate of missing values is randomly drawn from a specified range. Our models allow maximum flexibility. We can set the range from 0 to 100% missing values.
- **Optimal mask:** Once we identified subsets of grid points that are most relevant for successful reconstruction, we create a fixed mask from taking the remaining grid points as missing data. Here, we use optimal masks only for distinct rates of missing values.

Moreover, anomaly fields are scaled to $[0, 1]$ with minimum and maximum values derived from training data. After scaling, missing values are set to *zero*. While SLP data is defined everywhere, we don't have SST data over land masses. Permanently missing values are also set to *zero*. 80% of all samples are used for training, while the remaining 20% are reserved as test data for model evaluation.

3 Models and Methods

We give an overview of our U-Net models in Section 3.1. Furthermore, we describe our sampling strategy and introduce relative loss reduction maps in Section 3.2. For a detailed technical description of the DINEOF methodology, we refer to Beckers and Rixen [1]. In their work, they introduced DINEOF as a method to infer missing data from oceanographic data series using empirical orthogonal functions.

3.1 U-Nets

In this work, we focus on encoder-decoder-style CNNs, also referred to as U-Nets [13]. A thorough introduction to deep learning in general and CNNs in particular can be found e.g., in the textbook of Goodfellow et al. [14]. A sketch of our U-Net architecture with 4 convolution blocks in the encoder part is shown in Figure 1. As inputs, we use scaled two-dimensional *sparse* anomaly fields.

Scaled *complete* anomaly fields serve as targets. Each convolution block in the encoder path consists of two convolution operations with strides set to one and zero padding to conserve height and width dimensions of the resulting feature maps plus a rectified linear unit (ReLU) activation. The convolution blocks are followed by a 2×2 maximum pooling operation to reduce height and width. Maximum depth of 512 feature maps is reached after 4 convolution blocks. In the decoder path we use skip connections to prevent vanishing gradients [15] and 2×2 upsampling to restore former height and width dimensions. Ultimately, we have a 1×1 convolution operation to obtain a single output channel. The kernel size for convolution operations is set to 5×5 . The number of convolution filters F_i determines the depth of the resulting feature maps in the i -th convolution block. For our standard U-Net configuration, we have an increasing number of 64, 128, 256 and 512 filters in the four convolution blocks, respectively. This configuration is used throughout this work. Only for models trained on SST RW data, we add an additional fifth convolution block with $F_5 = 1,024$ filters and reduce kernel size for all convolutions to 4×4 to optimize performance. The U-Net models are trained over 10 epochs using the ADAM optimizer [16] with a mean squared error (mse) loss function. The learning rate is set to 10^{-4} and 10^{-5} for SLP and SST, respectively. The batch size is set to 10. All hyperparameters are the result of a grid search optimization. Therefore, we temporarily used 20% of the training data as validation data.

3.2 Relative Loss Reduction Maps

Assume that we have geospatial input data with a total number v of valid grid points. Furthermore, we assume to have a specific rate $m \in [0, 1]$ of missing values. In a first step, we aim to find the subset G_m^s of the most relevant grid points for a single sample s that are essential to restore complete information. Thus, G_m^s contains $g_m \equiv |G_m^s| = (1 - m) \cdot v$ grid points, where g_m is rounded to the nearest integer number. Moreover, we assume to have a U-Net model trained on samples over the full range of missing values $[0, 1]$ with a variable range mask. We then pick a single sample s and compute the mean state loss (MSL) and the minimum loss (MinL) by using an empty sample of only zeros and the complete sample, respectively, as input for our range model. The difference of MSL and MinL determines the maximum absolute loss reduction (MALR) that we can achieve for a specific sample s , as stated in Equation 1. Superscript s denotes the sample number.

$$MALR^s = MSL^s - MinL^s \quad (1)$$

To find the optimal subset G_m^s of grid points for a single sample, we start with an empty sample and successively add grid points $g = 1..g_m$. The more information we provide to the U-Net, the lower the reconstruction loss. To add grid point g , we choose the one from all $v - g + 1$ remaining valid grid points, that leads to maximum decline in reconstruction loss. The absolute loss reduction (ALR) for

E. A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

Reconstruct Geospatial Data from Ultra Sparse Inputs 7

adding a grid point g to an input sample s can be translated into relative loss reduction (RLR) as follows:

$$RLR_g^s = ALR_g^s / MALR^s \quad (2)$$

The relative loss reduction scores $\{RLR_g^s\}_{g=1..g_m}$ are assigned to the corresponding grid points and can be visualized as heat map to give an intuitive understanding of grid points' relevance for reconstruction. The subset of most relevant grid points can vary from sample to sample. Here, we average scores over the first and last 120 training samples to obtain a stable mean relative loss reduction map (MRLRM) for a given geospatial feature. To reduce computational effort, we stop adding grid points when the accumulated relative loss reduction exceeds a threshold t . Here, we set $t = 0.9$ as a tradeoff between computation time and information gain. Finally, we derive optimal masks $G_{m=0.999}$, $G_{m=0.99}$ and $G_{m=0.95}$ that contain most relevant grid points for 99.9%, 99% and 95% missing values, respectively, for a given feature (SLP or SST) and source (CESM or FOCI). Therefore, we fit a set of bivariate Gaussians to each MRLRM using Gaussian Mixture Models (GMMs) [17] and take the grid points with highest RLR as cluster representatives. To summarize the algorithm for deriving optimal masks:

- **Step 1:** Train U-Net with variable range mask.
- **Step 2:** Compute relative loss reduction maps (RLRMs) for a representative subset of training samples.
- **Step 3:** Get MRLRM by averaging over RLRMs obtained in previous step.
- **Step 4:** Derive optimal masks for desired rates of missing values using GMMs.

4 Application and Results

In Section 3.2, we introduced the methodology to derive subsets of most relevant grid points from MRLRMs for various rates of missing values. In Section 4.1, we transfer these subsets (also referred to as *optimal masks*) obtained from ESM data to RW data. We then apply our U-Net models and DINEOF to RW data in Section 4.2 to reconstruct complete information from sparse input data for different types of masks. Additionally, we evaluate model performance and compute several climate indices on reconstructed versus complete SLP and SST fields in Section 4.3.

4.1 Derive and Transfer Optimal Masks

Figure 2 shows MRLRMs for ESM data for both features, SLP and SST. From these MRLRMs, we derive optimal masks. In order to transfer optimal masks

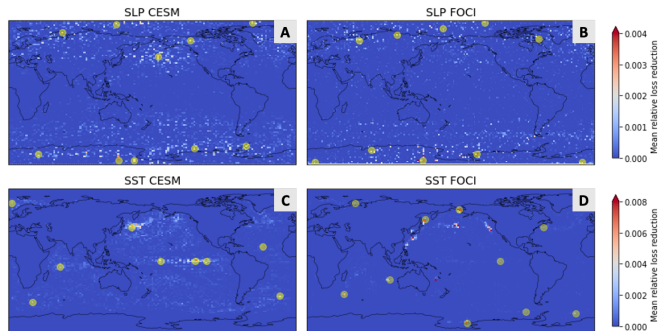


Fig. 2. The upper part shows MRLRMs for SLP obtained from CESM (A) and FOCI (B) data. The lower part shows results for SST CESM (C) and FOCI (D) data, respectively. The maps are averaged over first and last 120 training samples. The 10 most relevant grid points obtained from GMM are highlighted as yellow dots. These grid points represent 0.1% of valid grid points for SLP and SST RW data and hence, form the optimal masks for 99.9% missing values, transferred from ESM to RW data.

from ESM to RW data, we need to consider the number of valid grid points for our target grid (RW) from Table 1. Here, we focus on the ultra sparse regime with 99.9%, 99% and 95% missing values. Accordingly, for SLP RW data we aim to find the most relevant 10, 104 and 518 grid points, whereas for SST RW data, we look for the most relevant 10, 99 and 496 grid points, representing 0.1%, 1% and 5% of valid grid points, respectively. As an example, we highlight resulting cluster representatives from GMM for 99.9% missing data (10 grid points) in Figure 2. Eventually, we need to deal with the fact that ESM and RW data are on different grids. Therefore, we use nearest neighbour interpolation to transfer optimal masks to the target grid.

4.2 Reconstruction with U-Nets and DINEOF

In Section 2, we introduced different mask types to simulate missing data. In this section, we show results for U-Net models trained on RW data with fixed and variable masks with augmentation factors 1, 2 and 3, respectively, for distinct rates of missing values $m \in \{0.999, 0.99, 0.95, 0.9, 0.75, 0.5\}$. Additionally, we show the test loss for models trained on our optimal masks transferred from ESM to RW data in the ultra sparse regime for $m \in \{0.999, 0.99, 0.95\}$. As a benchmark, we include results from reconstruction with DINEOF for $m \in \{0.99, 0.95, 0.9, 0.75, 0.5\}$.

The main interest of this work lies in the attempt to optimally reconstruct information from ultra sparse inputs. From inspecting loss on test data in Figure

E. A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

Reconstruct Geospatial Data from Ultra Sparse Inputs 9

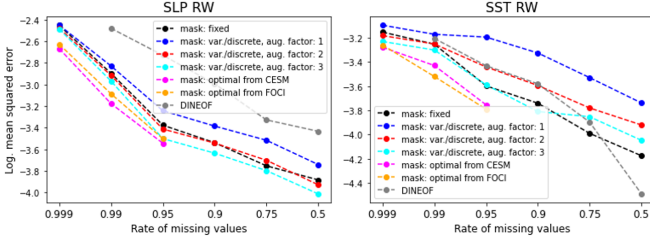


Fig. 3. Loss (mse) for test data on a logarithmic scale for U-Net models trained on SLP (left) and SST (right) RW data with various types of missing masks and distinct rates of missing values. All results are averaged over three runs with random seeds 1, 2 and 3, respectively. For comparison, we show corresponding loss for reconstruction with DINEOF.

3, best performance is found for models trained on optimal masks transferred from ESM data. Most competitive are the models trained on variable masks with augmentation factor 3. In Figure 4, we show reconstruction for an individual SLP RW test sample with 99.9% and 99% missing data, respectively, applying best performing U-Net models. For comparison, we also show the final reconstruction for DINEOF from 99% missing data. For 99.9% missing data, DINEOF fails.

The target field shows two dominant spots of positive SLP anomaly (red) in the upper center and upper left part of the sample. Moreover, we find alternating positive and negative SLP anomalies in high latitudes (40° to 70° N) and low latitudes (-40° to -70° N), respectively. The reconstruction of the sample from only 0.1% of the data (99.9% missing) using a model trained on a variable mask with augmentation factor 3 appears to be blurry and does not match the corresponding target. Whereas, reconstructions from models trained on optimal masks restore either the most dominant SLP anomaly spot in the upper center or upper left part and we also see a rudimentary replica of the alternating patterns of positive and negative SLP anomalies.

The reconstruction from 1% of the data (99% missing) using a model trained on a variable mask better fits the original target field, as it shows at least the dominant spot of positive SLP anomaly in the upper left part and alternating positive and negative anomalies in low latitudes. Compared to that, the reconstructions of both models trained on optimal masks are far more precise as they restore the correct location and shape for *both* dominant spots of positive SLP anomaly plus the alternating patterns of positive and negative anomalies. The final reconstruction with DINEOF appears to be noisy compared to results from our U-Nets. At least, the alternating positive and negative anomalies are restored in a rudimentary way. Similar results are obtained for reconstruction of SST samples. But for the sake of brevity, we only show results for SLP samples, here.

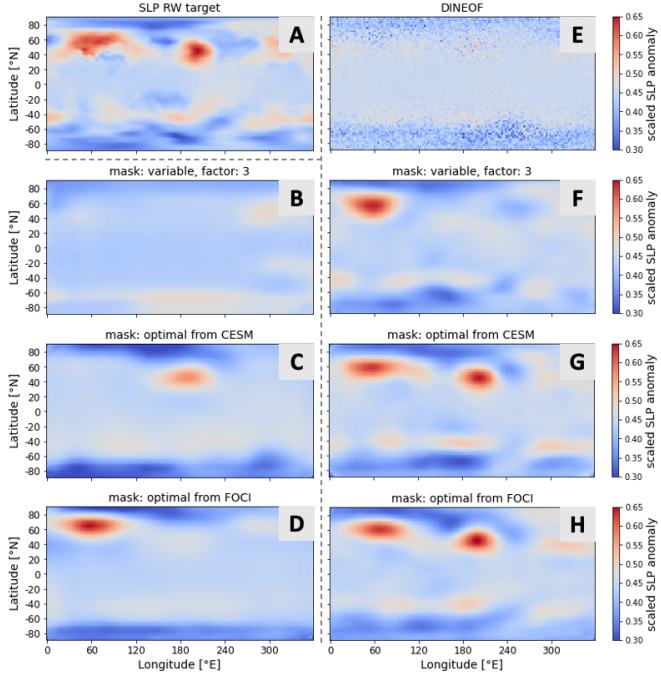


Fig. 4. Reconstruction of a complete SLP RW anomaly field used as target (A) from 0.1% and 1% of all data, hence, with 99.9% (left part) and 99% (right part) missing values, respectively. Here, we show results for best performing U-Net models. In particular, we use models trained on variable masks with augmentation factor 3 (B and F), models trained on optimal masks transferred from CESM (C and G) and FOCI (D and H). For comparison, we show the final reconstruction for DINEOF (E) from 99% missing data.

E. A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

4.3 Model Evaluation

In this section, we evaluate model performance. The features we use throughout this work (SLP and SST), reflect some of the main dynamics of the Earth system in form of known modes of climate variability, patterns and oscillations, like e.g., the Southern Annular Mode (SAM) [18], the North Atlantic Oscillation (NAO) [19], the Atlantic Multidecadal Oscillation (AMO) [20] or the El Niño Southern Oscillation (ENSO) [21].

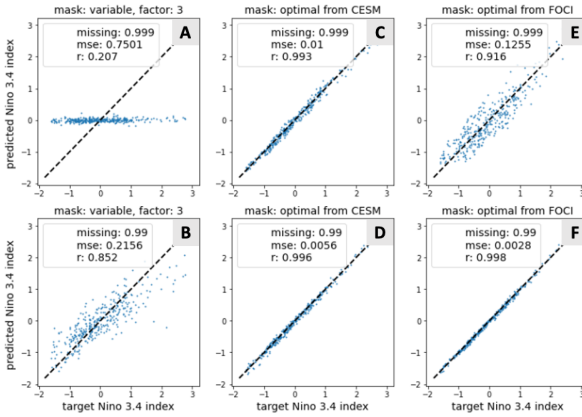


Fig. 5. Compare Niño 3.4 index obtained from reconstructed versus complete SST anomaly fields. Each blue dot represents a RW test sample. We compute indices on reconstructions from 0.1% (upper part) and 1% (lower part) of all data, hence, with 99.9% and 99% missing values, respectively. Here, we show results for best performing U-Net models. In particular, we include models trained on variable masks with augmentation factor 3 (A and B), models trained on optimal masks transferred from CESM (C and D) and FOCI (E and F). Loss (mse) and correlation coefficient are stated for each model and rate of missing values. The dashed black line is the identity.

To describe the Earth System dynamics in a compressed way, two-dimensional geospatial fields can be reduced to specific climate indices that capture the main processes. The SAM index was originally defined by Gong and Wang [18] as the difference of normalized monthly zonal mean SLP at 40°S and 65°S, respectively. The NAO index can be computed from SLP as the normalized difference between Reykjavik (64°9'N, 21°56'W) and Ponta Delgada (37°45'N, 25°40'W) [19]. AMO refers to a natural variability occurring in the SST of the North Atlantic with a multidecadal period of 60 to 80 years. The AMO index is computed from

area-weighted SST anomalies of the North Atlantic [22]. ENSO is a complex phenomenon that can be detected as periodic SST fluctuations in the Tropical Pacific. Several indices are defined to compute the current ENSO phase from area-averaged SST anomalies in certain regions. For instance, Morrow et al. [23] define the Niño 3.4 index from the Niño 3.4 region (5°N–5°S, 120°W–170°W) which is often used in the context of ENSO. In Figure 5, we show the Niño 3.4 index computed on reconstructed versus complete SST RW anomaly fields (test data). We again focus on the ultra sparse regime with 99.9% and 99% of missing data and compare results for models trained on variable masks with augmentation factor 3 with results for models trained on optimal masks derived from CESM and FOCI data. These models perform best, as shown in Section 4.2. As metric, we use the mse of predicted versus true Niño 3.4 index and correlation coefficient r . Clearly, highest fidelity in terms of low mse combined with a high degree of correlation is found for models trained on optimal masks. Similar results are obtained for SAM, NAO and AMO indices, as shown in Table 2.

5 Discussion and Conclusion

In this work, we presented encoder-decoder-style U-Net models to reconstruct complete information from sparse geospatial input data. To have consistent data over a sufficiently long time span, we favored to work with ESM data since RW data from reanalysis already contains reconstructed information from statistical methods.

To simulate missing data, we used different types of masks. Additionally, we have permanently missing data for SST in terms of land masses. All missing values were set to zero. Our U-Nets showed maximum flexibility in a sense, that

Table 2. Compare climate indices corresponding to SAM, NAO and AMO obtained from reconstructed versus complete SLP and SST fields, respectively. Overview of loss (mse) and correlation coefficient for RW test data with 99.9% and 99% missing values and models trained on different masks to simulate missing values. In particular, we compare results for models trained on variable masks with augmentation factor 3 to models trained on optimal masks transferred from CESM and FOCI data, respectively.

Feature	Index	Mask	Missing rate	Loss (mse)	Correlation
SLP	SAM	var., factor: 3	99.9% 99%	0.502 0.098	0.749 0.951
		opt. from CESM		0.230 0.050	0.885 0.975
		opt. from FOCI		0.288 0.048	0.856 0.976
SLP	NAO	var., factor: 3	99.9% 99%	2.171 0.677	0.573 0.883
		opt. from CESM		1.568 0.400	0.691 0.931
		opt. from FOCI		1.404 0.446	0.726 0.923
SST	AMO	var., factor: 3	99.9% 99%	0.031 0.018	0.180 0.675
		opt. from CESM		0.018 0.007	0.673 0.895
		opt. from FOCI		0.025 0.007	0.483 0.897

E. A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

different samples can have different sets of missing data plus the rate of missing values can vary over the full range [0,1]. This flexibility allowed us to find most relevant grid points for each sample by using a bottom-up sampling strategy. We successively added and fixed grid points that lead to largest drop in reconstruction loss and assigned the relative loss reduction to the corresponding grid points. Results vary for different samples. Therefore, we averaged over multiple training samples to obtain stable MRLRMs for both features (SLP and SST) and ESMs (CESM and FOCI). MRLRMs have been visualized as heat maps and give an intuitive understanding of grid points' relevance for successful reconstruction. Our MRLRM methodology allows to look inside our U-Nets and adds a new method of explainable artificial intelligence (xAI). For SLP CESM and FOCI data, we find highest values in MRLRMs at high and low latitudes. Whereas, for SST CESM and FOCI, we find concise spots of high values mostly at coastlines of the Northern Pacific and in the Niño 3.4 region, respectively. The location of dominant spots with high values in MRLRMs clearly differs for SST CESM and FOCI. Understanding these differences in detail is a matter of ongoing research and could help to reveal artifacts and biases for certain ESMs.

From the MRLRMs, we then derived optimal masks for distinct rates of missing values in the ultra sparse regime with 99.9%, 99% and 95% missing data. As a proof of concept, we transferred the obtained masks to RW data. For SLP and SST RW data, we have 10,368 and 9,913 valid grid points, respectively. Thus, for SLP RW data we aimed to find the most relevant 10, 104 and 518 grid points, whereas for SST RW data, we looked for the most relevant 10, 99 and 496 grid points, representing 0.1%, 1% and 5% of valid grid points, respectively. Assuming that we only have sparse measurements in practice, the derived masks help to answer the question, how to get most information from a limited number of survey stations and tells us *where* we should measure.

Although we refer to the masks as *optimal* masks, there is no proof for that optimality statement. For instance, finding the optimal 10 grid points from a total number of 10,368 valid grid points, gives us

$$\binom{10,368}{10} \approx 3.9 \cdot 10^{33}$$

possible combinations. Opposed to that, our methods of successively adding and fixing grid points, only requires to search through

$$\sum_{g=1..10} (10,368 - g + 1) \approx 10^5$$

possibilities. This keeps computational effort manageable and still leads to clear outperformance over models trained on fixed and variable masks. We showed that our models trained on the optimal masks succeed in reconstructing dominant large scale structures from only 0.1% and 1% of the original data, whereas all other models fail on this task, including DINEOF. We also found that the reconstructed geospatial fields from models trained on optimal masks

reflect the large scale dynamics, as shown for several climate indices obtained from reconstructed versus complete data.

As next steps, we plan to extend our method of identifying most relevant grid points to further geospatial data like e.g., geopotential height, surface air temperature, sea surface salinity or precipitation. We then aim to go beyond reconstructing missing information and will try to predict *future* geospatial fields from (ultra) sparse inputs. This requires additional information in form of multivariate and/or time lagged input features as multiple channels for our U-Net models.

Data Availability Statement

Our framework for reconstructing missing data is hosted on GitHub: https://github.com/MarcoLandtHayen/reconstruct_missing_data. Trained models and all results are stored in a separate Git repository: https://git.geomar.de/marco-landt-hayen/reconstruct_missing_data_results. Observational data used in this work are publicly available [12]. ESM data are stored on Zenodo: <https://doi.org/10.5281/zenodo.7774316>.

References

1. Beckers, J. M., and Rixen, M.: EOF Calculations and Data Filling from Incomplete Oceanographic Datasets, *Journal of Atmospheric and Oceanic Technology*, **20**(12), pp. 1839–1856 (2003)
2. Alvera-Azcárate, A., Barth, A., Rixen, M., and Beckers J.M.: Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature, *Ocean Modelling*, **9**(4), pp. 325–346 (2005) <https://doi.org/10.1016/j.ocemod.2004.08.001>
3. Alvera-Azcárate, A., Barth, A., Beckers J.M., and Weisberg, R.H.: Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields, *Journal of Geophysical Research Oceans*, **112**(C3) (2007) <https://doi.org/10.1029/2006JC003660>
4. Yang, Y.C., Lu, C.Y., Huang, S.J., Yang, T.Z., Chang, Y.C., and Ho, C.R.: On the Reconstruction of Missing Sea Surface Temperature Data from Himawari-8 in Adjacent Waters of Taiwan Using DINEOF Conducted with 25-h Data, *Remote Sensing*, **2022**(14) 2818 (2022) <https://doi.org/10.3390/rs14122818>
5. Chai, X., Gu, H., Li, F., Duan, H., Hu, X., and Lin, K.: Deep learning for irregularly and regularly missing data reconstruction, *Scientific Reports*, **10**, 3302 (2020) <https://doi.org/10.1038/s41598-020-59801-x>
6. Barth, A., Alvera-Azcárate, A., Troupin, C., and Beckers, J.M.: DINCAE 2.0: multivariate convolutional neural network with error estimates to reconstruct sea surface temperature satellite and altimetry observations, *Geoscientific Model Development*, **15**(5), pp. 2183–2196 (2022) <https://doi.org/10.5194/gmd-15-2183-2022>
7. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. and Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, *PLoS ONE*, **10** (2015)

E. A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs

8. Lundberg, S.M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA (2017) <https://doi.org/10.48550/arXiv.1705.07874>
9. Matthes K., Biastoch A., Wahl S., Harlaß J., Martin T., Brücher T., Drews A., Ehlert D., Getzlaff K., Krüger F., Rath W., Scheinert M., Schwarzkopf F.U., Bayr T., Schmidt H. and Park W.: The Flexible Ocean and Climate Infrastructure version 1 (FOCI1): mean state and variability, *Geoscientific Model Development*, **13**(6), pp. 2533–2568 (2020)
10. Hurrell J.W., Holland M.M., Gent P.R., Ghan S., Kay J.E., Kushner P.J., Lamarque J.-F., Large W.G., Lawrence D., Lindsay K., Lipscomb W.H., Long M.C., Mahowald N., Marsh D.R., Neale R.B., Rasch P., Vavrus S., Vertenstein M., Bader D., Collins W.D., Hack J.J., Kiehl J. and Marshall S.: The Community Earth System Model: A framework for collaborative research, *Bulletin of the American Meteorological Society*, **94**, pp. 1339–1360 (2013)
11. Marsh D.R., Mills M.J., Kinnison D.E., Lamarque J.F., Calvo N. and Polvani L.M.: Climate change from 1850 to 2005 simulated in CESM1(WACCM), *Journal of Climate*, **26**(19), pp. 7372–7391 (2013)
12. National Oceanic and Atmospheric Administration download center, https://downloads.psl.noaa.gov/Datasets/ncep_reanalysis_derived/surface/pres_sfc_mon_mean.nc, and https://downloads.psl.noaa.gov/Datasets/noaa_ersst.v5/sst_mnmean.nc
13. Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, **9351**, pp. 234–241 (2015) https://doi.org/10.1007/978-3-319-24574-4_28
14. Goodfellow, I., Bengio, Y. and Courville, A.: Deep Learning, MIT Press, <http://www.deeplearningbook.org> (2016)
15. He, K., Zhang, X., Ren, S. and Jian, S.: Deep residual learning for image recognition, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770–778 (2016) <https://doi.org/10.1109/CVPR.2016.90>
16. Kingma, D.P. and Ba, J.: Adam: A Method for Stochastic Optimization, In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA (2014) <https://arxiv.org/abs/1412.6980>
17. Reynolds, D.A., Gaussian Mixture Models, *Encyclopedia of Biometrics*, pp. 827–832 (2009) https://doi.org/10.1007/978-1-4899-7488-4_196
18. Gong D. and Wang S.: Definition of Antarctic oscillation index, *Geophysical Research Letters*, **26**(4), pp. 459–462 (1999)
19. Hurrell, J.W.: Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation, *Science*, **269**(5224), pp. 676–679 (1995)
20. Schlesinger M.E. and Ramankutty N.: An oscillation in the global climate system of period 65–70 years, *Nature*, **367**, pp. 723–726 (1994)
21. Philander S.G.: El Niño, La Niña, and the Southern Oscillation, Academic Press (1989)
22. Trenberth, K.E. and Shea, D.J.: Atlantic hurricanes and natural variability in 2005, *Geophysical Research Letters*, **33**(12) (2006)
23. Morrow R., Ward M.L., Hogg A.McC. and Pasquet S.: Eddy response to Southern Ocean climate modes, *Journal of Geophysical Research* **115**(C10) (2010) <https://doi.org/10.1029/2009JC005894>

Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events

Marco Landt-Hayen, Peer Kröger, Willi Rath, Martin Claus

Accepted for Publication at IEEE eScience 2023

F. Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events

Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events

Marco Landt-Hayen

Ocean Circulation and Climate Dynamics
GEOMAR Helmholtz Centre for Ocean Research
Kiel, Germany
mlandt-hayen@geomar.de

Peer Kröger

Information Systems and Data Mining
Christian-Albrechts-Universität zu Kiel
Kiel, Germany
pkr@informatik.uni-kiel.de

Willi Rath

Ocean Circulation and Climate Dynamics
GEOMAR Helmholtz Centre for Ocean Research
Kiel, Germany
wrath@geomar.de

Martin Claus

Faculty of Mathematics and Natural Sciences
Christian-Albrechts-Universität zu Kiel
Kiel, Germany
mclaus@geomar.de

Abstract—Working with observational data in the context of geophysics can be challenging, since we often have to deal with missing data. This requires imputation techniques in pre-processing to obtain data-mining-ready samples. Here, we present a convolutional neural network approach from the domain of deep learning to reconstruct complete information from sparse inputs. As data, we use various two-dimensional geospatial fields. To have consistent data over a sufficiently long time span, we favor to work with output from control simulations of two Earth System Models, namely the Flexible Ocean and Climate Infrastructure and the Community Earth System Model. Our networks can restore complete information from incomplete input samples with varying rates of missing data. Moreover, we apply a bottom-up sampling strategy to identify the most relevant grid points for each input feature. Choosing the optimal subset of grid points allows us to successfully reconstruct current fields and to predict future fields from ultra sparse inputs. As a proof of concept, we predict El Niño Southern Oscillation and rainfall in the African Sahel region from sea surface temperature and precipitation data, respectively. To quantify uncertainty, we compare corresponding climate indices derived from reconstructed versus complete fields.

Index Terms—missing value imputation, geospatial data, convolutional neural networks, predict climate events

I. INTRODUCTION

Geospatial data in the context of ocean and atmosphere are often provided on a two-dimensional latitude-longitude grid. Examples are sea level pressure (SLP) or sea surface temperature (SST). Observational data are obtained e.g., from remote sensing by satellites or from in situ measurements by survey stations, ships and planes. Observations are usually incomplete due to technical, physical or economical reasons. For instance, instrumental errors and malfunctions can lead to missing data. Moreover, infrared radiation is scattered and reflected by clouds leading to missing data in samples of SST obtained by remote sensing. For other data, we only have

sparse measurements due to a restricted number of survey stations or limited funding periods for research missions.

As prerequisite for advanced data mining, we need consistent and complete data. This requires imputation of missing values. Beckers and Rixen [1] introduced DINEOF as a method to infer missing data from oceanographic data series using empirical orthogonal functions. DINEOF has been and still is a widely used technique in the context of geospatial data. To name a few, Alvera-Azcárate et al. [2] applied the technique to SST data with rates of missing data typically in the range of 40 – 80%. DINEOF has also been used for multivariate reconstruction of missing data [3] with a focus on specific regions of interest [4]. The method is powerful, self-consistent and easy to use. But it has its limits in the allowed rates of missing values. For instance, SST samples with cloud coverage exceeding 95% are eliminated in pre-processing, since DINEOF fails in this regime.

In our earlier work [5], we presented encoder-decoder-style convolutional neural networks (CNNs) that can reconstruct missing data from inputs with varying rates of missing values and clearly outperform DINEOF. Moreover, we proposed a bottom-up sampling strategy to identify grid points in geospatial input fields that are most relevant for successful reconstruction. With this optimal subset of grid points, we restored complete information from ultra sparse inputs and set the benchmark in this regime for SLP and SST data. We introduced mean relative loss reduction maps (MRLRMs) to give a visual and intuitive understanding of grid points' relevance for reconstruction.

In this work, we go beyond reconstructing *current* two-dimensional geospatial fields from incomplete input data and try to predict *future* fields. This allows us to predict climate events like e.g., the El Niño Southern Oscillation (ENSO) [6] or rainfall in the African Sahel region [7] several months into the future. To improve the prediction quality, we add further variables. In particular, we add geopotential height at pressure

This work was supported by the Helmholtz School for Marine Data Science (MarDATA) funded by the Helmholtz Association (Grant HIDSS-0005).

TABLE I
 OVERVIEW OF CESM AND FOCI DATA USED THROUGHOUT THIS WORK. THE GRID RESOLUTION DETERMINES THE GRID DIMENSIONS IN LATITUDE AND LONGITUDE. THE NUMBER OF VALID GRID POINTS EXCLUDES PERMANENTLY MISSING DATA, DUE TO LAND MASSES. FINALLY, WE SHOW THE TOTAL NUMBER OF SAMPLES AND THE TIME SPAN IN YEARS.

Source	Feature	Grid resolution	Grid dimensions	Valid grid points	Samples (years)
			$N_{lat} \times N_{lon}$		
CESM	SLP, Z500, SAT, PREC SST, SSS	$1.8^\circ \times 2.5^\circ$	96×144	13,824 8,276	11,988 (999)
FOCI	SLP, Z500, SAT, PREC SST, SSS	$1.8^\circ \times 1.8^\circ$	96×192	18,432 12,949	12,000 (1,000)

level 500 millibar (Z500), surface air temperature (SAT), sea surface salinity (SSS) and total precipitation (PREC). To have consistent data over a sufficiently long time span, we favor to work with output from control simulations of two Earth System Models (ESMs), namely the Flexible Ocean and Climate Infrastructure (FOCI) [8] and the Whole Atmosphere Community Climate Model as extension of the Community Earth System Model (CESM) [9], [10]. Both are coupled, global climate models that provide state-of-the-art computer simulations of the past, present and future states of the Earth system. In particular, we work with SLP, SST, Z500, SAT, SSS and PREC anomaly fields over 1,000 and 999 years obtained from FOCI and CESM, respectively. Our main contributions are as follows:

- We apply our bottom-up sampling strategy to identify most relevant grid points for SLP, SST, Z500, SAT, SSS and PREC anomaly fields from CESM and FOCI data.
- We reconstruct current and predict future SST and PREC data from ultra sparse inputs.
- We compare results for models trained on univariate, multivariate and time-lagged inputs.
- As a proof of concept, we predict ENSO and Sahel rainfall several months into the future from reconstructed data.
- Our framework is provided as open-source project that can be extended and customized to individual needs, e.g., by including further methods or data.

The rest of this work is structured as follows. In Section II we provide a short description of ESM data and outline data pre-processing. An overview of our U-Net models is given in Section III. Moreover, we recap our sampling strategy and relative loss reduction maps in Section III. In Section IV we apply our U-Net models to ESM data. Additionally, we evaluate model performance and compute several climate indices on reconstructed versus complete SST and PREC fields. Discussion of all results and a conclusion are found in Section V.

II. DATA

In this section, we briefly describe ESM data before we outline data pre-processing. We start with monthly mean SLP, SST, Z500, SAT, SSS and PREC fields from FOCI and CESM control runs, respectively. Both simulations were run using pre-industrial external forcing that is representative for the

year 1850. The FOCI pre-industrial control simulation has been initialized from an ocean at rest with a salinity and temperature distribution based on observations approximately from the last 30 years and then ran for 1,500 years. Here, we only use the latter 1,000 years and skip the first 500 years to allow the model to find its equilibrium. The CESM control simulation has been initialized from another multi-centennial pre-industrial control run provided by the core development team of the National Center for Atmospheric Research (NCAR) [11] and is therefore already in equilibrium. FOCI and CESM are based on very different component models (see [8]–[10] for details) with different strengths and weaknesses in simulating various aspects of the global climate. Table I gives an overview of all data. Anomaly fields are obtained from raw data by removing the seasonal cycle. In particular, we subtract the mean over time separately for each month. To simulate missing data, we use different settings:

- **Fixed mask:** Randomly choose a subset of grid points as missing data with a discrete rate of missing values. This subset is then identical for *all* samples.
- **Variable discrete mask:** Only fix the discrete rate of missing values, then, randomly choose subsets of grid points as missing values. Thus, different samples have different subsets of missing values.
- **Variable range mask:** The set of missing values still differs for different samples, as for the variable discrete mask. Additionally, the rate of missing values is randomly drawn from a specified range. Our models allow maximum flexibility. We can set the range from 0 to 100% missing values.
- **Optimal mask:** Once we identified subsets of grid points that are most relevant for successful reconstruction, we create a fixed mask from taking the remaining grid points as missing data. Here, we use optimal masks only for distinct rates of missing values.

Moreover, anomaly fields are scaled to $[0, 1]$ with minimum and maximum values derived from training data. After scaling, missing values are set to *zero*. While SLP, Z500, SAT and PREC data are defined everywhere, we don't have SST and SSS data over land masses. Permanently missing values due to land masses are also set to *zero*. Of all samples, 80% are used for training, 10% serve as validation data for tuning hyperparameters and the remaining 10% are reserved as test data for model evaluation.

F. Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events

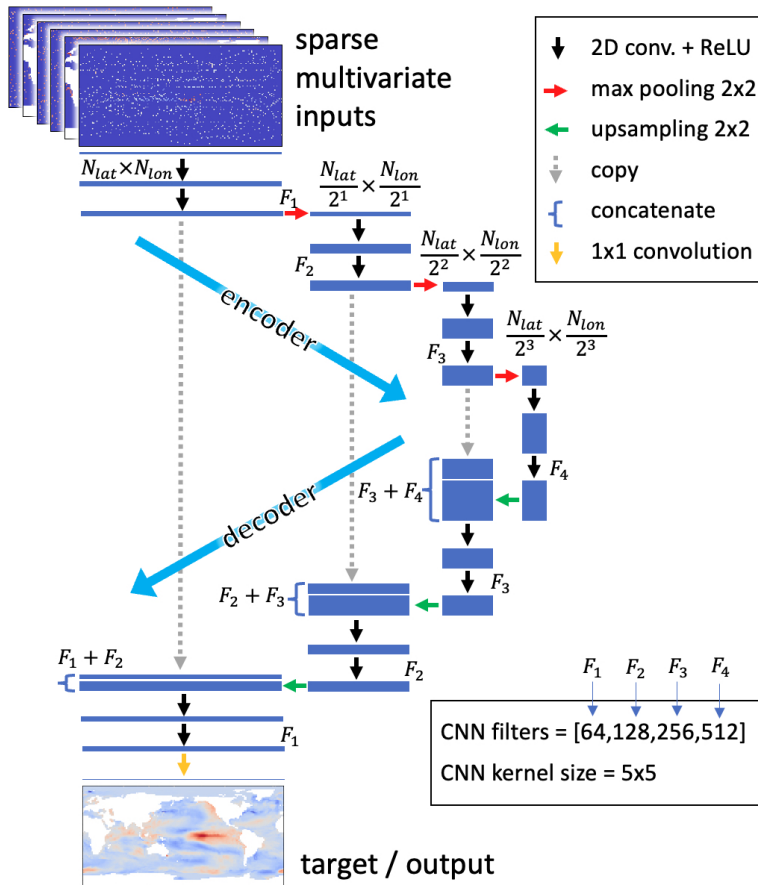


Fig. 1. Encoder-decoder-style U-Net architecture used for reconstructing complete two-dimensional geospatial fields from sparse inputs. In this sketch, we show a multivariate input sample with its 6 input channels (SLP, SST, Z500, SAT, SSS and PREC) with 95% missing values and the corresponding complete SST anomaly field used as target. Blue rectangles symbolize resulting feature maps. Grid dimensions (N_{lat} and N_{lon}) in combination with pooling or upsampling operations determine height and width of feature maps. The depth is specified by the number of CNN filters $F_1..F_4$.

III. MODELS AND METHODS

We give an overview of our U-Net models and the training process in Section III-A. Furthermore, we recap our sampling strategy and relative loss reduction maps in Section III-B.

A. U-Nets

In this work, we focus on encoder-decoder-style CNNs, also referred to as U-Nets [12]. A thorough introduction to deep learning in general and CNNs in particular can be found e.g., in the textbook of Goodfellow et al. [13]. The

U-Net architecture we use throughout this work consists of 4 convolution blocks in the encoder part as shown in Figure 1. Each convolution block in the encoder path consists of two convolution operations with strides set to one and zero padding to conserve height and width dimensions of the resulting feature maps plus a rectified linear unit (ReLU) activation. The convolution blocks are followed by a 2×2 maximum pooling operation to reduce height and width. Maximum depth of 512 feature maps is reached after 4 convolution blocks. In the decoder path we use skip connections to prevent vanishing

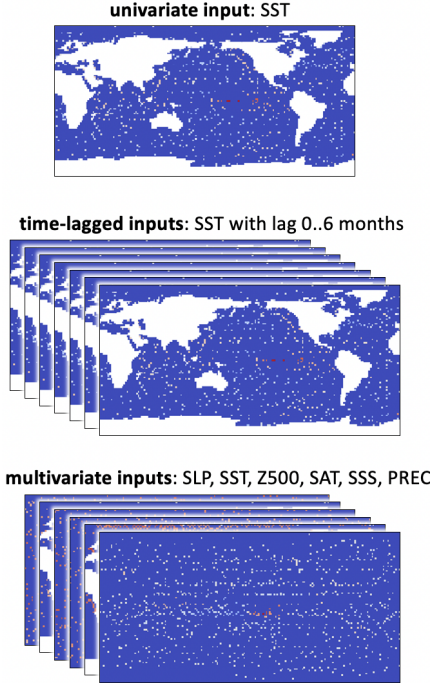


Fig. 2. A sketch of different input types for our U-Net models. A sparse univariate SST input sample is shown in the topmost part. In the middle we show time-lagged SST inputs, combining current (lag 0 months) and historic fields (lag 1..6 months) to have a total number of 7 channels per sample. In the bottom part, we sketch a multivariate input sample as stacked sparse SLP, SST, Z500, SAT, SSS and PREC data to have 6 channels per sample.

gradients [14] and 2×2 upsampling to restore former height and width dimensions. Ultimately, we have a 1×1 convolution operation to obtain a single output channel. The kernel size for convolution operations is set to 5×5 . The number of convolution filters F_i determines the depth of the resulting feature maps in the i -th convolution block. For our U-Net configuration, we have an increasing number of 64, 128, 256 and 512 filters in the four convolution blocks, respectively. As inputs, we use scaled two-dimensional *sparse* anomaly fields. Scaled *complete* anomaly fields serve as targets. We train models on univariate inputs using sparse preprocessed samples as a single input channel. Furthermore, we train models on time-lagged and multivariate inputs where we have multiple input channels. In particular, we combine current fields (lag 0 months) and historic fields (lag 1..6 months) to have a total number of 7 channels per sample for time-lagged inputs for a specific input feature. For multivariate inputs, we stack

sparse SLP, SST, Z500, SAT, SSS and PREC data to have 6 channels per sample. Different types of input data are sketched in Figure 2. The U-Net models are trained over 100 epochs using the ADAM optimizer [15] with a mean squared error (mse) loss function. The learning rate is set to $0.5 \cdot 10^{-4}$. As an exception, models trained to infer SST CESM data require a higher learning of 10^{-4} . The batch size is set to 10. All hyperparameters are the result of a grid search optimization. In the training process, we track validation loss and stop when validation loss reaches its minimum.

B. Sampling Strategy

Assume that we have geospatial input data with a total number v of valid grid points. Furthermore, we assume to have a specific rate $m \in [0, 1]$ of missing values. In a first step, we aim to find the subset G_m^s of the most relevant grid points for a single sample s that are essential to restore complete information. Thus, G_m^s contains $g_m \equiv |G_m^s| = (1 - m) \cdot v$ grid points, where g_m is rounded to the nearest integer number. Moreover, we assume to have a U-Net model trained on samples over the full range of missing values $[0, 1]$ with a variable range mask. We then pick a single sample s and compute the mean state loss (MSL) and the minimum loss (MinL) by using an empty sample of only zeros and the complete sample, respectively, as input for our range model. The difference of MSL and MinL determines the maximum absolute loss reduction (MALR) that we can achieve for a specific sample s , as stated in Equation 1. Superscript s denotes the sample number.

$$MALR^s = MSL^s - MinL^s \quad (1)$$

To find the optimal subset G_m^s of grid points for a single sample, we start with an empty sample and successively add grid points $g = 1..g_m$. The more information we provide to the U-Net, the lower the reconstruction loss. To add grid point g , we choose the one from all $v - g + 1$ remaining valid grid points, that leads to maximum decline in reconstruction loss. The absolute loss reduction (ALR) for adding a grid point g to an input sample s can be translated into relative loss reduction (RLR) as follows:

$$RLR_g^s = ALR_g^s / MALR^s \quad (2)$$

The relative loss reduction scores $\{RLR_g^s\}_{g=1..g_m}$ are assigned to the corresponding grid points and can be visualized as heat map to give an intuitive understanding of grid points' relevance for reconstruction. The subset of most relevant grid points can vary from sample to sample. Here, we average scores over the first and last 120 training samples to obtain a representative mean relative loss reduction map (MRLRM) for a given geospatial feature. To reduce computational effort, we stop adding grid points when the accumulated relative loss reduction exceeds a threshold t . Here, we set $t = 0.9$ as a tradeoff between computation time and information gain. Finally, we derive optimal masks $G_{m=0.999}$, $G_{m=0.99}$ and $G_{m=0.95}$ that contain most relevant grid points for 99.9%,

F. Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events

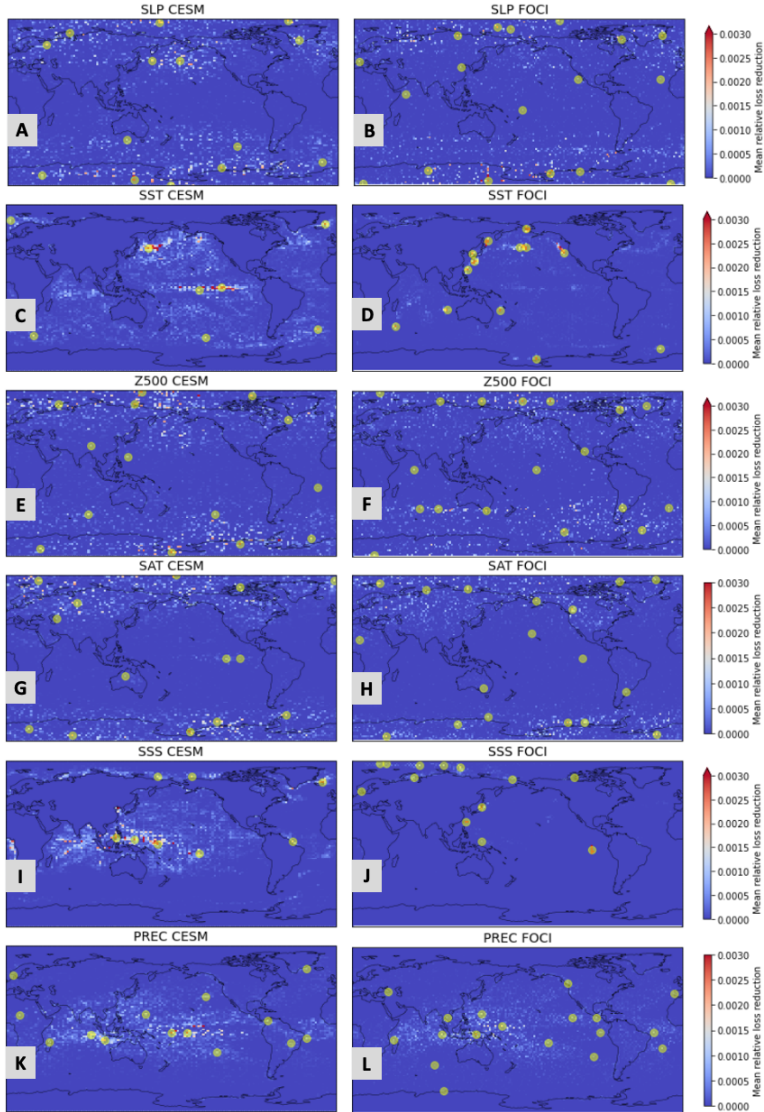


Fig. 3. Individual MRLMs for all variables obtained from CESM (left part) and FOCI (right part) data. All maps are averaged over the first and last 120 training samples. The most relevant grid points obtained from GMM are highlighted as yellow dots. These grid points represent 0.1% of valid grid points and hence, form the optimal masks for 99.9% missing values for the respective feature and source. A and B: SLP. C and D: SST. E and F: Z500. G and H: SAT. I and J: SSS. K and L: PREC.

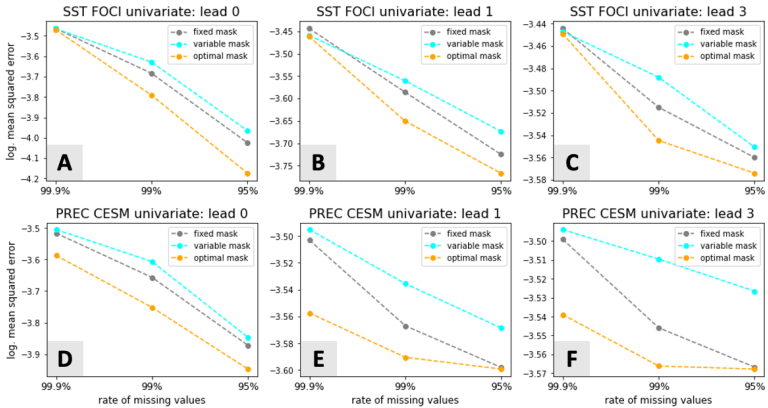


Fig. 4. Loss (mse) for test data on a logarithmic scale for U-Net models trained on SST FOCI (upper part) and PREC CESM (lower part) data with various types of missing masks, distinct rates of missing values and different lead times for corresponding targets. A and D: No lead time (lead 0). B and E: One month into the future (lead 1). C and F: Three months into the future (lead 3).

99% and 95% missing values, respectively, for a given feature and source. Therefore, we fit a set of bivariate Gaussians to each MRLRM using Gaussian Mixture Models (GMMs) [16] and take the grid points with highest RLR as cluster representatives. To summarize the algorithm for deriving optimal masks:

- **Step 1:** Train U-Net with variable range mask on specific input feature.
- **Step 2:** Compute relative loss reduction maps (RLRMs) for a representative subset of training samples.
- **Step 3:** Get MRLRM by averaging over RLRMs obtained in previous step.
- **Step 4:** Derive optimal masks for desired rates of missing values using GMMs.

IV. APPLICATION AND RESULTS

In Section III-B, we recapped the methodology to derive subsets of most relevant grid points from MRLRMs for various rates of missing values. In Section IV-A, we apply this methodology and present individual MRLRMs for all input features. We then apply our U-Net models to univariate, multivariate and time-lagged ESM data in Section IV-B to reconstruct current and future geospatial fields from sparse inputs. Additionally, we evaluate model performance and compute several climate indices on reconstructed versus complete SST and PREC fields in Section IV-C.

A. Derive Optimal Masks

Figure 3 shows MRLRMs for ESM data for all variables and both sources, CESM and FOCI. From these MRLRMs, we derive individual optimal masks for each feature. Therefore, we need to consider the number of valid grid points for the

respective feature and source from Table I. Here, we focus on the ultra sparse regime with 99.9%, 99% and 95% missing values. Accordingly, we aim to find the most relevant 0.1%, 1% and 5% of valid grid points, respectively. As an example, we highlight resulting cluster representatives from GMM for 99.9% missing data in Figure 3.

B. Reconstruction with U-Nets

In Section II, we introduced different mask types to simulate missing data. In this section, we show results for U-Net models trained on univariate CESM and FOCI data with fixed, variable and optimal masks, respectively, for distinct rates of missing values $m \in \{0.999, 0.99, 0.95\}$. As a first step, we try to reconstruct current SST and PREC anomaly fields with no lead time and we try to predict future fields with lead times one and three months, respectively, from the corresponding univariate sparse inputs. The main interest of this work lies in the attempt to optimally reconstruct information from ultra sparse inputs. In Figure 4, we show results for SST FOCI and PREC CESM. Similar results can be obtained for SST CESM and PREC FOCI. From inspecting loss on test data, best performance for univariate inputs is found for models trained on optimal masks for both, SST and PREC.

In a second step, we focus on best performing models trained on optimal masks and try to improve the reconstruction quality by adding additional input data in form of multiple channels. In particular, we train U-Net models on time-lagged inputs adding historic information over the past six months for the respective input feature (SST or PREC). Thereby, we simulate missing values using identical optimal masks for both, current and historic fields. Ultimately, we train U-Net models on multivariate inputs where we simultaneously use current sparse fields of *all* variables as multichannel input

F. Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events

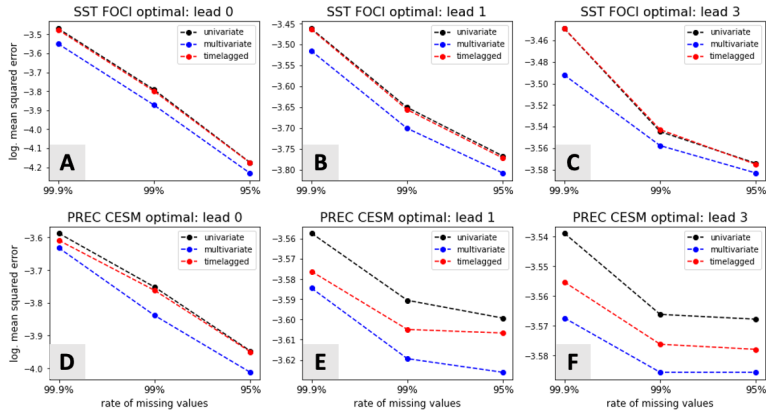


Fig. 5. Loss (mse) for test data on a logarithmic scale for U-Net models trained on SST FOCI (upper part) and PREC CESM (lower part) data with optimal masks applied to various input types, distinct rates of missing values and different lead times for corresponding targets. A and D: No lead time (lead 0). B and E: One month into the future (lead 1). C and F: Three months into the future (lead 3).

data. Thereby, we apply individual optimal masks for each feature. In Figure 5, we show results for SST FOCI and PREC CESM. Similar results can again be obtained for SST CESM and PREC FOCI. From inspecting results, we find best results for models trained on multivariate inputs.

To get an impression of the reconstruction quality, we show the reconstructions for a single SST FOCI test sample obtained from the best performing models trained on multivariate inputs with optimal masks in Figure 6. In particular, we show the reconstruction from models trained on targets without lead time compared to models trained on targets looking three months into the future for distinct rates of missing values $m \in \{0.999, 0.99, 0.95\}$.

The selected target field shows a dominant spot of positive SST anomaly (red) in the Tropical Pacific which is typical for El Niño as the positive phase of ENSO. The strength of the positive anomaly is better matched in reconstructions from models trained to reconstruct current samples without lead time. The shape and the extent of the positive SST anomaly is best restored from an input sample with 95% missing data.

C. Model Evaluation

In this section, we evaluate model performance. As a proof of concept, we aim to predict certain climate events from reconstructed geospatial data. To describe the Earth System dynamics in a compressed way, multi-dimensional physical fields can be reduced to specific climate indices that capture the main processes. For instance, ENSO is a complex phenomenon that can be detected as periodic SST fluctuations in the Tropical Pacific. Several indices are defined to compute the current ENSO phase from area-averaged SST anomalies in certain regions. For instance, the Niño 3.4 index defined from the Niño 3.4 region (5°N–5°S, 120°W–170°W) by the

National Oceanic and Atmospheric Administration (NOAA) is often used in the context of ENSO [17]. While ENSO is a large-scale driver of the climate system, other indices aim to capture regional variability in specific features, like the Sahel precipitation index (SPI) [7]. This index measures anomalies of rainfall in the African Sahel region (10°N–20°N, 20°W–10°E). To quantify uncertainty and to evaluate the model performance, we compute the Niño 3.4 index and SPI on reconstructed versus complete SST FOCI and PREC CESM anomaly fields, respectively. In particular, we compare results for best performing models trained on multivariate inputs with optimal masks trained with distinct rates of missing values in the ultra sparse regime and various lead times. In Figures 7 and 8, we show results for the Niño 3.4 index and SPI on test data. As metric, we use the mse of predicted versus true indices and correlation coefficient r . Clearly, highest fidelity in terms of low mse combined with a high degree of correlation is found for models trained samples with 95% missing data to restore current fields without lead time. The higher the rate of missing data and the further we look into the future, the worse the prediction accuracy. Similar results are obtained for indices computed on SST CESM and PREC FOCI data but not shown, here, for the sake of brevity.

V. DISCUSSION AND CONCLUSION

In this work, we presented encoder-decoder-style U-Net models to restore complete information from sparse geospatial input data. We tried to reconstruct current and to predict future SST and PREC anomaly fields. To have consistent data over a sufficiently long time span, we favored to work with ESM data since observational data from reanalysis already contains reconstructed information from statistical methods.

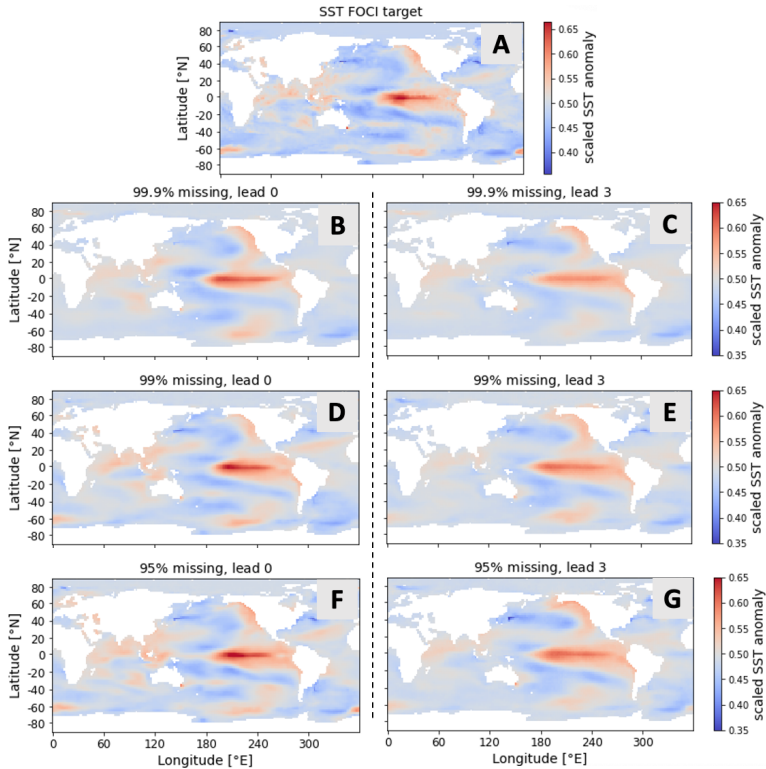


Fig. 6. Reconstruction of a complete SST FOCI test sample used as target (A) from models trained to reconstruct current fields without lead time (lead 0, left part) and models trained to predict future fields three months into the future (lead 3, right part). Here, we show results for best performing U-Net models. In particular, we use models trained on multivariate inputs with optimal masks and various rates of missing values. B and C: 99.9% missing, D and E: 99% missing, F and G: 95% missing.

To simulate missing data, we used different types of masks. Additionally, we have permanently missing data for SST and SSS in terms of land masses. All missing values were set to zero. Our U-Nets showed maximum flexibility in a sense, that different samples can have different sets of missing data plus the rate of missing values can vary over the full range $[0,1]$. This flexibility allowed us to find most relevant grid points for each sample by using a bottom-up sampling strategy. We successively added and fixed grid points that lead to largest drop in reconstruction loss and assigned the relative loss reduction to the corresponding grid points. Results vary for different samples. Therefore, we averaged over multiple training samples to obtain stable MRLRMs for all features (SLP, SST, Z500, SAT, SSS and PREC) and both ESMs (CESM and FOCI). MRLRMs have been visualized

as heat maps and give an intuitive understanding of grid points' relevance for successful reconstruction. Our MRLRM methodology allows to look inside our U-Nets and adds a new method of explainable artificial intelligence (xAI).

For SLP, Z500 and SAT CESM and FOCI data, we find highest values in MRLRMs at high and low latitudes. For SST and SSS CESM, we find regions of high values at coastlines of the Northern and Southern Pacific and in the Niño 3.4 region. Whereas, for SST and SSS FOCI, we find concise spots of high values mostly at coastlines of the Northern Pacific and in the Niño 3.4 region, respectively. For PREC CESM and FOCI, highest values are found around the equator. The location of dominant spots with high values in MRLRMs clearly differs for CESM and FOCI. Understanding these differences in detail is a matter of ongoing research and could help to reveal

F. Reconstruct Geospatial Data from Ultra Sparse Inputs to Predict Climate Events

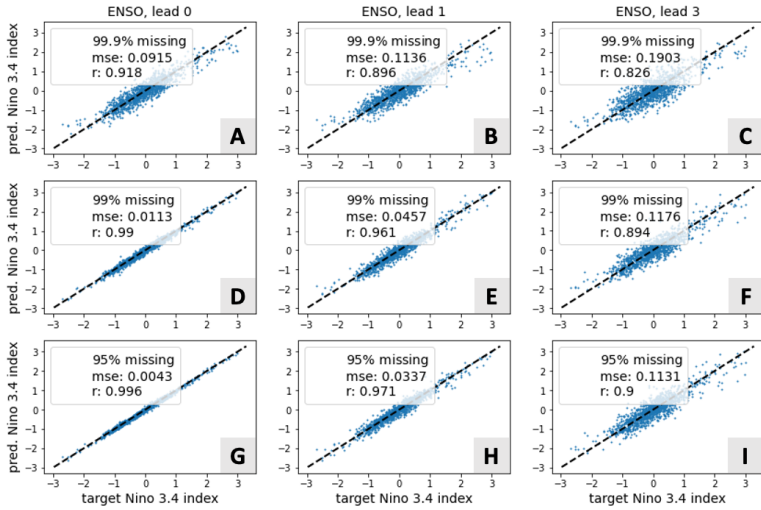


Fig. 7. Compare Niño 3.4 index obtained from reconstructed versus complete SST anomaly fields. Each blue dot represents a FOCI test sample. We compute indices on reconstructions from 0.1% (first row), 1% (second row) and 5% (third row) of all data, hence, with 99.9%, 99% and 95% missing values, respectively. Here, we show results for best performing U-Net models. In particular, we compare models trained on multivariate inputs with optimal masks and targets with various lead times. A, D and G: No lead time (lead 0). B, E and H: One month into the future (lead 1). C, F and I: Three months into the future (lead 3). Loss (mse) and correlation coefficient (r) are stated for each model. The dashed black line is the identity.

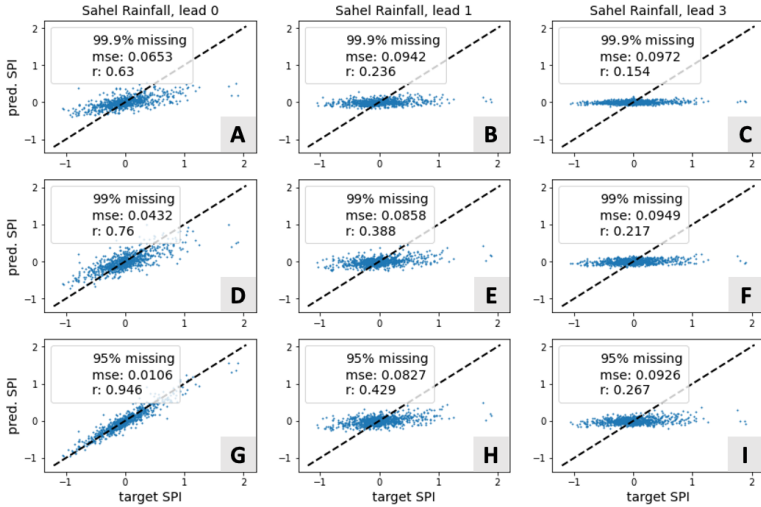


Fig. 8. Compare SPI obtained from reconstructed versus complete PREC anomaly fields. Each blue dot represents a CESM test sample. We compute indices on reconstructions from 0.1% (first row), 1% (second row) and 5% (third row) of all data, hence, with 99.9%, 99% and 95% missing values, respectively. Here, we show results for best performing U-Net models. In particular, we compare models trained on multivariate inputs with optimal masks and targets with various lead times. A, D and G: No lead time (lead 0). B, E and H: One month into the future (lead 1). C, F and I: Three months into the future (lead 3). Loss (mse) and correlation coefficient (r) are stated for each model. The dashed black line is the identity.

artifacts and biases for certain ESMs.

From the MRLRMs, we then derived individual optimal masks for all variables and both ESMs for distinct rates of missing values in the ultra sparse regime with 99.9%, 99% and 95% missing data. Assuming that we only have sparse measurements in practice, the derived masks could help to answer the question, how to get most information from a limited number of survey stations and tells us *where* we should measure.

Although we refer to the masks as *optimal* masks, there is no proof for that optimality statement. For instance, finding the optimal 14 grid points from a total number of 13,824 valid grid points, gives us

$$\binom{13,824}{14} \approx 10^{47}$$

possible combinations. Opposed to that, our methods of successively adding and fixing grid points, only requires to search through

$$\sum_{g=1..14} (13,824 - g + 1) \approx 2 \cdot 10^5$$

possibilities. This keeps computational effort manageable and still leads to clear outperformance over models trained on fixed and variable masks. We showed that our models trained on multivariate inputs with optimal masks succeed in reconstructing and predicting dominant large scale structures from only 0.1%, 1% and 5% of the original data. As a proof of concept, we found that the restored geospatial fields reflect the large scale dynamics and can be used to attribute climate events, as shown for ENSO and Sahel rainfall. In general, the fidelity of predicted indices drops with an increasing rate of missing data and the further we look into the future, as expected. However, predicting ENSO with a lead time of three months still yields reasonably high correlation of predicted versus true index values. Whereas, predicting SPI fails, at least from samples in the ultra sparse regime.

So far, we worked with two-dimensional geospatial data on a global scale and used a rectangular projection as prerequisite for our U-Net models with two-dimensional convolutions. Like this, we pretend to have equal distance of grid points in latitude and longitude directions. However, distances of neighboring grid points in longitude direction depend on latitude. As next steps, we therefore plan to extend our method. One idea is to use a latitude-weighted mse loss for training our U-Net models. Another idea is to try alternative projections for our geospatial input data in combination with spherical convolution or graph neural networks.

DATA AVAILABILITY STATEMENT

Our framework for reconstructing missing data is hosted on GitHub: https://github.com/MarcoLandtHayen/reconstruct_missing_data. Furthermore, we maintain a Docker container providing a Python environment with Jupyter notebooks, Tensorflow and our framework as pre-installed python package. Trained models and all results are stored in a separate

Git repository allowing large file storage: https://git.geomar.de/marco-landt-hayen/reconstruct_missing_data_results. ESM data used in this work are stored on Zenodo: <https://doi.org/10.5281/zenodo.7774316>. Details on the climate indices used for model evaluation can be found at: https://github.com/MarcoLandtHayen/climate_index_collection.

REFERENCES

- [1] Beckers, J. M., and Rixen, M.: EOF Calculations and Data Filling from Incomplete Oceanographic Datasets, *Journal of Atmospheric and Oceanic Technology*, **20**(12), pp. 1839–1856 (2003)
- [2] Alvera-Azcárate, A., Barth, A., Rixen, M., and Beckers, J.M.: Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature, *Ocean Modelling*, **9**(4), pp. 325–346 (2005) <https://doi.org/10.1016/j.ocemod.2004.08.001>
- [3] Alvera-Azcárate, A., Barth, A., Beckers, J.M., and Weisberg, R.H.: Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields, *Journal of Geophysical Research Oceans*, **112**(C3) (2007) <https://doi.org/10.1029/2006JC003660>
- [4] Yang, Y.C., Lu, C.Y., Huang, S.J., Yang, T.Z., Chang, Y.C., and Ho, C.R.: On the Reconstruction of Missing Sea Surface Temperature Data from Himawari-8 in Adjacent Waters of Taiwan Using DINEOF Conducted with 25-h Data, *Remote Sensing*, **2022**(14) 2818 (2022) <https://doi.org/10.3390/rs14122818>
- [5] Landt-Hayen, M., Wölker, Y., Rath, W. and Claus, M.: A Bottom-Up Sampling Strategy for Reconstructing Geospatial Data from Ultra Sparse Inputs, In Proceedings of the 19th International Conference on Advanced Data Mining and Applications (2023)
- [6] Philander, S.G.: El Niño, La Niña, and the Southern Oscillation, Academic Press (1989)
- [7] Badr, H.S., Zaitchik, B.F. and Guikema, S.D.: Application of Statistical Models to the Prediction of Seasonal Rainfall Anomalies over the Sahel, *Journal of Applied Meteorology and Climatology*, **53**(3), 614–636 <https://doi.org/10.1175/JAMC-D-13-0181.1> (2014)
- [8] Matthes, K., Biastoch, A., Wahl, S., Harlaß, J., Martin, T., Brücher, T., Drews, A., Ehlert, D., Getzlaff, K., Krüger, F., Rath, W., Scheinert, M., Schwarzkopf, F.U., Bayr, T., Schmidt, H. and Park, W.: The Flexible Ocean and Climate Infrastructure version 1 (FOCI1): mean state and variability, *Geoscientific Model Development*, **13**(6), pp. 2533–2568 (2020)
- [9] Hurrell, J.W., Holland, M.M., Gent, P.R., Ghan, S., Kay, J.E., Kushner, P.J., Lamarque, J.-F., Large, W.G., Lawrence, D., Lindsay, K., Lipscomb, W.H., Long, M.C., Mahowald, N., Marsh, D.R., Neale, R.B., Rasch, P., Vavrus, S., Verneken, M., Bader, D., Collins, W.D., Hack, J.J., Kiehl, J. and Marshall, S.: The Community Earth System Model: A framework for collaborative research, *Bulletin of the American Meteorological Society*, **94**, pp. 1339–1360 (2013)
- [10] Marsh, D.R., Mills, M.J., Kinnison, D.E., Lamarque, J.F., Calvo, N. and Polvani, L.M.: Climate change from 1850 to 2005 simulated in CESM1(WACCM), *Journal of Climate*, **26**(19), pp. 7372–7391 (2013)
- [11] National Center for Atmospheric Research homepage, <https://ncar.ucar.edu>
- [12] Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, **9351**, pp. 234–241 (2015) https://doi.org/10.1007/978-3-319-24574-4_28
- [13] Goodfellow, I., Bengio, Y. and Courville, A.: Deep Learning, MIT Press, <http://www.deeplearningbook.org> (2016)
- [14] He, K., Zhang, X., Ren, S. and Jian, S.: Deep residual learning for image recognition, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770–778 (2016) <https://doi.org/10.1109/CVPR.2016.90>
- [15] Kingma, D.P. and Ba, J.: Adam: A Method for Stochastic Optimization, In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA (2014) <https://arxiv.org/abs/1412.6980>
- [16] Reynolds, D.A., Gaussian Mixture Models, *Encyclopedia of Biometrics*, pp. 827–832 (2009) https://doi.org/10.1007/978-1-4899-7488-4_196
- [17] National Oceanic and Atmospheric Administration, *Climate Diagnostics Bulletin*, **96**(4) (1996)

