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

CAMBRIDGE
UNIVERSITY PRESS

DATA PAPER

A climate index collection based on model data

Marco Landt-Hayen[1,2], Willi Rath[2], Sebastian Wahl[2], Nils Niebaum[2], Martin Claus[2,3] and Peer Kröger[1]

[1]Information Systems and Data Mining, Christian-Albrechts-Universität zu Kiel, Kiel, Germany
[2]Ocean Circulation and Climate Dynamics, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[3]Faculty 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.

(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:

- 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-piCtl” 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° × 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 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

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

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

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

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.

			Spatial domain	
	Index	Acronym	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

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

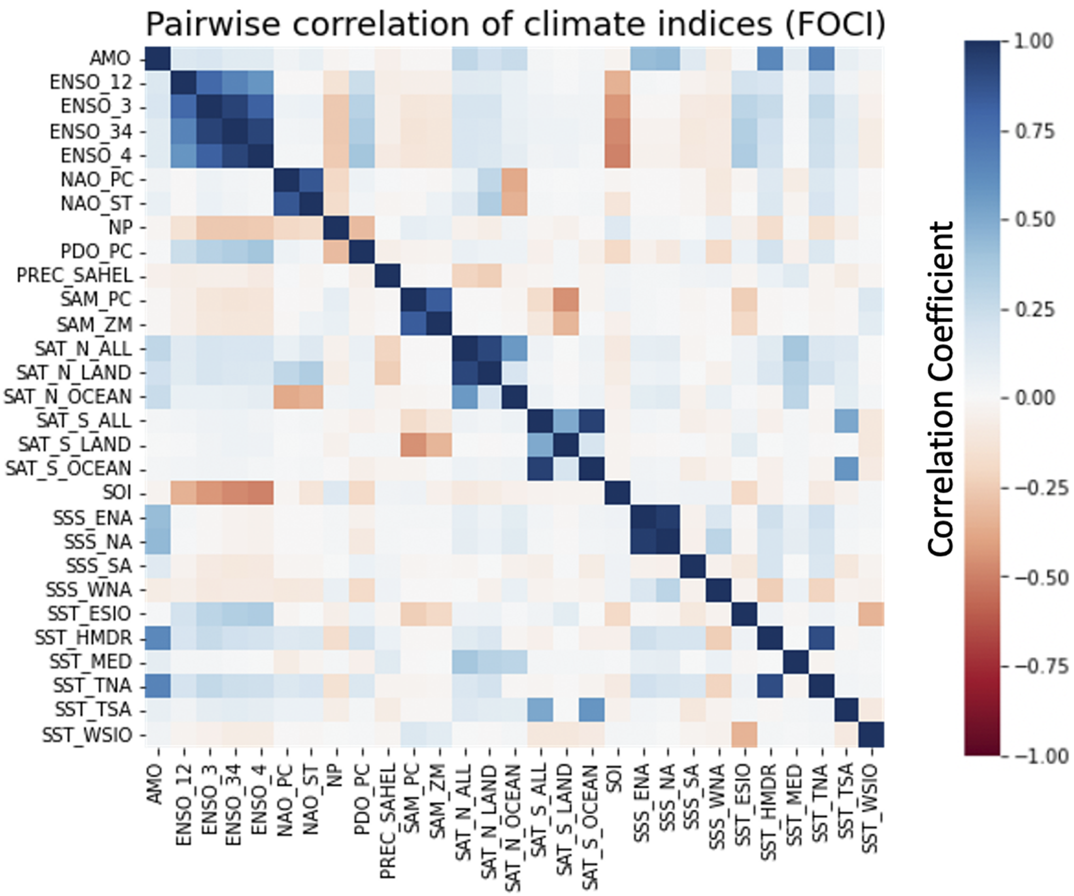

Figure 1. *Pairwise correlation coefficients of all CICMoD indices derived from FOCI data.*

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

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

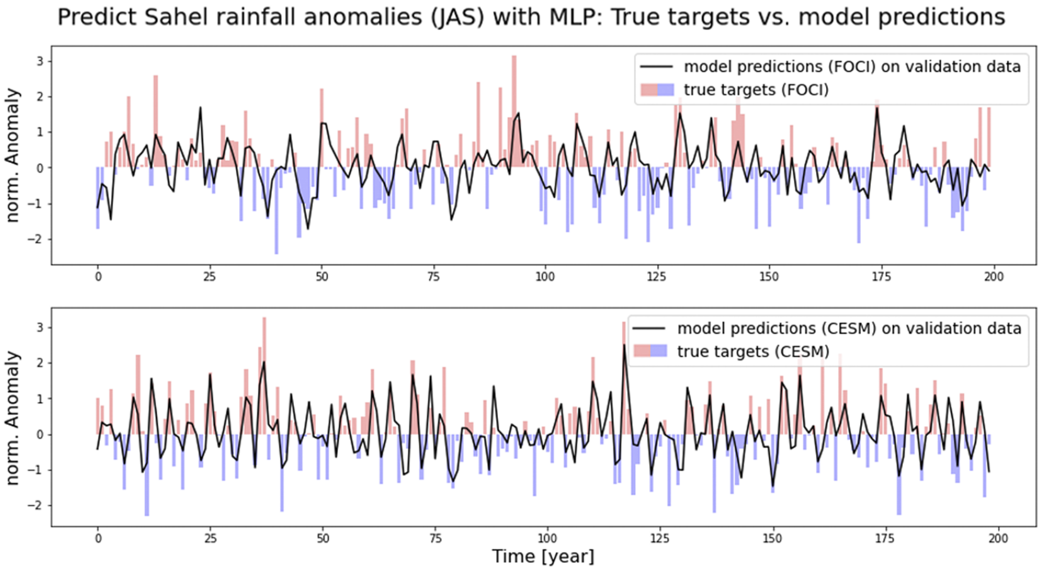

Figure 2. *Fidelity check on validation data: Sahel rainfall predictions (black line) from MLP models on FOCI data (upper part) and CESM data (lower part), respectively, compared to true targets shown as a bar plot.*

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

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

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
$\mathrm{MSE}_{\mathrm{train}}$	0.86	0.88	0.49	0.51
$\mathrm{MSE}_{\mathrm{val}}$	0.83	0.78	0.62	0.60
$\mathrm{Correl}_{\mathrm{train}}$	0.55	0.53	0.70	0.69
$\mathrm{Correl}_{\mathrm{val}}$	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.

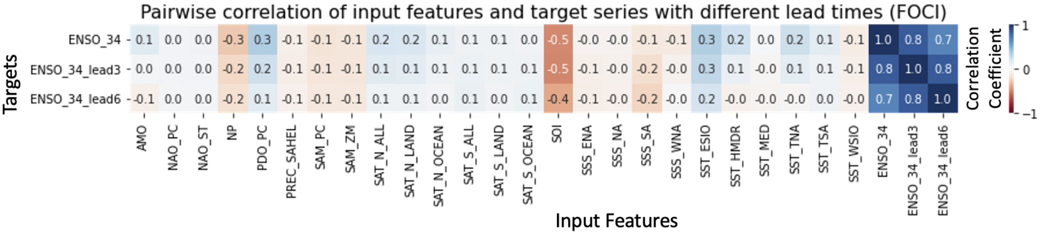

Figure 3. *Pairwise correlation coefficients of Nino 3.4 index with various lead times (current phase, 3 and 6 months into the future) used as targets and input features, both derived from FOCI data.*

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

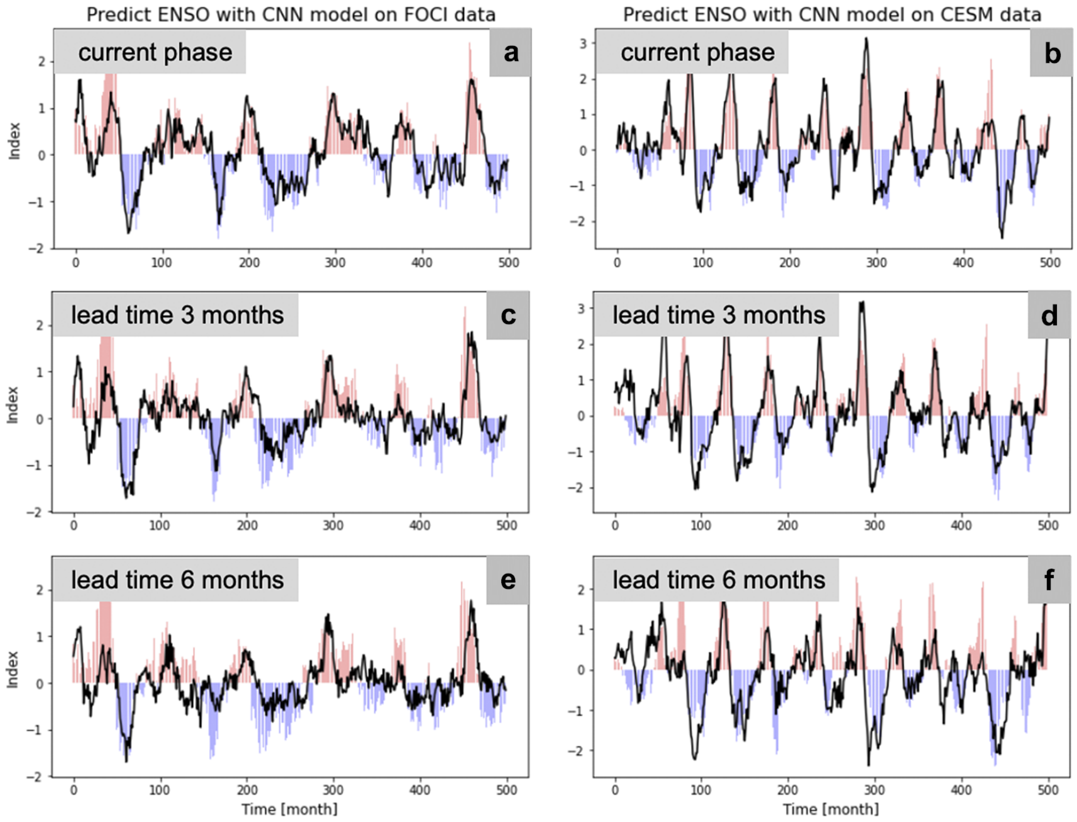

Figure 4. *Fidelity check on the first 500 months of validation data: Compare predictions (black line) from CNN models on FOCI (left-hand side) and CESM data (right-hand side), respectively, compared to true targets shown as bar plot for various lead times. (a,b) Current phase. (c,d) Three months into 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	$\text{MSE}_{\text{train}}$	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
	$\text{Correl}_{\text{train}}$	0.84	0.79	0.72	0.84	0.77	0.71
	$\text{Correl}_{\text{val}}$	0.81	0.68	0.56	0.79	0.64	0.47
CESM	$\text{MSE}_{\text{train}}$	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
	$\text{Correl}_{\text{train}}$	0.88	0.80	0.72	0.88	0.79	0.72
	$\text{Correl}_{\text{val}}$	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.

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 to 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

Abernathey R, **Squire DNT**, **Nicholas T**, **Bourbeau JJG**, **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.

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

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

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 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