


APPLICATION PAPER  

Towards learned emulation of interannual water isotopologue variations in General Circulation Models

Jonathan Wider^{1,2,5,6} , Jakob Kruse^{1,2}, Nils Weitzel^{1,3}, Janica C. Bühler³, Ullrich Köthe² and Kira Rehfeld^{1,3,4}

¹Institut für Umweltp Physik, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany

²Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany

³Department of Geosciences, University of Tübingen, Tübingen, Germany

⁴Department of Physics, University of Tübingen, Tübingen, Germany

⁵Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany

⁶Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany

Corresponding author: Jonathan Wider; Email: jonathan.wider@ufz.de

Received: 20 February 2023; **Revised:** 26 July 2023; **Accepted:** 18 August 2023



Keywords: Climate models; convolutional neural networks; paleoclimate; spherical networks

Abstract

Simulating abundances of stable water isotopologues, that is, molecules differing in their isotopic composition, within climate models allows for comparisons with proxy data and, thus, for testing hypotheses about past climate and validating climate models under varying climatic conditions. However, many models are run without explicitly simulating water isotopologues. We investigate the possibility of replacing the explicit physics-based simulation of oxygen isotopic composition in precipitation using machine learning methods. These methods estimate isotopic composition at each time step for given fields of surface temperature and precipitation amount. We implement convolutional neural networks (CNNs) based on the successful UNet architecture and test whether a spherical network architecture outperforms the naive approach of treating Earth's latitude-longitude grid as a flat image. Conducting a case study on a last millennium run with the iHadCM3 climate model, we find that roughly 40% of the temporal variance in the isotopic composition is explained by the emulations on interannual and monthly timescale, with spatially varying emulation quality. The tested CNNs outperform simple baseline models such as random forest and pixel-wise linear regression substantially. A modified version of the standard UNet architecture for flat images yields results that are as good as the predictions by the spherical CNN. Variations in the implementation of isotopes between climate models likely contribute to an observed deterioration of emulation results when testing on data obtained from different climate models than the one used for training. Future work toward stable water-isotope emulation might focus on achieving robust climate–oxygen isotope relationships or exploring the set of possible predictor variables.

Impact Statement

Information on the hydrological cycle is imprinted onto the isotopic composition of precipitation, which subsequently is preserved in natural climate archives like speleothems or glaciers. Some climate models, so-called isotope-enabled General Circulation Models (iGCMs), simulate isotopes explicitly and, thus, allow

  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.

comparing climate model output under paleoclimate scenarios to samples taken from natural climate archives. However, isotopes are not included in most climate simulations due to computational constraints and the complexity of their implementation. We test the possibility of using machine learning methods to infer the isotopic composition from surface temperature and precipitation amounts, which are standard outputs for a wide range of climate models.

1. Introduction

Reliable analysis of current climate change, as well as robust prediction of future Earth system behavior, has become a crucial foundation for all endeavors to protect humanity's prosperity, mitigate ecological disasters, or formulate plans for adaptation (IPCC, 2023). This analysis hinges on an accurate understanding and modeling of complex mechanisms in the climate system, which in turn relies on knowledge of the system's past behavior. To analyze past climatic conditions outside the comparatively short period of instrumental measurements, we depend on environmental processes recording and preserving information on the climate system in natural "climate archives." One way to recover past climate information from such archives is to measure the relative abundance of isotopes, particularly of the isotopes of the constituents of water molecules (Mook and Rozanski, 2000). Due to differences in mass, molecules with varying isotopic compositions, so-called isotopologues, differ in their behavior in chemical reactions and phase transitions. For the special case of water, molecules containing heavy ^{18}O atoms, in the following denoted heavy isotopes, evaporate slower but condensate faster than ones containing the lighter ^{16}O . These effects are imprinted on the global hydrological cycle. The resulting patterns of the isotopic composition of precipitation depend on many variables such as precipitation amount, temperature, relative humidity, and the circulation of the atmosphere (Dansgaard, 1964). This makes heavy isotopes in water an important tracer of the hydrological cycle and consequently a valuable proxy for past climatic changes.

Isotope abundances are canonically expressed in the delta notation. For stable oxygen isotopes ^{18}O and ^{16}O , this is given by

$$\delta^{18}\text{O} = \left(\frac{[\text{}^{18}\text{O}]_{\text{sample}}}{[\text{}^{16}\text{O}]_{\text{sample}}} / \frac{[\text{}^{18}\text{O}]_{\text{reference}}}{[\text{}^{16}\text{O}]_{\text{reference}}} \right) - 1 [\text{‰}]. \quad (1)$$

Here the ratio of concentrations of the isotopic species in a given sample is compared to a defined reference standard. For $\delta^{18}\text{O}$ of precipitation, this standard is an artificially created sample with an isotopic composition that is typical for ocean surface water (Baertschi, 1976).

One important task in paleoclimatology is to test whether hypotheses about the past climate are compatible with proxy data like $\delta^{18}\text{O}$ measured in natural climate archives (e.g. Bühler et al., 2022). To compare simulations of hypothetical climate states to those measurements, a special sub-type of climate models, so-called isotope-enabled General Circulation Models (iGCMs), was developed. They explicitly simulate isotopic compositions by following the isotopic water species through the hydrological cycle (Yoshimura et al., 2008; Tindall et al., 2009; Colose et al., 2016; Werner et al., 2016; Brady et al., 2019). However, many climate models and climate model simulations exist that do not include information on water isotopologues. Simulating $\delta^{18}\text{O}$ is costly because it typically requires duplicating large parts of the water cycle for each simulated water species (Tindall et al., 2009). In light of recent advances in data science, the question arises whether this isotopic output can instead be emulated using machine learning (ML) models that infer the $\delta^{18}\text{O}$ at each location from other climate variables after a model run is finished. We thus call this approach "offline-emulation." Conducting the emulation "offline," that is, not coupled to the climate simulation, is possible because isotopes are passive tracers of the hydrological cycle that reflect climatic variations, but have no feedback onto the climate system. Exploratory work in this direction has been conducted by Fiorella et al. (2021), who used random forest regression to infer isotope ratios in precipitation. Their study assessed whether and to what extent potential climate effects on the isotopic composition can be verified in data simulated by an isotope-

enabled climate model. There is an important difference between their study and our work: while our ML methods use standard output variables of climate models as inputs, Fiorella et al. (2021) relied on tracers being implemented for the inputs to the random forest regression. While this is a suitable choice for their study design and research question, it limits the utility of their random forest model as an emulator.

Within this study, we narrow the broad task of “offline-emulation” by making a number of choices for the learned isotope emulation. The first choice is to only emulate the isotopic composition of precipitation, neglecting subsequent processes that might disturb the signal until it is stored in a climate archive (see e.g., Casado et al., 2018). Systematic observations of oxygen isotopes in precipitation did not begin until the 1960s (IAEA/WMO, 2020), and data are spatially sparse. A line of research (Bowen and Revenaugh, 2003; Bowen, 2010; Vachon et al., 2010; Terzer et al., 2013) constructs so-called iso-scapes (isotopic landscapes) for $\delta^{18}\text{O}$ from observation data (e.g., IAEA/WMO, 2020). These studies often address climatological rather than meteorological questions (Bowen, 2010), and provide, for instance, multi-year averages of annual and monthly mean $\delta^{18}\text{O}$. In contrast, we exclusively utilize simulation data in our experiments and aim to learn and emulate the relationship between a given atmospheric state and the related spatial distribution of $\delta^{18}\text{O}$.

We limit ourselves to using surface temperature and precipitation amount as the two fundamental predictor variables, since these variables possess strong correlations to $\delta^{18}\text{O}$ that are well known experimentally (Dansgaard, 1964) and from simulations (see Figure 2c) and are frequently simulated in climate models. We decided to emulate yearly $\delta^{18}\text{O}$ data from the last millennium (850 CE to 1849 CE) climate simulations. This is motivated by the combination of the high data availability of simulation runs of sufficient length, and the archiving resolution of paleoclimate records during this time period which is typically between monthly and sub-decadal. We also contrast the yearly emulation results with experiments using monthly resolution.

As a measure of emulator performance, we will use the R^2 score, which measures the fraction of explained temporal variance, as detailed in Section 2.2.5. While we use ML methods that exploit spatial correlations in the data by design, we leave explicit incorporation of temporal correlations largely to future investigation.

Working within these constraints, our article presents the following contributions:

- We train a deep neural network to estimate stable oxygen isotopes in precipitation ($\delta^{18}\text{O}$), given surface temperature and precipitation, and compare to common regression baselines.
- To respect the underlying geometry of the climate model data, we investigate the performance of a spherical network architecture.
- We present cross-model results, where a regressor trained on simulated data from one climate model is used to emulate $\delta^{18}\text{O}$ in a run from a different model.

2. Data and Methodology

Our approach to emulating $\delta^{18}\text{O}$ is sketched in Figure 1. For each time step, we start with variables that we know to be statistically related to $\delta^{18}\text{O}$, namely surface temperature and precipitation amount. All variables are standardized pixel-wise, that is, we subtract the mean and divide by the standard deviation, both calculated on the training set. We then estimate the standardized spatial field of $\delta^{18}\text{O}$ from the predictor variables by training a machine learning (ML) regression model. Subsequently, the standardization is inverted for the inferred $\delta^{18}\text{O}$, resulting in our estimate for the isotopic composition.

2.1. Data

We use data from the isotope-enabled version of the Hadley Center Climate Model version 3 (hereafter iHadCM3, Tindall et al., 2009). iHadCM3 is a fully coupled atmosphere–ocean general circulation model (AOGCM). The horizontal resolution of iHadCM3 is 3.75° in the longitudinal direction, and 2.5° in the latitudinal direction. We exclude -90° and 90° from the latitudinal values because $\delta^{18}\text{O}$ is not simulated at

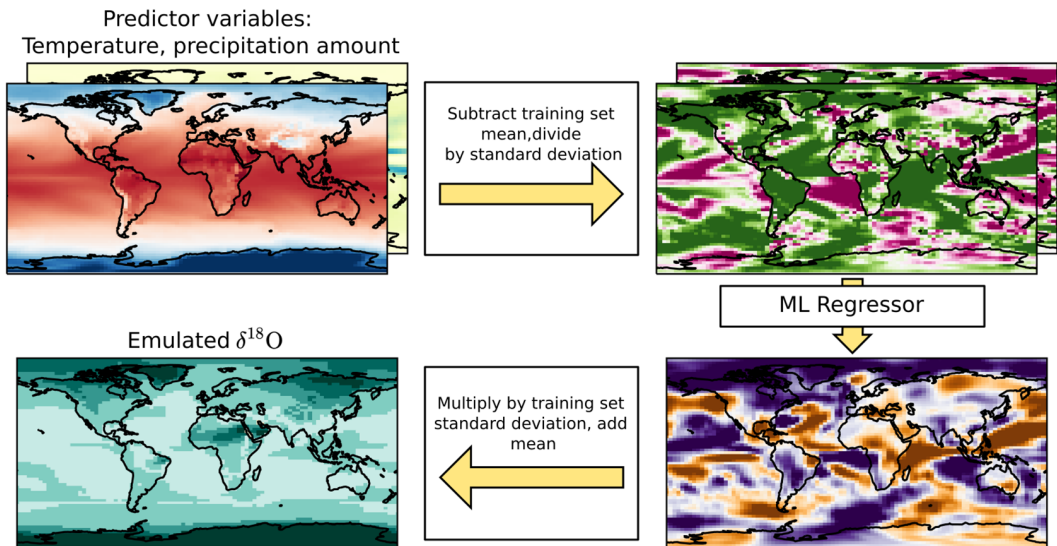


Figure 1. Our approach to the emulation of $\delta^{18}\text{O}$ in precipitation: for each time step, we use surface temperature and precipitation amount as predictor variables. Subsequently, the data is standardized pixel-wise by subtracting the mean and dividing it by the standard deviation at each pixel (top right). Means and standard deviations are based on the training set of the investigated climate model simulation. We use a machine learning emulation model (ML Regressor) to obtain a standardized estimate for $\delta^{18}\text{O}$. The emulator output (bottom right) is then de-standardized using the training set mean and standard deviation of $\delta^{18}\text{O}$ at every pixel, to arrive at the final emulation result (bottom left). When applying the ML model to data from climate models other than the one that was used for training (e.g., in the cross-comparison experiment in Section 3.4), we use the mean and standard deviation from the training set of the new model.

these latitudes. We focus on the last millennium (850 CE to 1849 CE), which is characterized by a stable climate with variability on interannual-to-centennial timescales, but no major trends (Jungclauss et al., 2017). Additionally, the last millennium is well documented in climate archives and observations (Morice et al., 2012; PAGES2k-Consortium, 2019; Comas-Bru et al., 2020; Konecky et al., 2020).

Diagnostics of the iHadCM3 data set are visualized in Figure 2. As can be seen from Figure 2b, the standard deviation of the simulated $\delta^{18}\text{O}$ is large over dry regions like the Sahara desert or the Arabian peninsula. This is partly related to the way $\delta^{18}\text{O}$ is computed in the climate models: in these regions, the abundances of ^{18}O and ^{16}O are both small because of generally low precipitation amounts, leading to numerically unstable ratios and missing values on the monthly time scale. Overall, 0.3% of the $\delta^{18}\text{O}$ values are missing on the monthly timescale, with a strong clustering in the regions with numerical instabilities described above (compare Supplementary Figure A.10). We take this into account by adapting the loss we use to train our ML methods to deal with missing values, as described in Section 2.2.3.

To test the extrapolation and robustness of our emulator, we use last-millennium simulations of three other climate models: Scripps Experimental Climate Prediction Center's Global Spectral Model (hereafter isoGSM, Yoshimura et al., 2008), iCESM version 1.2 (hereafter iCESM, Brady et al., 2019), and ECHAM5/MPI-OM (hereafter ECHAM5-wiso, Werner et al., 2016). While iCESM and ECHAM5-wiso are fully coupled AOGCMs, isoGSM is an atmospheric GCM forced by sea-surface temperatures and sea-ice distributions of a last millennium run with the CCSM4 climate model (Landrum et al., 2013). We re-grid the data of the other climate model simulations ($\delta^{18}\text{O}$, surface temperature, precipitation amount) to the iHadCM3 grid using bilinear interpolation from the CDO tool set (Schulzweida, 2020).

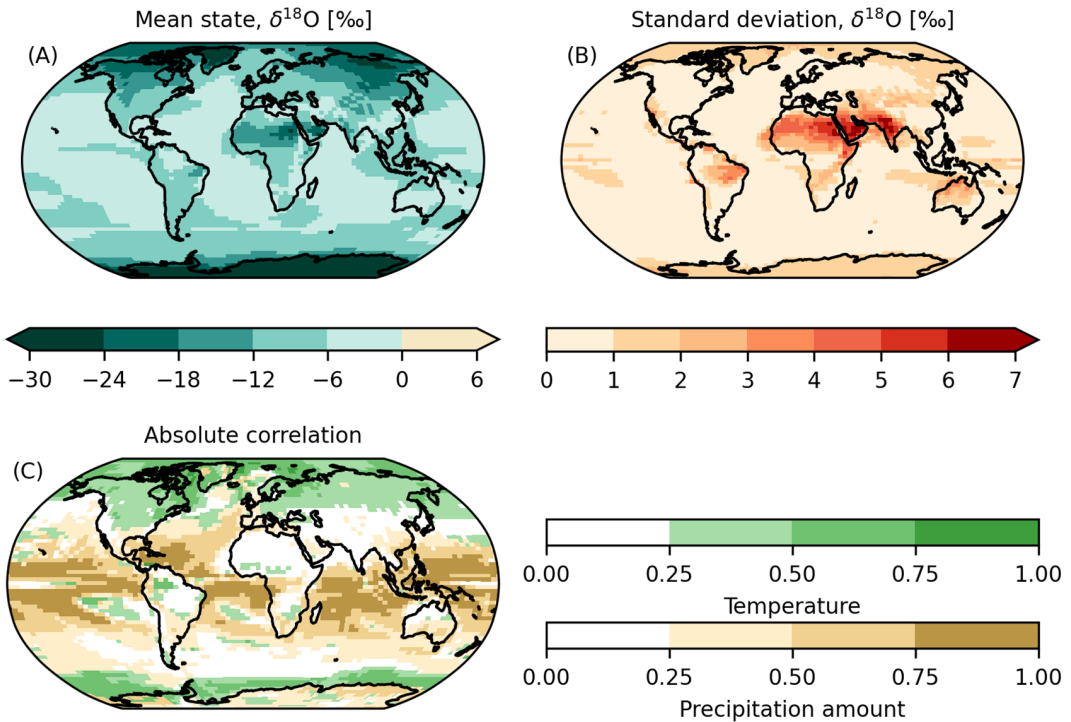


Figure 2. Statistical properties of the iHadCM3 $\delta^{18}\text{O}$ data: (a) mean state of isotopic composition ($\delta^{18}\text{O}$) in precipitation and (b) standard deviation of $\delta^{18}\text{O}$ on an annual timescale; (c) absolute correlations of $\delta^{18}\text{O}$ with temperature (green) and precipitation amount (brown) on interannual timescale—for each grid cell only the stronger of the two is shown.

All datasets are freely available at <https://doi.org/10.5281/zenodo.7516327> and described in detail in Bühler et al. (2022).¹

2.1.1. Pre-processing

We apply the following pre-processing steps to the climate simulation data:

- We set valid ranges for all variables, thereby excluding implausibly large or small values, using the following choices: surface temperature range: [173,373] K, $\delta^{18}\text{O}$ range: [−100,100], precipitation amount: [−1,10000] $\frac{\text{mm}}{\text{month}}$. Wide ranges are chosen because we aim to exclude only implausible values that might deteriorate emulator performance without artificially removing model deficiencies. Thus, we also keep small negative precipitation values that climate models might produce due to numerical inaccuracies in rare occasions.
- Time steps with missing values in the predictor variables are excluded from the dataset. This leads to the exclusion of 31 of the 12,000 monthly time steps of iHadCM3.
- We form yearly averages from monthly data. Missing $\delta^{18}\text{O}$ data points are omitted in the yearly averaging. We argue that this does not impact our results negatively, because the invalid 0.3% of $\delta^{18}\text{O}$ values cluster in regions, where due to numerical instabilities in the “ground truth” iHadCM3 simulation, learning a physically consistent emulation would not have been possible anyway (compare Supplementary Figure A.10).

¹ Bühler et al. (2022) also investigate a fifth climate model, GISS ModelE2-R (Colose et al., 2016), which we excluded from our study because of physically implausible trends in polar regions in the corresponding model run.

- We re-grid the yearly datasets to the irregular grid on which the investigated spherical network operates (see Section 2.2) using a first-order conservative remapping scheme (Schulzweida, 2020).
- We split the data into test and training sets. We use 850–1750 CE for training and 1751–1849 CE for testing. The data are split chronologically instead of randomly to make the test and training set as independent as possible, and prevent the network from exploiting auto-correlations from previous or subsequent time steps. If a validation set (used for making choices of ML hyperparameters) is needed, we split off 10% of the training set randomly unless specified otherwise.
- Before the ML methods are applied, the data are standardized pixel-wise by subtracting the training set mean and dividing by the standard deviation of the corresponding climate model, as visualized in Figure 1.

2.2. Methodology

To obtain a spatially consistent emulation, and to utilize the fact that the local statistical relations between $\delta^{18}\text{O}$ and the predictor variables are similar in many grid boxes on the Earth's surface, we choose two approaches based on convolutional neural networks (CNNs). Both utilize the successful UNet architecture (Ronneberger et al., 2015), whose multi-scale analysis can simultaneously capture fine structure variations and utilize large-scale contextual information. UNet architectures have been successfully applied in a climate science context before (e.g., Kadow et al., 2020). The first of our two approaches treats data on the latitude-longitude grid as a flat image. The second explicitly incorporates the spherical geometry of the data.

2.2.1. Flat network

Because our data naturally lie on the surface of a sphere, distortions arise when treating the equally spaced longitude-latitude grid as a flat image using, for example, a plate carrée projection (lat/lon projection). We test if we can still obtain reasonable results with this naive setup. Furthermore, we try to partially remedy the effects of the distortions within the “flat” approach, by modifying the standard UNet architecture in three ways:

- We use area-weighted loss functions.
- We use periodic padding in the longitudinal direction, that is, we append the rightmost column to the very left of the plate carrée map (and vice versa) before computing convolutions. Thereby, we assure continuity along the 0° – 360° coordinate discontinuity.
- We incorporate CoordConv (Liu et al., 2018), a tweak to convolutional layers that appends the coordinates to the features input into each convolution, thus allowing networks to learn to break translational symmetry if necessary.

2.2.2. Spherical network

As a more sophisticated technique, a multitude of approaches to directly incorporate the spherical nature of data into a neural network architecture has been proposed (Cohen et al., 2018; Coors et al., 2018; Cohen et al., 2019; Defferrard et al., 2020; Esteves et al., 2020; Lam et al., 2022). We reproduce the approach of Cohen et al. (2019), where the network operates on an icosahedral grid, with grid boxes centered on the vertices. Using the icosahedron offers a straightforward way to increase or decrease resolution for a UNet-like design, as we can recursively subdivide each of its triangles into four smaller triangles, projecting all newly created vertices onto the sphere again. We denote the number of recursive refinements of the grid as r , with $r = 0$ identifying the grid containing only the 12 vertices of the regular icosahedron. As the refined icosahedral grid is locally very similar to a flat hexagonal grid, we can use an appropriately adapted implementation of the usual efficient way to compute convolutions. Additionally, the architecture of Cohen et al. (2019) is equivariant to a group of symmetry transformations, meaning that if the input to the CNN is transformed by an element of the symmetry group, the output transforms accordingly. This fits well with the approximate symmetries present in the Earth system, like symmetry to reflections on the

equatorial plane or rotations around the polar axis. We validate our implementation of the method on a toy problem described by Cohen et al. (2019): the classification of handwritten digits projected onto a spherical surface. We obtain results that are comparable to those reported by Cohen et al. (2019); see Supplementary Appendix A.1.1 for more details.

2.2.3. Loss function

To train our UNet architectures for isotope emulation, we use a weighted mean squared error loss between the standardized $\delta^{18}\text{O}$ ground truth Y and the predicted values \hat{Y} :

$$L(Y, \hat{Y}) = \frac{1}{b} \sum_{i=1}^b \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i} w_j (Y_{ij} - \hat{Y}_{ij})^2, \tag{2}$$

where the loss is averaged over a batch of size b and the set of valid grid boxes \mathcal{G}_i at time step i . A grid box is valid if the simulated ground truth data has no missing value at this time step in this grid box. $|\mathcal{G}_i|$ denotes the cardinality of \mathcal{G}_i , and w_j are weighting coefficients. For the convolutional UNet working on the plate carrée projection, we choose w_j to be proportional to the cosine of the latitude of the center of grid cell j , which is an approximation of the area of the grid cell. We rescale the weights, such that they sum to the total number of grid boxes. For the icosahedral UNet, all grid boxes are of approximately equal size. Therefore, no weighting is applied and w_j is a constant independent of j .

2.2.4. Baselines

In addition to the UNet models, we implement three simple baseline models to assess the relative benefit of complex and deep models in our emulation problem. These baselines are as follows:

- **Grid-box-wise linear regression**, the simplest conceivable model: regress $\delta^{18}\text{O}$ on temperature and precipitation amount in a separate model for each grid box.
- **Grid-box-wise random forest regression model**: in contrast to the linear regression baseline, we train a single random forest (Breiman, 2001) to make predictions on all grid boxes. To allow the model to learn spatially varying relationships, we include the coordinates as predictor variables.²
- **Grid-to-grid approach (PCA regression)**: relations between $\delta^{18}\text{O}$ and other climatic variables tend to behave similarly over large areas (see Figure 2c), justifying a dimension reduction of the input and output spaces before applying a multivariate linear regression. This is implemented by computing the principal components of the input and output spaces. Schematically, the computation goes as follows: $X \xrightarrow{\text{PCA}_X} C_X \xrightarrow{\text{lin.reg.}} \hat{C}_Y \xrightarrow{\text{PCA}_Y^{-1}} \hat{Y}$. Approximately optimal numbers of principal components are obtained as follows: we iterate over a 50×50 logarithmically spaced grid of candidate values for the number of input and output principal components. For each configuration, the emulation model is trained and its performance is measured on a held-out validation set. We then select the combination of numbers of input and output principal components which yields the best results on the validation set. As a last step, the selected model is retrained, now including the validation set data. Principal component analysis can be performed on arbitrary grids, which makes it equally applicable to the projected 2D data and the icosahedral representation.

2.2.5. Metrics

The metric we use for evaluating emulation approaches is the R^2 score, also called the ‘‘coefficient of determination,’’ which quantifies what fraction of the temporal variance in the test set is explained by the

² We encode each longitude ϕ as two values $\sin(\phi)$ and $\cos(\phi)$ to avoid the discontinuity at $0^\circ/360^\circ$.

ML estimate in each grid box. The R^2 score compares the $\delta^{18}\text{O}$ ground truth Y_j and an estimate \hat{Y}_j in a given grid box j as

$$R^2(Y_j, \hat{Y}_j) = 1 - \frac{\text{MSE}(Y_j, \hat{Y}_j)}{\sigma_j^2}, \quad (3)$$

where $\text{MSE}(Y_j, \hat{Y}_j)$ is the mean squared error and σ_j^2 the variance of the test set ground truth, both taken over the time axis at grid box j . A value of $R^2 = 1$ indicates perfect emulation. $R^2 = 0$ can, for instance, be obtained by a trivial baseline model that returns the true temporal mean at every time step. The score can become arbitrarily negative.

Additionally, we compute the Pearson correlation coefficient between the true and emulated time series at selected grid boxes. To choose time steps in which a method's performance is particularly strong or weak, we calculate the anomaly correlation coefficient (ACC) between emulation and ground truth. ACC is defined as the Pearson correlation coefficient between the true and emulated anomaly patterns for a given time step. Anomalies are computed with respect to the training set mean.

If error intervals on performance metrics are given, they are 1σ intervals computed over a set of 10 runs, unless stated otherwise. Thus, the uncertainties only account for the uncertainty of the stochastic aspects of the ML model parameter optimization, disregarding any uncertainty that is related to the data.

Implementation details for training and configuration of the ML methods are provided in [Supplementary Appendix A.1](#), and the code to reproduce our experiments is freely available at <https://github.com/paleovar/isoEm/releases/v1.0>.

3. Results

We structure the Results section as follows. First, we give a detailed spatiotemporal overview to illustrate the characteristics of the ML-based emulation results. To this purpose, we use the best-performing emulation method as an example. Subsequently, we compare emulation methods amongst each other, contrasting deep architectures and baselines as well as “flat” and “spherical” approaches. We follow up with a range of sensitivity experiments and conclude by conducting a cross-model experiment, that is, we train an ML model on data from one climate model and then use the trained model to emulate $\delta^{18}\text{O}$ in other climate model simulations.

3.1. Spatiotemporal overview of emulation results

In [Section 3.3](#), we will discover that the best-performing ML emulation method, a deeper version of the flat UNet architecture, reaches an average R^2 score of 0.389 ± 0.006 on the plate carrée grid. This means that in the global average, almost 40% of the temporal variance in the test set is explained by our emulation on the interannual timescale. We use this best ML method to introduce spatial and temporal characteristics of the emulation.

The prediction quality varies spatially, as shown in [Figure 3a](#). R^2 scores of 0.6 or larger are reached in 18.5% of the grid cells, and $R^2 \leq 0$ for only 5.4% of grid cells. The best results are achieved over tropical oceans, which are regions with strong correlations of $\delta^{18}\text{O}$ and precipitation amounts. Performance is good over large parts of the Arctic and over western Antarctica as well, which is important because these regions are especially relevant for the comparison with $\delta^{18}\text{O}$ measurements from ice cores. We illustrate the performance in these regions by comparing emulated and ground truth time series in the grid boxes closest to two ice core drilling sites in panels b and c of [Figure 3](#): the North Greenland Ice Core Project (“NGRIP,” 75.1° N, 42.3° W, North Greenland Ice Core Project Members, 2004) and the West Antarctic Ice Sheet Divide ice core project (“WAIS Divide,” 79.5° S, 112.1° W, Buizert et al., 2015). For these drilling sites, the correlation between our emulation and the exact output time series of the isotope-enabled climate model exceeds 0.7.

In general, spatial variations in performance follow the correlation structure between $\delta^{18}\text{O}$ and the predictor variables ([Figure 2c](#)): in regions with strong absolute correlations between $\delta^{18}\text{O}$ and surface

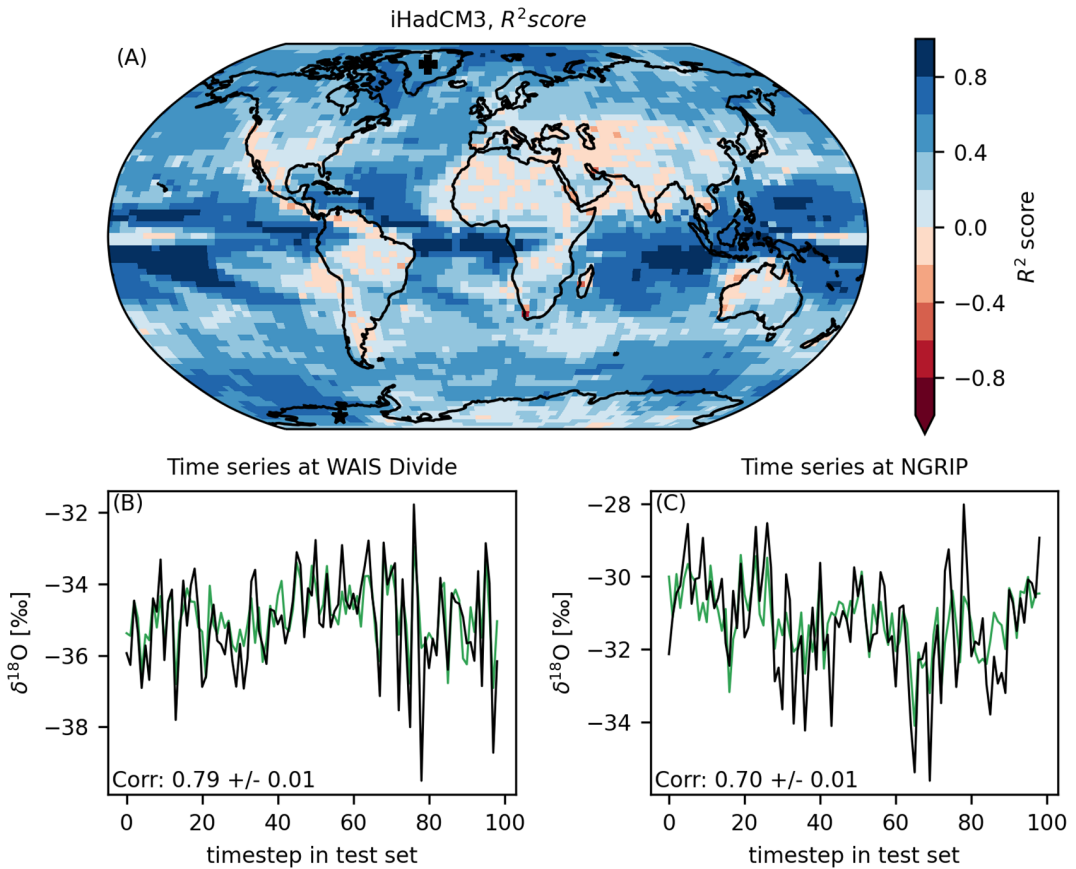


Figure 3. Test set emulation performance of the best ML emulation method. The bluer the colors, the better the emulation. Blue colors indicate regions in which the performance is better than a trivial baseline model that returns the correct test set mean at every time step. This plot displays the average of the R^2 scores over 10 runs. Additionally, we show the time series of the ML emulation (green, mean over ten runs) and the true simulation data (black) for grid boxes next to two ice core drilling sites. Panel (b) “NGRIP” (Greenland). Panel (c) “WAIS Divide” (West Antarctica).

temperature or precipitation amount, the R^2 scores are higher than in regions where none of the predictor variables is strongly correlated with $\delta^{18}\text{O}$. Thus, performance is worse over landmasses, especially in the low and mid-latitudes. Next, we visualize emulation and climate model output for individual time steps. For a year with typical emulator performance³, we plot emulated (panel a) and simulated (panel b) anomalies in Figure 4. We can see that the large-scale patterns match well between emulation and simulation: there are strong positive anomalies over the Arctic, related to positive temperature anomalies in this time step, and the large-scale structure over the Pacific is captured as well. Strong negative anomalies over parts of South America and northern India and Pakistan are reproduced. Emulation and ground truth simulation differ in their fine-scale structure: the ground truth is generally less smooth than the emulation and seems particularly noisy over some dry regions like the Sahara and the Arabic deserts. In these regions, there is a potential for numerical inaccuracies in the isotopic component of climate models, due to small abundances of each isotopic species, and it is hard to untangle which parts of the “noisy” signal have a climatic origin and which parts are simulation

³ We chose the median in terms of anomaly correlation coefficient (ACC).

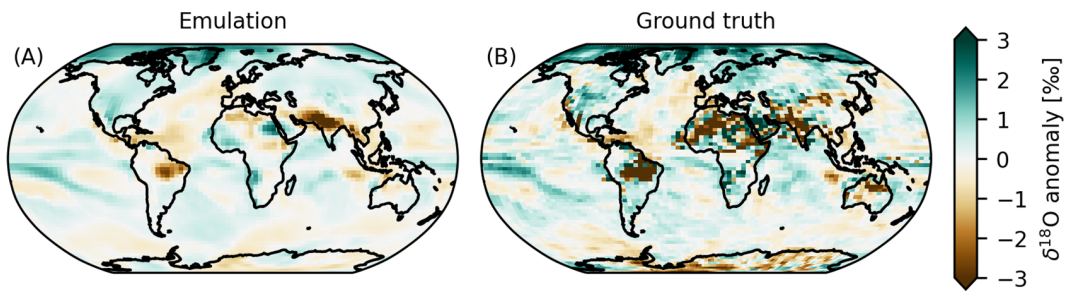


Figure 4. Typical emulation results on iHadCM3 dataset: we show anomalies as they are output by the ML emulator (“Emulation”) and the “true” result in the simulation data set (“Ground truth”). The anomalies are computed with respect to the training set mean. For the selected time step, the anomaly correlation coefficient (ACC) reaches its median value.

artifacts. A part of the overall smoother nature of the UNet regression results can be attributed to the MSE Loss giving a large (quadratic) penalty for strong deviations from the true values, thus, priming the network against predicting values in the tails of the distribution. [Supplementary Figure A.4](#) compares emulation and simulation for three additional time steps: time steps in which the emulation works particularly well or poorly, and a climatically interesting year—1816 CE, the “year without a summer” (Luterbacher and Pfister, 2015), which is caused by a volcanic eruption included in the volcanic forcing of the iHadCM3 simulation. For 1816 CE, we observe that the emulator reproduces a strong negative $\delta^{18}\text{O}$ anomaly in regions where $\delta^{18}\text{O}$ is primarily influenced by temperature, namely in the Arctic, northern North America, and Siberia.

3.2. Comparing machine learning methods

The ML emulation models (UNet architectures and simpler baselines) differ in the quality of their emulation. In the following, we compare the methods amongst each other. For details on the training procedures, network architectures, and method implementations, see [Supplementary Appendix A.1](#). We also address the question of whether using an inherently spherical approach is beneficial over treating the latitude-longitude grid as “flat.” However, the comparison is not trivial: the approaches are developed for data on different grids (plate carrée and icosahedral) and the necessary interpolations may deteriorate

Table 1. Globally averaged R^2 scores for the different ML emulation methods.

Emulation method	R^2 score, plate carrée grid	R^2 score, icosahedral grid
Flat UNet, unmodified	0.352 ± 0.015	0.374 ± 0.017
Flat UNet, modified	0.377 ± 0.005	0.402 ± 0.006
Flat random forest baseline	0.212	0.256
Flat linear regression baseline	0.251	0.274
Flat PCA regression baseline	0.303	0.332
Icosahedral UNet	0.126 ± 0.011	0.396 ± 0.009
Icosahedral PCA regression baseline	0.076	0.339

Note. Bold indicates the best performing methods (highest R^2 values) on each model grid (= each column). Results are calculated for the icosahedral grid that the method of Cohen et al. (2019) operates on and the plate carrée grid. When a method works with data on the other grid, the emulated data is interpolated. “Flat UNet, unmodified” and “Flat UNet, modified” refer to the flat network architecture described in [Section 2.2.1](#), either not applying or applying the modifications to remedy projection artifacts described in that chapter.

performance. Thus, we compute performances on both grids, interpolating the predictions from one grid to another. Results for the globally averaged R^2 scores are shown in Table 1. The best model in the comparison is the “modified” version of the flat UNet that includes the three modifications described in Section 2.2 (area-weighted loss, adapted padding, CoordConv). The effects of the individual modifications are detailed in Supplementary Table A.2 and Supplementary Figure A.6.

All UNet architectures outperform all baseline architectures that operate on the same grid. The best UNet method explains 7% more of the test set variance than the best baseline model, PCA regression. The other baseline models perform worse. In particular, it seems that the random forest baseline, which regresses on a pixel-to-pixel level is not able to capture the spatially varying relationships between $\delta^{18}\text{O}$ and the predictor variables surface temperature and precipitation amount well enough, even when including coordinates as additional inputs. The spatial performance differences between the UNet methods and the best baselines are visualized in Supplementary Figure A.5. The improvements by the UNets are largest over oceans.

On the icosahedral grid, the icosahedral UNet and the modified flat UNet achieve R^2 scores that are not significantly different. On the plate carrée grid, however, the results of the icosahedral UNet are much worse. This drop can largely be attributed to the interpolation method (see Supplementary Figure A.7): on the plate carrée grid, neither training data nor results of the flat UNet are interpolated, while interpolations are necessary in both cases for the icosahedral UNet.

3.3. Sensitivity experiments

We conduct a range of sensitivity experiments, to test (a) the influence of each predictor variable on the results, (b) whether we can further improve the performance of our ML method, and (c) whether emulation quality varies with timescale.

First, we use the modified flat UNet architecture as employed in Section 3.2 and test how the results differ if we exclude one of the predictor variables. The globally averaged R^2 score on the plate carrée grid drops from 0.377 ± 0.005 if both precipitation and temperature are used to 0.327 ± 0.006 when using only precipitation and to 0.251 ± 0.004 when only using temperature. The spatial differences in emulation quality follow the large-scale behavior of the correlation structure in panel c of Figure 2. When precipitation is excluded, the performance decreases most over low latitudes, while the R^2 score drops over polar regions without temperature. This is visualized in Supplementary Figure A.8.

To potentially improve the emulation results even further, we create variations of the modified flat UNet architecture: a “wider” version in which the number of computed features per network layer is doubled ($R^2 = 0.386 \pm 0.008$, plate carrée grid), and a “deeper” version with six additional network layers⁴, which obtains $R^2 = 0.389 \pm 0.006$ (plate carrée grid), both improving over the default choice by roughly 0.01. Additionally, we test whether results could be improved by tuning the learning rate of the employed optimizer by testing a grid of 20 logarithmically spaced values between 10^{-4} and 10^{-1} . The performance is best for learning rates between 10^{-3} and 10^{-2} . However, no substantial improvements over the default parameter choice were reached in the limited range of tested values.

The monthly timescale differs from the interannual scale by a pronounced seasonal cycle of $\delta^{18}\text{O}$ in many regions. Thus, even a simple climatology can explain a part of the variability in $\delta^{18}\text{O}$. To exclude this trivially explainable part from the computation of the R^2 score, we compute the score separately for each month. Results are similar to the results on the interannual scale with roughly 40% of variance explained. The higher time resolution suggests exploring whether the emulation can profit from taking the temporal context into account. We test this by including the temperature and precipitation of not only the current time step but also the previous month as inputs to the emulation of $\delta^{18}\text{O}$. Results do not improve strongly, however, possibly because the investigated timescale is still larger than the average atmospheric moisture residence time (Trenberth, 1998).

⁴I.e. one additional “depth step” in Figure A.11.

3.4. Cross-model comparison

For practical applicability, it is essential that an emulator's performance is robust under varying climatic conditions and under potential biases of the climate model that produces the training data for the emulator. We address these questions by testing how well our emulation generalizes to data generated with different climate models (iCESM, ECHAM5-wiso, isoGSM). To do so, we train the best model architecture so far, the deeper modified flat UNet, on data from iHadCM3. Subsequently, the trained network is used to emulate $\delta^{18}\text{O}$ for the test sets of the other climate model datasets. Results of the emulation are visualized in Figure 5. For all datasets the mean R^2 score is positive, meaning that in the global average, the emulation is preferable to predicting the mean state of the corresponding training set. The R^2 score is highest for the ECHAM5-wiso simulation and lowest for isoGSM, where 80% less variance is explained than on iHadCM3.

In all three cross-prediction cases, the performance drops strongly in the Pacific Ocean west of South America, a region that is important for the El Niño–Southern Oscillation (ENSO). This might hint at inter-model differences in the spatial pattern of ENSO variability. For isoGSM, the emulation quality over Antarctica is considerably worse than for all other models. The Antarctic in isoGSM is much less depleted in $\delta^{18}\text{O}$ (less negative $\delta^{18}\text{O}$) than in the other models while showing similar equator-to-pole temperature gradients (Bühler et al., 2022). This can potentially impact the relationship between the temporal variations of temperature and $\delta^{18}\text{O}$.

For isoGSM and iCESM, R^2 is negative over large areas of the mid-latitude oceans. As synoptic-scale variability of moisture transport pathways might be an important factor for $\delta^{18}\text{O}$ in the mid-latitudes, adding predictor variables that encode information on the atmospheric circulation in the respective models could improve the results. The independence of the isoGSM and iCESM runs in these regions must be assessed carefully: isoGSM is forced by sea-surface temperatures and sea-ice distributions of a last-millennium run with CCSM4, which is a predecessor model of iCESM. Therefore, characteristics of iCESM might also be present in the isoGSM results.

We also test how well the baseline ML models generalize when employed to estimate $\delta^{18}\text{O}$ for other climate models. The very simplistic pixel-wise linear regression yields better results than the PCA

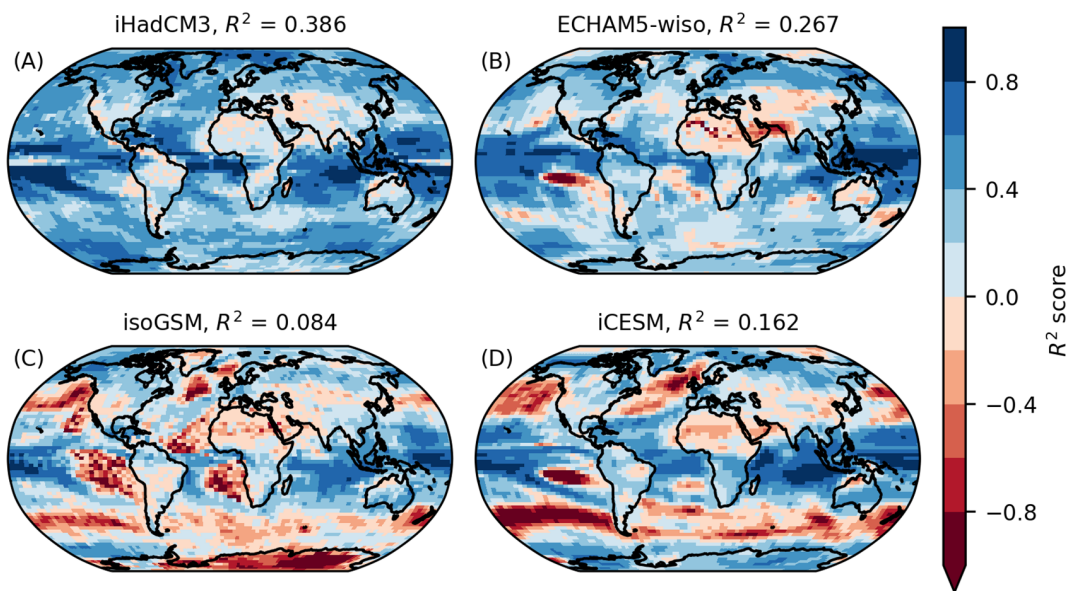


Figure 5. Results for the cross-prediction task: a UNet is trained on the iHadCM3 training data set. The performance is then evaluated on the test set of various climate models; shown R^2 scores are averages over 10 runs.

regression baseline. [Supplementary Figure A.9](#) shows the cross-model performance of the linear regression baseline. While the R^2 score for iHadCM3 itself is significantly smaller than for the UNet model, the loss of performance when doing cross-prediction is much smaller. For iCESM, the results are even better than those obtained with the UNet model. Especially over mid-latitude oceans, the R^2 scores of the linear regression are better than the ones obtained with the UNet.

4. Discussion

In a first step toward data-driven emulation of water-isotope variability in precipitation from standard climate model output variables, we show that in a simulated dataset 40% of the interannual $\delta^{18}\text{O}$ variance can be explained by ML models. The emulation quality follows patterns of the correlation between $\delta^{18}\text{O}$ and the predictor variables, precipitation amount and surface temperature. This hints at the possibility of further improving the emulation by including other variables that are statistically connected to $\delta^{18}\text{O}$ as predictors. $\delta^{18}\text{O}$ composition depends on atmospheric moisture transport, which in turn, depends on atmospheric circulation. Thus, variables encoding information on atmospheric circulation, such as sea-level pressure, are promising candidates which should be explored in future research. This could be particularly relevant in the mid-latitudes, where the comparably poor performance of the emulators might be due to synoptic-scale moisture transport variability which is not well captured by annual or monthly means of precipitation and temperature. In addition, relative humidity seems a promising candidate as it is important for the evolution of $\delta^{18}\text{O}$ during the evaporation process.

It should be noted that correlation structures between predictor variables and $\delta^{18}\text{O}$ are likely timescale dependent. Our results suggest that temperature, precipitation, and atmospheric circulation variations due to internal variability in the climate system and short-scale external forcing such as volcanic eruptions and solar variability are the most important factors controlling interannual $\delta^{18}\text{O}$ variability. On the other hand, changes in long-term external forcings such as greenhouse gas concentrations and Earth's orbital configuration, and variations in oceanic circulation have been found to explain $\delta^{18}\text{O}$ changes on millennial and orbital (10,000 years and longer) timescales (He et al., 2021). This varying importance of factors controlling climate variations can also result in timescale-dependent relationships between the predictor variables surface temperature and precipitation amount (Rehfeld and Laepple, 2016), which limits the generalization of emulators between timescales. Meanwhile, on timescales from hours to weeks, the memory in the atmosphere is higher. Thus, taking into account previous time steps and explicitly tracking moisture pathways, for example, in tropical or extratropical cyclones could improve the emulation performance. On these timescales, ML methods to model sequences of data, like long short-term memory (LSTM), recurrent neural networks (RNNs), or transformer models could be good alternatives.

A tested spherical CNN architecture shows no clear benefit over a modified version of the standard flat UNet for our task of emulating $\delta^{18}\text{O}$ in precipitation globally. We suppose that this is partly due to the strong latitudinal dependence of the statistical relationships between $\delta^{18}\text{O}$ and the predictor variables (as indicated by correlations in [Figure 2c](#)). Thus, the strength of the spherical network architecture, namely its equivariance to rotations, possibly does not offer a strong benefit. Additionally, the interpolation between the plate carrée grid and the icosahedral grid deteriorates the results. This might be remedied by “differentiating through” the interpolation or directly learning the interpolation, as is done by Lam et al. (2022). Using ML architectures that are equivariant to approximate symmetries in the Earth system might still be beneficial in many applications, since adapting the ML approach to symmetries of the problem reduces overfitting and the demands for training data. One might use Cohen and Welling (2016), for example, as a starting point and test a network that is equivariant under rotations around the polar axis and reflections on the equatorial plane.

The cross-model emulations can be seen as a supplement to test for the generalization to the (unavailable) real-world $\delta^{18}\text{O}$ data. Assuming that each climate model possesses different deficiencies in its $\delta^{18}\text{O}$ simulation, robustness under varying models would hint at robustness in the generalization to real-world data. Additionally, reliable $\delta^{18}\text{O}$ emulations for climate models that do not possess an

implementation of water isotopologues would ideally be done with an emulator that does not overfit to a certain climate model it was trained on. Two reasons that might make an ML emulator perform poorly under cross-emulations are (a) weak statistical connections between $\delta^{18}\text{O}$ and the predictor variables in the training set and (b) differences in the statistical connections of $\delta^{18}\text{O}$ and the predictor variables between climate models. Variations between the climate models' isotope modules likely affect these statistical connections. Particularly, the models differ in the formulation of kinetic fractionation: iHadCM3 is based on Cappa (2003), while isoGSM, ECHAM5-wiso, and iCESM use results of Merlivat (1978). We investigate whether drops in cross-prediction performance can be attributed to causes (a) and (b) in [Supplementary Appendix A.2](#) and [Supplementary Figure A.3](#). Indeed, most regions, in which there is a drop in emulation performance, coincide with regions of differing correlation structures or weak correlations between $\delta^{18}\text{O}$ and the predictor variables in the iHadCM3 dataset ([Supplementary Figures A.3, B4 to D4 and B2 to D2](#)). In regions with weak correlations between $\delta^{18}\text{O}$ and the predictor variables in the iHadCM3 dataset, such as the Southern Hemisphere mid-latitudes, the UNet has to predict $\delta^{18}\text{O}$ based on spatial similarity structures (teleconnections). The poor performance in the Southern Hemisphere mid-latitudes in IsoGSM and iCESM suggest that the spatial similarity structures differ between those two GCMs and iHadCM3. Here, predictors that encode atmospheric circulation more directly such as sea-level pressure could be beneficial in future studies.

This interpretation is supported by a much sharper drop in performance of the UNet architectures than simple linear regression when methods were trained on the iHadCM3 climate model and then used to emulate other climate model data. As a result, the R^2 scores on the other climate models were comparable between UNet and linear regression. This suggests that the UNet might overfit to the spatial anomaly patterns in iHadCM3, given the limited information provided by the predictor variables. This overfitting will partly reproduce deficiencies of the respective dataset used for the training of the emulator. It was shown previously that the models used in our study differ in their mean climate state. For example, iHadCM3 and ECHAM5-wiso show a similar global temperature state, but iHadCM3 $\delta^{18}\text{O}$ is much more negative in the global mean (Bühler et al., 2022). Similar differences in the spatial anomaly patterns between models need to be explored further to understand their contribution to poor cross-model emulation performance. To obtain a more robust emulator that is applicable across models, one might utilize data from multiple climate models and climate states (e.g., Last Glacial Maximum, mid-Holocene, Pliocene) in the training set. The cross-prediction performance might also be influenced by the interpolations that are necessary to re-grid all climate model datasets to the resolution of iHadCM3. We would expect interpolation artifacts to appear as small-scale noise. However, we mostly find differences in large-scale patterns, that are structurally similar to the results on the iHadCM3 dataset, in which no interpolation has been applied ([Figure 5a](#)). This indicates that interpolation artifacts are of minor relevance for the reduced performance in the cross-model emulations.

Spatially, ML estimates are smoother than the true simulated data. The ground truth data show very noisy behavior over dry regions, part of which is likely due to numerical instabilities in the computation of $\delta^{18}\text{O}$ for very low precipitation amounts. Missing data points also occur more frequently in these regions, thus potentially biasing the emulator and its measured performance. Because of these inconsistencies in the input data, it might be beneficial to focus on particular regions when developing an emulator with the aim of comparing to a certain natural climate archive. Examples are the polar regions for comparisons to ice core data or the mid-latitudes for speleothem records. Restricting the spatial extent would also alleviate artifacts of the map projection and render spherical approaches unnecessary. Alternatively, one might think about the application of ML to do in-painting of missing values of $\delta^{18}\text{O}$ for the training of emulators, similar to Kadow et al. (2020). In this case, the incomplete $\delta^{18}\text{O}$ would serve as an input to the ML method in addition to precipitation and temperature.

Training an isotope emulator on real-world data would avoid uncertainties originating from climate models and the implementation of isotopes within them. It would also increase the emulator's utility for research areas that work with observational isotope data. For instance, the local isotopic composition of precipitation can be valuable when studying human influences on hydrology (Good et al., 2014). The

isotopic signature of precipitation is archived in the composition of plants or the tissue of wild animals. Understanding spatial and temporal variations of stable water isotopes can, therefore, help in studying plant origins and wildlife migration patterns, and contribute to food authentication (Bowen et al., 2005; West et al., 2007; Cernusak et al., 2016). Databases of observed $\delta^{18}\text{O}$ in precipitation (IAEA/WMO, 2020) or $\delta^{18}\text{O}$ from natural climate archives (Konecky et al., 2020) are publicly available. However, challenges arise from the spatial scarcity and unequal distribution of data, and the short temporal coverage of observations. Here, using graph networks like the one developed by Defferrard et al. (2020) might be an option, and likely strong prior constraints would need to be used to compensate for small dataset sizes. For the future goal of comparing emulations to $\delta^{18}\text{O}$ measured in natural climate archives, archive-specific processes need to be taken into account. This is because $\delta^{18}\text{O}$ in precipitation is not archived directly, but always as the response of a sensor of the archiving medium. For example, precipitation $\delta^{18}\text{O}$ is archived in speleothem records as calcite carbonate in accumulating layers that form from cave drip water (Fairchild and Baker, 2012).

We calculate yearly $\delta^{18}\text{O}$ as the unweighted average of monthly $\delta^{18}\text{O}$. In most natural climate archives, yearly $\delta^{18}\text{O}$ is weighted by precipitation amount. We tested the influence of such a weighting and found that it does not impact the emulator performance negatively (not shown). However, climate archives can also show seasonal preference in their sensitivity to $\delta^{18}\text{O}$ (Wackerbarth et al., 2010; Fohlmeister et al., 2017; Baker et al., 2019) such that there is likely no optimal way for computing yearly values. Including archive-specific processes could either be a second step in a two-step approach, where an ML emulator is trained to predict $\delta^{18}\text{O}$ in precipitation and then a proxy system model (Evans et al., 2013) is used to forward-model archive-specific processes. Alternatively, one might include a differentiable proxy system model in the ML pipeline. This would make it possible to train the ML architecture directly with proxy data instead of $\delta^{18}\text{O}$ measured in precipitation.

5. Conclusion

In this study, we explored the ability of machine learning methods to emulate oxygen isotopes as simulated by isotope-enabled General Circulation Models (GCMs). Focussing on interannual variability in a last-millennium simulation, we show that UNet neural networks improve the emulation performance compared to baseline methods such as pixel-wise linear regression and PCA regression. Averaged over all grid cells, our best-performing UNet architecture explains 40% of the temporal $\delta^{18}\text{O}$ variance. The emulation performs best in polar regions, where $\delta^{18}\text{O}$ is strongly controlled by surface temperature variations, and in low latitude ocean areas, where $\delta^{18}\text{O}$ is highly correlated with precipitation amounts. Lowest performances occur in arid regions, partly because of numerical instabilities in the simulation of $\delta^{18}\text{O}$ for very low precipitation amounts. Using a spherical network architecture does not improve the results compared to a modified flat architecture, which accounts better for Earth's spherical geometry than a default UNet architecture. This might be because our spherical UNet architecture is not optimized to capture latitudinal dependences in the relationships between $\delta^{18}\text{O}$ and the predictor variables.

We tested the generalization of the emulator trained on output from the iHadCM3 GCM to last-millennium simulations with other GCMs. While the performance is better than predicting the model's climatology for all GCMs, the explained variance is substantially lower than for iHadCM3. Performances are especially poor in regions where the correlation structure between $\delta^{18}\text{O}$ and the predictor variables differs from the correlation structure in iHadCM3 and in regions with low correlations between $\delta^{18}\text{O}$ and the predictor variables in iHadCM3. In the latter case, the UNet architecture learns spatial dependence structures to improve the emulation of $\delta^{18}\text{O}$. This improves the performance within iHadCM3 compared to pixel-wise regression. However, these spatial structures seem to differ too much between GCMs to facilitate skillful cross-model emulations, especially in the mid-latitudes where encoding synoptic-scale circulation variations could be important to capture $\delta^{18}\text{O}$ variations.

To further improve emulation performance, adding more predictor variables could be a promising next step. In particular, variables such as sea-level pressure, which capture characteristics of the atmospheric

circulation more directly than surface temperature and precipitation amount, could help in regions with currently poor performance. To compare emulated isotopes to $\delta^{18}\text{O}$ measured in natural climate archives such as ice cores and speleothems, a way of incorporating archive-specific processes needs to be investigated. This could be done by incorporating differentiable proxy system models into UNet architectures or by applying proxy system models to the emulator output in a two-step approach. For comparison with $\delta^{18}\text{O}$ measurements in natural climate archives, the timescales of variations recorded by the archives are important. While we focused on interannual timescales in this study, shorter as well as longer timescales could be explored in future research to understand the importance of synoptic-scale processes, local predictor variables, and external forcings for $\delta^{18}\text{O}$ emulation across timescales.

Abbreviations

ACC	anomaly correlation coefficient
AOGCM	Atmosphere–Ocean General Circulation Model
CDO	Climate Data Operators tool set
CNN	convolutional neural network
ECHAM5-wiso	ECHAM5/MPI-OM, an investigated climate model
ENSO	El Niño–Southern Oscillation
GCM	General Circulation Model
iCESM	iCESM1 version 1.2, an investigated climate model
iGCM	isotope-enabled General Circulation Model
iHadCM3	Hadley Center Climate Model version 3, an investigated climate model
isoGSM	Scripps Experimental Climate Prediction Center’s Global Spectral Model, an investigated climate model
ML	machine learning
NAO	North Atlantic Oscillation
NGRIP	North Greenland Ice Core Project
PCA	principal component analysis
WAIS Divide	West Antarctic Ice Sheet Divide ice core project

Acknowledgments. We thank Nadine Theisen for help with the implementation and testing of baseline models. We are grateful for the contribution and standardization of model data from Jesper Sjolte, Kei Yoshimura, Madhavan Midhun, Martin Werner, and Josefine Axelsson. Kira Rehfeld is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764. Jonathan Wider is supported by BMBF (Federal Ministry of Education and Research) through the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI).

Author contribution. Conceptualization: U.K., K.R., J.W., N.W.; Data curation: J.B., K.R.; Formal Analysis: J.W.; Funding Acquisition: K.R.; Investigation: J.W.; Methodology: J.K., U.K., K.R., J.W., N.W.; Project administration: K.R., U.K.; Resources: K.R.; Software: J.W.; Supervision: J.B., J.K., U.K., K.R., N.W.; Validation: J.K., J.W.; Visualization: J.W.; Writing original draft: J.K., J.W.; Writing—review & editing: all authors. All authors approved the final submitted draft.

Competing interest. The authors declare no competing interests exist.

Data availability statement. The data used in this study can be freely downloaded here: <https://doi.org/10.5281/zenodo.7516327>. Code to reproduce our experiments is publicly available at <https://github.com/paleovar/isoEm/releases/v1.0>; it is subject to the license statements in the GitHub repository.

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

Funding statement. This research was supported by grants from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the STACY (project no. 395588486) and CLIMAIC (project no. 442926051) projects.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/eds.2023.29>.

References

- Baertschi P** (1976) Absolute $\delta^{18}\text{O}$ content of standard mean ocean water. *Earth and Planetary Science Letters* 31(3), 341–344. [https://doi.org/10.1016/0012-821X\(76\)90115-1](https://doi.org/10.1016/0012-821X(76)90115-1).
- Baker A, Hartmann A, Duan W, Hankin S, Comas-Bru L, Cuthbert MO, Treble PC, Banner J, Genty D, Baldini LM, Bartolomé M, Moreno A, Pérez-Mejías C and Werner M** (2019) Global analysis reveals climatic controls on the oxygen isotope composition of cave drip water. *Nature Communications* 10(1), 2984. <https://doi.org/10.1038/s41467-019-11027-w>.
- Bowen GJ** (2010) Isoscapes: Spatial pattern in isotopic biogeochemistry. *Annual Review of Earth and Planetary Sciences* 38(1), 161–187. <https://doi.org/10.1146/annurev-earth-040809-152429>.
- Bowen GJ and Revenaugh J** (2003) Interpolating the isotopic composition of modern meteoric precipitation. *Water Resources Research* 39(10), 1299. <https://doi.org/10.1029/2003WR002086>.
- Bowen GJ, Wassenaar LI and Hobson KA** (2005) Global application of stable hydrogen and oxygen isotopes to wildlife forensics. *Oecologia* 143(3), 337–348. <https://doi.org/10.1007/s00442-004-1813-y>.
- Brady E, Stevenson S, Bailey D, Liu Z, Noone D, Nusbaumer J, Otto-Bliesner BL, Tabor C, Tomas R, Wong T, Zhang J and Zhu J** (2019) The connected isotopic water cycle in the community earth system model version 1. *Journal of Advances in Modeling Earth Systems* 11(8), 2547–2566. <https://doi.org/10.1029/2019MS001663>.
- Breiman L** (2001) Random forests. *Machine Learning* 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Bühler JC, Axelsson J, Lechleitner FA, Fohlmeister J, LeGrande AN, Midhun M, Sjolte J, Werner M, Yoshimura K and Rehfeld K** (2022) Investigating stable oxygen and carbon isotopic variability in speleothem records over the last millennium using multiple isotope-enabled climate models. *Climate of the Past* 18(7), 1625–1654. <https://doi.org/10.5194/cp-18-1625-2022>.
- Buizert C, Cuffey KM, Severinghaus JP, Baggenstos D, Fudge TJ, Steig EJ, Markle BR, Winstrup M, Rhodes RH and Brook EJ** (2015) The WAIS divide deep ice core WD2014 chronology—Part 1: Methane synchronization (68–31 ka BP) and the gas age–age difference. *Climate of the Past* 11(2), 153–173. <https://doi.org/10.5194/cp-11-153-2015>.
- Cappa CD** (2003) Isotopic fractionation of water during evaporation. *Journal of Geophysical Research* 108(D16), 4525. <https://doi.org/10.1029/2003JD003597>.
- Casado M, Landais A, Picard G, Münch T, Laepple T, Stenni B, Dreossi G, Ekaykin A, Arnaud L, Genthon C, Touzeau A, Masson-Delmotte V and Jouzel J** (2018) Archival processes of the water stable isotope signal in East Antarctic ice cores. *The Cryosphere* 12(5), 1745–1766. <https://doi.org/10.5194/tc-12-1745-2018>.
- Cernusak LA, Barbour MM, Arndt SK, Cheesman AW, English NB, Feild TS, Helliker BR, Holloway-Phillips MM, Holtum JA, Kahmen A, McInerney FA, Munksgaard NC, Simonin KA, Song X, Stuart-Williams H, West JB and Farquhar GD** (2016) Stable isotopes in leaf water of terrestrial plants: Stable isotopes in leaf water. *Plant, Cell & Environment* 39(5), 1087–1102. <https://doi.org/10.1111/pce.12703>.
- Cohen TS, Geiger M, Koehler J and Welling M** (2018) Spherical CNNs. In *International Conference on Learning Representations*. arXiv:1801.10130.
- Cohen TS, Weiler M, Kicanaoglu B and Welling M** (2019) Gauge equivariant convolutional networks and the icosahedral CNN. In Chaudhuri K and Salakhutdinov R (eds), *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. PMLR, pp. 1321–1330. arXiv:1902.04615
- Cohen T and Welling M** (2016) Group equivariant convolutional networks. In Balcan MF and Weinberger KQ (eds), *Proceedings of the 33rd International Conference on Machine Learning*, Vol. 48. New York, NY: PMLR, pp. 2990–2999.
- Colose CM, LeGrande AN and Vuille M** (2016) The influence of volcanic eruptions on the climate of tropical South America during the last millennium in an isotope-enabled general circulation model. *Climate of the Past* 12(4), 961–979. <https://doi.org/10.5194/cp-12-961-2016>.
- Comas-Bru L, Rehfeld K, Roesch C, Amirnezhad-Mozhdehi S, Harrison SP, Atsawaranunt K, Ahmad SM, Brahim YA, Baker A, Bosomworth M, Breitenbach SFM, Burstyn Y, Columbu A, Deininger M, Demény A, Dixon B, Fohlmeister J, Hatvani IG, Hu J, Kaushal N, Kern Z, Labuhn I, Lechleitner FA, Lorrey A, Martrat B, Novello VF, Oster J, Pérez-Mejías C, Scholz D, Scroton N, Sinha N, Ward BM, Warken S and Zhang H** (2020) SISALv2: A comprehensive speleothem isotope database with multiple age–depth models. *Earth System Science Data* 12(4), 2579–2606. <https://doi.org/10.5194/essd-12-2579-2020>.
- Coors B, Condurache AP and Geiger A** (2018) Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds), *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer, pp. 518–533. https://doi.org/10.1007/978-3-030-01240-3_3.
- Dansgaard W** (1964) Stable isotopes in precipitation. *Tellus* 16(4), 436–468. <https://doi.org/10.3402/tellusa.v16i4.8993>.
- Defferrard M, Milani M, Guset F and Perraudin N** (2020) DeepSphere: A graph-based spherical CNN. In *International Conference on Learning Representations (ICLR)* arXiv:2012.15000.
- Esteves C, Allen-Blanchette C, Makadia A and Daniilidis K** (2020) Learning SO(3) Equivariant representations with spherical CNNs. *International Journal of Computer Vision* 128(3), 588–600. <https://doi.org/10.1007/s11263-019-01220-1>.
- Evans M, Tolwinski-Ward S, Thompson D and Anchukaitis K** (2013) Applications of proxy system modeling in high resolution paleoclimatology. *Quaternary Science Reviews* 76, 16–28. <https://doi.org/10.1016/j.quascirev.2013.05.024>.
- Fairchild IJ and Baker A** (2012) *Speleothem Science: From Process to Past Environments*. Chichester: John Wiley & Sons. <https://doi.org/10.1002/9781444361094>.

- Fiorella RP, Siler N, Nusbaumer J and Noone DC (2021) Enhancing understanding of the hydrological cycle via pairing of process-oriented and isotope ratio tracers. *Journal of Advances in Modeling Earth Systems* 13(10). <https://doi.org/10.1029/2021MS002648>.
- Fohlmeister J, Plessen B, Dudashvili AS, Tjallingii R, Wolff C, Gafurov A and Cheng H (2017) Winter precipitation changes during the medieval climate anomaly and the little ice age in arid central asia. *Quaternary Science Reviews* 178, 24–36. <https://doi.org/10.1016/j.quascirev.2017.10.026>.
- Good SP, Kennedy CD, Stalker JC, Chesson LA, Valenzuela LO, Beasley MM, Ehleringer JR and Bowen GJ (2014) Patterns of local and nonlocal water resource use across the western U.S. determined via stable isotope intercomparisons. *Water Resources Research* 50(10), 8034–8049. <https://doi.org/10.1002/2014WR015884>.
- He C, Liu Z, Otto-Bliesner BL, Brady E, Zhu C, Tomas R, Clark P, Zhu J, Jahn A, Gu S, Zhang J, Nusbaumer J, Noone D, Cheng H, Wang Y, Yan M and Bao Y (2021) Hydroclimate footprint of pan-Asian monsoon water isotope during the last deglaciation. *Science Advances* 7(4), eabe2611. <https://doi.org/10.1126/sciadv.abe2611>
- IAEA/WMO (2020) Global Network of Isotopes in Precipitation. The GNIP Database. Retrieved July 25, 2023 from <https://www.iaea.org/services/networks/gnip>.
- IPCC (2023) *Climate Change 2022 – Impacts, Adaptation and Vulnerability [Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change]*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009325844>.
- JungCLAUS JH, Bard E, Baroni M, Braconnot P, Cao J, Chini LP, Egorova T, Evans M, González-Rouco JF, Goussé H, HurrT GC, Joos F, Kaplan JO, Khodri M, Klein Goldewijk K, Krivova N, LeGrande AN, Lorenz SJ, Luterbacher J, Man W, Maycock AC, Meinshausen M, Moberg A, Muscheler R, Nehrbass-Ahles C, Otto-Bliesner BI, Phipps SJ, Pongratz J, Rozanov E, Schmidt GA, Schmidt H, Schmutz W, Schurer A, Shapiro AI, Sigl M, Smerdon JE, Solanki SK, Timmreck C, Toohey M, Usoskin IG, Wagner S, Wu C-J, Yeo KL, Zanchettin D, Zhang Q and Zorita E (2017) The PMIP4 contribution to CMP6 – Part 3: The last millennium, scientific objective, and experimental design for the PMIP4 past1000 simulations. *Geoscientific Model Development* 10(11), 4005–4033. <https://doi.org/10.5194/gmd-10-4005-2017>.
- Kadow C, Hall DM and Ulbrich U (2020) Artificial intelligence reconstructs missing climate information. *Nature Geoscience* 13 (6), 408–413. <https://doi.org/10.1038/s41561-020-0582-5>.
- Konecky BL, McKay NP, Churakova (Sidorova) OV, Comas-Bru L, Dassié EP, DeLong KL, Falster GM, Fischer MJ, Jones MD, Jonkers L, Kaufman DS, Leduc G, Managave SR, Martrat B, Opel T, Orsi AJ, Partin JW, Sayani HR, Thomas EK, Thompson DM, Tyler JJ, Abram NJ, Atwood AR, Conroy JL, Kern Z, Porter TJ, Stevenson SL, von Gunten L and Iso2k Project Members (2020) The iso2k database: A global compilation of paleo-d18o and dh records to aid understanding of common era climate. *Earth System Science Data* 12(3), 2261–2288. <https://doi.org/10.5194/essd-12-2261-2020>.
- Lam R, Sanchez-Gonzalez A, Willson M, Wirnsberger P, Fortunato M, Pritzel A, Ravuri S, Ewalds T, Alet F, Eaton-Rosen Z, Hu W, Merose A, Hoyer S, Holland G, Stott J, Vinyals O, Mohamed S and Battaglia P (2022) GraphCast: Learning skillful medium-range global weather forecasting. Preprint. [arXiv:2212.12794](https://arxiv.org/abs/2212.12794).
- Landrum L, Otto-Bliesner BL, Wahl ER, Conley A, Lawrence PJ, Rosenbloom N and Teng H (2013) Last millennium climate and its variability in CCSM4. *Journal of Climate* 26(4), 1085–1111. <https://doi.org/10.1175/JCLI-D-11-00326.1>.
- Liu R, Lehman J, Molino P, Such FP, Frank E, Sergeev A and Yosinski J (2018) An intriguing failing of convolutional neural networks and the coordconv solution. In Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. [arXiv:1807.03247](https://arxiv.org/abs/1807.03247).
- Luterbacher J and Pfister C (2015) The year without a summer. *Nature Geoscience* 8(4), 246–248. <https://doi.org/10.1038/ngeo2404>.
- Merlivat L (1978) Molecular diffusivities of H216O, HD16O, and H218O in gases. *Journal of Chemical Physics* 69(6), 2864–2871. <https://doi.org/10.1063/1.436884>.
- Mook W (ed.) (2000) Environmental isotopes in the hydrological cycle: Principles and applications, Vol. I: Introduction: Theory, Methods, Review. In *Technical Documents in Hydrology*. Paris: UNESCO.
- Morice CP, Kennedy JJ, Rayner NA and Jones PD (2012) Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres* 117 (D8). <https://doi.org/10.1029/2011JD017187>.
- North Greenland Ice Core Project Members (2004) High-resolution record of northern hemisphere climate extending into the last interglacial period. *Nature*, 431(7005), 147–151. <https://doi.org/10.1038/nature02805>
- PAGES2k-Consortium (2019) Consistent multidecadal variability in global temperature reconstructions and simulations over the common era. *Nature Geoscience* 12(8), 643–649. <https://doi.org/10.1038/s41561-019-0400-0>.
- Rehfeld K and Laepple T (2016) Warmer and wetter or warmer and dryer? Observed versus simulated covariability of holocene temperature and rainfall in Asia. *Earth and Planetary Science Letters* 436, 1–9. <https://doi.org/10.1016/j.epsl.2015.12.020>.
- Ronneberger O, Fischer P and Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
- Schulzweida U (2020) CDO User Guide. <https://doi.org/10.5281/ZENODO.5614769>.

- Terzer S, Wassenaar LI, Araguás-Araguás LJ and Aggarwal PK** (2013) Global isoscapes for $\delta^{18}\text{O}$ and $\delta^2\text{H}$ in precipitation: Improved prediction using regionalized climatic regression models. *Hydrology and Earth System Sciences* 17(11), 4713–4728. <https://doi.org/10.5194/hess-17-4713-2013>.
- Tindall JC, Valdes PJ and Sime LC** (2009) Stable water isotopes in HadCM3: Isotopic signature of El Niño–Southern oscillation and the tropical amount effect. *Journal of Geophysical Research* 114(D4), D04111. <https://doi.org/10.1029/2008JD010825>.
- Trenberth KE** (1998) Atmospheric moisture residence times and cycling: Implications for rainfall rates and climate change. *Climatic Change* 39(4), 667–694. <https://doi.org/10.1023/A:1005319109110>.
- Vachon RW, Welker JM, White JWC and Vaughn BH** (2010) Monthly precipitation isoscapes ($\delta^{18}\text{O}$) of the United States: Connections with surface temperatures, moisture source conditions, and air mass trajectories. *Journal of Geophysical Research* 115(D21), D21126. <https://doi.org/10.1029/2010JD014105>.
- Wackerbarth A, Scholz D, Fohlmeister J and Mangini A** (2010) Modelling the $\delta^{18}\text{O}$ value of cave drip water and speleothem calcite. *Earth and Planetary Science Letters* 299(3–4), 387–397. <https://doi.org/10.1016/j.epsl.2010.09.019>.
- Werner M, Haese B, Xu X, Zhang X, Butzin M and Lohmann G** (2016) Glacial–interglacial changes in H_2^{18}O , HDO and deuterium excess—results from the fully coupled ECHAM5/MPI-OM earth system model. *Geoscientific Model Development* 9(2), 647–670. <https://doi.org/10.5194/gmd-9-647-2016>.
- West JB, Ehleringer JR and Cerling TE** (2007) Geography and vintage predicted by a novel GIS model of wine $\delta^{18}\text{O}$. *Journal of Agricultural and Food Chemistry*, 55(17), 7075–7083. <https://doi.org/10.1021/jf071211r>.
- Yoshimura K, Kanamitsu M, Noone D and Oki T** (2008) Historical isotope simulation using reanalysis atmospheric data. *Journal of Geophysical Research* 113(D19), D19108. <https://doi.org/10.1029/2008JD010074>.