DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

Evaluation of statistical climate reconstruction methods based on pseudoproxy experiments using linear and machine learning methods

Zeguo Zhang, Sebastian Wagner, Marlene Klockmann, Eduardo Zorita

Helmholtz Zentrum Hereon, Institute of Coastal Systems, 21502 Geesthacht, Germany

Correspondence to: Zeguo Zhang (zeguo.zhang@hereon.de)

Abstract. Three different climate field reconstruction (CFR) methods employed to reconstruct North Atlantic-European (NAE) and Northern Hemisphere (NH) summer season temperature over the past millennium from proxy records are tested in the framework of pseudoproxy experiments derived from three climate simulations with Earth System Models. Two of these methods are traditional multivariate linear methods (Principal Components Regression, PCR and Canonical Correlation Analysis, CCA), whereas the third method (Bidirectional Long-Short-Term Memory Neural Network, Bi-LSTM) belongs to the category of machine learning methods. The Bi-LSTM method does not need to assume linear and temporally stable relationships between the underlying proxy network and the targeted climate field, in contrast to PCR and CCA. In addition, Bi-LSTM incorporates information on the serial correlation of the time series. All three methods tested herein achieve reasonable reconstruction performance in both spatial and temporal scale. Generally, the reconstruction skill is higher in regions with denser proxy coverage, but reconstruction skill is also achieved in proxy-free areas due to climate teleconnections. All three CFR methodologies generally tend to more strongly underestimate the target temperature variations as more noise is introduced into the pseudoproxies. The Bi-LSTM method tested in our experiments shows relatively worse reconstruction skills compared to PCR and CCA, yet it brings some encouraging results on capturing extreme cooling climate signals. This indicates that this nonlinear CFR method could be a potential methodology for past climate extremes analysis.

1 Introduction

The reconstruction of past climates helps to better understand past climate variability and pose the projected future climate evolution against the backdrop of natural climate variability (PAGES 2k Consortium, 2013, 2017, 2019; PAGES Hydro2k Consortium, 2017; Schmidt, 2010; Evans et al., 2014; Christiansen and Ljungqvist, 2016). Paleoclimate reconstruction also provide us with a deeper perspective to better understand the effect of external forcing on climate (Smerdon et al., 2011, 2016; Smerdon, 2012; Wang et al., 2017). However, systematic observational/instrumental climate records are only available starting from the middle of the 19th century, which hinders to capture the full spectrum of past climate variations. Consequently, our understanding of climate variations in earlier centuries is mainly based on indirect proxy records (such as tree rings, ice cores), etc.

Evaluation of statistical climate reconstruction methods based on pseudoproxy experiments using linear and machine learning methods

Zeguo Zhang, Sebastian Wagner, Marlene Klockmann, Eduardo Zorita

Helmholtz Zentrum Hereon, Institute of Coastal Systems, 21502 Geesthacht, Germany

Correspondence to: Zeguo Zhang (zeguo.zhang@hereon.de)

Abstract. Three different climate field reconstruction (CFR) methods are employed to reconstruct spatially resolved North Atlantic-European (NAE) and Northern Hemisphere (NH) summer temperature over the past millennium from proxy records These are tested in the framework of pseudoproxy experiments derived from two climate simulations with comprehensive Earth System Models. Two of these methods are traditional multivariate linear methods (Principal Components Regression, PCR and Canonical Correlation Analysis, CCA), whereas the third method (Bidirectional Long-Short-Term Memory Neural Network, Bi-LSTM) belongs to the category of machine learning methods. In contrast to PCR and CCA, the Bi-LSTM does not need to assume a linear and temporally stable relationships between the underlying proxy network and the target climate field,. In addition, Bi-LSTM naturally incorporates information of the serial correlation of the time series. Our working hypothesis is that the Bi-LSTM method will achieve a better reconstruction of the amplitude of past temperature variability. In all tests, the calibration period was set to the observational period, and the validation period was set to the pre-industrial centuries. All three methods tested herein achieve reasonable reconstruction performance on both spatial and temporal scales, with the exception of an overestimation of the interannual variance by PCR, which may be due to overfitting resulting from the rather short length of calibration period and the large number of predictors. Generally, the reconstruction skill is higher in regions with denser proxy coverage, but it is also reasonable high in proxy-free areas due to climate teleconnections. All three CFR methodologies generally tend to more strongly underestimate the variability of spatially averaged temperature indices as more noise is introduced into the pseudoproxies. The Bi-LSTM method tested in our experiments using a limited calibration dataset shows relatively worse reconstruction skills compared to PCR and CCA and, therefore, our working hypothesis that a more complex machine-learning method would provide better reconstructions for temperature fields was not confirmed.

1 Introduction

The reconstruction of past climates helps to better understand past climate variability and pose the projected future climate evolution against the backdrop of natural climate variability (Mann and Jones, 2003; Jones and Mann, 2004; Jones et al., 2009; Frank et al., 2010; Schmidt, 2010; Christiansen and Ljungqvist, 2012; Evans et al., 2014; Smerdon and Pollack, 2016;

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

The reconstruction of past climates based on proxy data requires the application of statistical methods to translate the information contained in the proxy records into climate variables such as temperature. These methods add an additional layer of statistical uncertainty and bias to the final reconstruction, in addition to the uncertainties originating in the sparse data coverage and in the presence of non-climatic variability in the proxy records. All these sources of error impact the quality of climate reconstructions. One way to estimate this impact is the test of reconstruction methods in the controlled conditions provided by climate simulations with state-of-the-art Earth System models. These models provide virtual climate trajectories, which although possibly not completely realistic, are from the model's perspective physically consistent. The skill of the statistical method, the impact of proxy network coverage and of the amount of climate signal present in the proxy records can thus be evaluated in that virtual reality of climate models, once adequate synthetic proxy records are constructed. These tests are generally denoted pseudo-proxy experiments (PPEs, Smerdon, 2012, Gómez-Navarro et al, 2017).

Many scientific studies that employ pseudo-proxies and real proxies have focused on global and hemisphere climate field or climate index reconstructions (Mann et al., 2002, 2005; von Storch et al., 2004; Smerdon, 2012; Michel et al., 2020; Hernández et al., 2020). These studies have identified several deficiencies that are common to most climate reconstructions methods, such as a general tendency to 'regress to the mean', which results in an underestimation of the reconstructed climate variability. This underestimation becomes more evident when the available proxy information becomes of less quality - diminishing the climate signal contained in the proxy records. In addition, sparser networks - shrinking proxy network coverage - may lead to biased reconstructions (Wang et al., 2014; Evans et al., 2014; Amrhein et al., 2020; Po-Chedley et al., 2020). Thus, significant scope still remains for further developing and evaluating CFR methodologies and in designing methods that are less prone to those common deficiencies (Christiansen and Ljungqvist, 2016).

In the present study, we test a new non-linear CFR method that belongs to the machine learning family, a Bidirectional Long-Short-Term Neural Network (Bi-LSTM). We compare the performance of this method with two well-established classical multi-variate linear regression methods, Principal Component Regression (PCR) and Canonical Correlation Analysis (CCA). Traditional CFRs usually assume linear and temporally stable relationships between the local variables captured by the proxy network and the target climate field. Likewise, the spatial patterns of climate variability are considered as stationary (Pyrina et al., 2017; Wang et al., 2014; Smerdon et al., 2016). However, climate change is dynamic and chaotic, and many links between climate fields can be non-linear (Schneider et al., 2018; Dueben and Bauer, 2018; Huntingford et al., 2019; Nadiga, 2020). Nonlinear machine leaning-based CFR methods (for instance, Artificial Neural Networks-ANN) could help capture capture underlying linear and nonlinear relationships between proxy records and the large-scale climate as realistically as possible (Rasp and Lerch, 2018; Schneider et al., 2018; Rolnick et al., 2019; Huang et al., 2020; Nadiga, 2020; Chattopadhyay et al., 2020; Lindgren et al., 2021). Moreover, machine-learning methods do not necessarily rely on statistical methods to first obtain the principal spatial climate patterns, such as Principal Component Analysis-PCA. The full inherent variability in the original dataset is sequentially and dynamically adjusted and captured with optimized hyper-parameters during the model training process (Goodfellow et al., 2016). The classical recurrent neural network (RNN) and Long Short-Term Memory Network (LSTM) can usually only receive and process information from prior forward inference steps, whereas the Bi-LSTM

Christiansen and Ljungqvist, 2017). Paleoclimate reconstructions also provide us with a deeper perspective to better understand the effect of external forcing on climate (Hegerl et al., 2006, 2007; Schurer et al., 2013, 2014; Anchukaitis et al., 2012, 2017; Tejedor et al., 2021). However, systematic observational/instrumental climate records are only available starting from the middle of the 19th century, which fails to capture the full spectrum of past climate variations. Consequently, our understanding of climate variations prior to 1850 is mainly based on indirect proxy records (such as tree rings, ice cores, etc. Jones and Mann, 2004). The reconstruction of past climates based on proxy data requires the application of statistical methods to translate the information contained in the proxy records into climate variables such as temperature. These methods add an additional layer of statistical uncertainty and bias to the final reconstruction, in addition to the uncertainties originating in the sparse data coverage and in the presence of non-climatic variability in the proxy records. All these sources of error impact the quality of climate reconstructions. One way to estimate this impact is the test of reconstruction methods in the controlled conditions provided by climate simulations with state-of-the-art Earth System Models. These models provide virtual climate trajectories, which although possibly not completely realistic, are from the model's perspective physically consistent. The skill of the statistical method, the impact of proxy network coverage and of the amount of climate signal present in the proxy records can thus be evaluated in that virtual reality of climate models, once adequate synthetic proxy records are constructed. These tests are generally denoted pseudo-proxy experiments (PPEs; Smerdon, 2012; Gómez-Navarro et al., 2017).

Many scientific studies that employ pseudo-proxies and real proxies have focused on global, hemispheric climate field or climate index reconstructions (Mann et al., 2002, 2005; von Storch et al., 2004; Smerdon, 2012; Michel et al., 2020; Hernández et al., 2020). These studies have identified several deficiencies that are common to most climate reconstructions methods, such as a general tendency to 'regress to the mean', which results in an underestimation of the reconstructed climate variability. This underestimation becomes more evident when the available proxy information becomes of less quality - diminishing the climate signal contained in the proxy records. In addition, sparser networks - shrinking proxy network coverage - may lead to biased reconstructions (Wang et al., 2014; Evans et al., 2014; Amrhein et al., 2020; Po-Chedley et al., 2020). Thus, significant scope still remains for further developing and evaluating climate field reconstructions (CFR) methodologies and in designing methods that are less prone to those common deficiencies (Christiansen and Ljungqvist, 2017).

In the present study, we test a non-linear CFR method that belongs to the machine learning family, a Bidirectional Long-Short-Term Neural Network (Bi-LSTM) and that, to our knowledge, has not been applied to CFR yet. We compare the performance of this method to two well-established classical multi-variate linear regression methods, Principal Component Regression (PCR) and Canonical Correlation Analysis (CCA). Traditional CFRs usually assume linear and temporally stable relationships between the local variables captured by the proxy network and the target climate field. Likewise, the spatial patterns of climate variability are considered as stationary (Coats et al., 2013; Pyrina et al., 2017; Wang et al., 2014; Smerdon et al., 2016; Yun et al., 2021). However, links between climate fields can be non-linear (Schneider et al., 2018; Dueben and Bauer, 2018; Huntingford et al., 2019; Nadiga, 2020). Nonlinear machine leaning-based CFR methods (for

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

handles information from both forward and backward temporal directions (Graves and Schmidhuber, 2005). It has been demonstrated that the Bi-LSTM model can achieve better performance for some classification and prediction tasks (Su et al., 2021; Biswas and Sinha, 2021; Biswas et al., 2021). Since climate dynamics usually exhibit temporal dependencies, the Bi-LSTM method might learn these dependencies better, which can provide another advantage to capture the time evolution of the reconstructed climate field. To our knowledge, Bi-LSTM method is applied for the first time in the context of paleo CFRs. In this evaluation of climate reconstruction methods, we focus on the whole Northern Hemisphere temperature field and on the temperature field of the North Atlantic European region. In the North Atlantic region, the most important mode of temperature variations at longer time series is the Atlantic Multidecadal Variability (AMV). The index of the AMV is defined as decadal filtered surface temperature anomaly over North Atlantic regions (95°W–30°E, 0–70°N, excluding the Mediterranean and Hudson Bay following Knight et al., 2006). It has been shown that AMV is related to many prominent examples of regional or even hemispheric multidecadal climate variability, for example European and North America summer climate variability (Knight et al., 2006; Qasmi et al., 2017). In this context, we test the reconstruction skill for the spatial resolved summer seasonal temperature anomalies over NH and NAE, as well as for the spatially averaged AMV and NH summer temperature anomalies, calculated from the spatially resolved reconstructed fields. The reconstruction of mean temperature series could provide a general assessment of the skill to reconstruct extreme temperature phases (e.g. related to volcanic eruptions or changes in solar activity) serving as benchmarks to test the potential capability of different CFR methods on those anomalies.

Regarding the networks of real proxies used so far, St. George and Esper (2019) reviewed contemporary studies on previous NH temperature reconstructions based on tree ring proxies (Mann et al., 1998, 2008, 2007, 2009a, 2009b; PAGES2k Consortium). St. George and Esper concluded that the present-day generation of tree ring proxy based reconstructions exhibit high correlations with seasonal hemispheric summer temperatures and display relatively better skills in tracking year-to-year climatic variabilities and decadal fluctuations than former proxy networks, as found by Wilson et al., (2016) and Anchukaitis et al., (2017). Thus, we test NH summer temperature CFRs employing a pseudo-proxy continental network that is the result of blending two networks: the PAGES2k Consortium multiproxy network, and the climate-tree-ring network of St. George (2014).

In the oceanic realm in the North Atlantic, additional marine proxy records based on mollusc shell bands (Pyrina et al., 2017) have been also used for climate reconstructions. These records, similarly to the dendroclimatological records, are based on annual growth bands, are annually resolved, and usually represent surface or subsurface water temperature. Compelling evidence has already been provided by earlier studies that Atlantic Ocean variability is an important driver of European summer climate variability (Jacobeit et al., 2003; Sutton and Hodson, 2005; Folland et al., 2009). Thus, we also employ an updated proxy network by combining the locations of marine proxies and tree ring proxies (Pyrina et al., 2017; PAGES 2k consortium, 2017; Luterbacher et al., 2016) to test the NAE summer seasonal temperature reconstructions.

The choice of climate model to run pseudo-experiments may plausibly have an impact on the estimation of method skills, since the spatial and temporal cross-correlations between climate variables may be model dependent. Thus, it is advisable to use

instance, Artificial Neural Networks-ANN) could help capture underlying linear and nonlinear relationships between proxy records and the large-scale climate more possible (Rasp and Lerch, 2018; Schneider et al., 2018; Rolnick et al., 2019; Huang et al., 2020; Nadiga, 2020; Chattopadhyay et al., 2020; Lindgren et al., 2021). Moreover, machine-learning methods do not necessarily rely on statistical methods to first obtain the principal spatial climate patterns, such as Principal Component Analysis-PCA. The full inherent variability in the original dataset is sequentially and dynamically adjusted and captured with optimized hyper-parameters during the model training process (Goodfellow et al., 2016).

Within the family of machine learning methods, recurrent neural networks (RNN) and Long Short-Term Memory networks (LSTM) are characterized by specifically incorporating the sequential structure of the predictors to estimate the predictand (Bengio et al. 1994). This property makes them promising methods to ameliorate the underestimation of variability that affects many other methods. Our assumption here is that the methods would be able to better capture episodes of larger deviations from the mean, especially those that stretch over several time steps. However, this assumption is not guaranteed to be realistic in practical situations and needs to be tested. The classical recurrent neural network and Long Short-Term Memory Network can usually only receive and process information from prior forward inference steps. A variant of the LSTM network is the bidirectional Bi-LSTM. It handles information from both forward and backward temporal directions (Graves and Schmidhuber, 2005). It has been demonstrated that the Bi-LSTM model is capable of learning and capturing long-term dependencies from a sequential dataset (Hochreiter and Schmidhuber, 1997) and that it achieves better performance for some classification and prediction tasks (Su et al., 2021; Biswas and Sinha, 2021; Biswas et al., 2021). Since climate dynamics usually exhibit temporal dependencies, the Bi-LSTM method might learn these dependencies better, which can provide another advantage to capture the time evolution of the reconstructed climate field.

The Bi-LSTM combines two independent LSTMs together, which allows the network to incorporate both backward and forward information for the sequential time series at every time step. Our working hypothesis is, that a more sophisticated type of RNN could better replicate the past variability, and perhaps even more so for extreme values. Thus, we would like to test whether this property of the Bi-LSTM is useful for paleo climate research in the future based on our experiments, especially by employing only a limited calibration/training dataset that could also be a challenge for training deep neural networks (Najafabadi et al. 2015).

This calibration period, which is usually chosen in the real reconstructions as the observational period (or the overlap period between observations and proxy records) can represent a challenge not only for a parameter-rich method such as the Bi-LSTM, but also for the usual linear methods. For instance, a global or hemispheric proxy network may span of the order of 100 sites, and a regional proxy network may span a few tenths of sites. If the calibration period spans at most 150 independent time steps, a method like Principal Component Regression, in which one principle component is predicted by the whole proxy network, is rather close to overfitting conditions, especially in a global or hemispheric case. Canonical Correlation Analysis with a PCA-prefiltering would be much more robust to the potential overfitting if only a few leading PCs are retained in the prefiltering step (sse Methods). Here, we test the methods in our pseudo-proxy experiments in the conditions as they are usually applied in real reconstructions, in which overfitting may be a real risk.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

several 'numerical laboratories' and employ several comprehensive Earth System Models (ESMs) to evaluate reconstructions methods. Besides, constructing PPEs based on different ESMs will highlight model-based impacts on the reconstructed magnitude and spatial patterns (Smerdon et al., 2011, Smerdon, 2012; Amrhein et al., 2020). Accordingly, in this study three different comprehensive Earth System Models are employed as 'surrogate climate database for setting up PPEs: the Community Climate System Model CCSM4, the Max-Planck-Institute climate model MPI-ESM-P and the Community Earth System Model CESM1-CAM5. Among the models providing climate simulations of the past millennium, these three models are the ones with the highest horizontal resolution.

2 Data and Method

2.1 Data

2.1.1 Proxy data locations

The pseudoproxies are constructed from the simulated grid-cell summer mean temperature sampled from three climate model simulations over the past millennium (see following subsections). In this context, 11 real proxy locations in the North Atlantic-European region (Pyrina et al., 2017; PAGES 2k consortium, 2017; Luterbacher et al., 2016) are selected for regional NAE (60°W-30°E, 0-88°N) PPEs and 48 proxy locations across the Northern Hemisphere are chosen from the PAGES 2k network. The original North Hemisphere PAGES network was trimmed down by removing proxies that may show a combined temperature-moisture response and by selecting only one proxy among those deemed to be too closely located (and thus redundant from the climate model perspective). Specifically, the 48 dendrochronology locations are selected according to Figure 4 of St. George, S. (2014) which shows the correlation coefficient between the dendroclimatological proxy records and summer season temperature. At most of the retained locations, the correlation between the dendroclimatological record and regional temperature is higher than 0.5.

2.1.2 Climate Models

One of the climate models utilized in our study is the Max-Planck-Institute Earth System model MPI-ESM-P with a spatial horizontal resolution of about 1.9 degree in longitude and 1.9 degree in latitude. The simulation covers the period from 100 BC to 2000 CE. The model MPI-ESM-P consists of the spectral atmospheric model ECHAM6 (Stevens et al., 2013), the ocean model MPI-OM (Jungclaus et al., 2013), the land model JSBACH (Reick et al., 2013) and the bio-geophysical model HAMOCC (Ilyina et al., 2013). The setup of our simulations corresponds to the MPI-ESM-P LR setup in the CMIP5 simulations suite. However, since the present simulations does not belong to the CMIP5 project, the forcings used in this simulation and additional technical details are shown in the Appendix A.

The second climate model is the Community Earth System Model CESM1 Paleoclimate model CAM5 from the National Centre for Atmospheric Research (NCAR) (Otto-Bliesner et al., 2016) with a spatial resolution of 2.5 degree in longitude and

For the sake of completeness, we briefly mention here reconstruction methods that combine information from proxy and from climate simulations - data assimilation (Steiger et al., 2014; Carrassi et al., 2018). The family of data-assimilation methods constrain or modify the spatially complete output of climate simulations conditional on the available locally sparse information provided by proxy records. Therefore, they are not so strongly constrained, in principle, by the assumption that the spatial covariance is stationary over time. Another advantage is that they provide estimation of reconstruction uncertainties in a more straightforward way, especially those methods formally based on a Bayesian framework. On the other hand, the underlying data-assimilation equations do require the estimation of large cross-covariance matrices, e.g., based on Kalman Filters, and this usually makes necessary the application of some sort of, subjective, regularization of the error-covariance matrices (Harlim, 2017; Janjić et al., 2018). They also might be computationally much more demanding than purely data-driven methods. Considering the replication of the amplitude of past variations, it depends on factors that are independent of the method itself, such as the variance generated by the climate model and also on the inherent uncertainties of the proxy data. Therefore, an under- or overestimation of reconstructed variance cannot be as characterized as a systemic property of these methods. They have the very important advantage in that they combine all the available information about past climate (simulations, forcings, proxy data) into a powerful tool. However, the prior use of information from climate models precludes a posterior critical comparison between simulations and reconstructions, and thereby the resulting reconstructions lose one appeal of climate reconstructions in general.

In this evaluation of three climate reconstruction methods, we focus on the whole Northern Hemisphere temperature field and on the temperature field of the North Atlantic European region. In the North Atlantic region, the most important mode of temperature variations at longer time series is the Atlantic Multidecadal Variability (AMV). The index of the AMV is defined as decadal filtered surface temperature anomaly over North Atlantic regions 95°W–30°E, 0–70°N, excluding the Mediterranean and Hudson Bay following Knight et al. (2006). It has been shown that AMV is related to many prominent features of regional or even hemispheric multidecadal climate variability, for example European and North America summer climate variability (Knight et al., 2006; Qasmi et al., 2017). In this context, we test the reconstruction skill for the spatial resolved summer temperature anomalies over Northern Hemisphere-NH (180°W-180°E, 0-90°N) and North Atlantic European region-NAE (60°W-30°E, 0-88°N), as well as for the spatially averaged AMV and NH summer temperature anomalies, calculated from the spatially resolved reconstructed fields. The reconstruction of mean temperature series could provide a general assessment of the skill to reconstruct extreme temperature phases (e.g. related to volcanic eruptions or changes in solar activity) serving as benchmarks to test the potential capability of different CFR methods on those anomalies.

Regarding the networks of real proxies used so far, St. George and Esper (2019) reviewed contemporary studies on previous NH temperature reconstructions based on tree ring proxies (Mann et al., 1998, 2008, 2007, 2009a, 2009b; Emile-Geay et al., 2017). St. George and Esper (2019) concluded that the present-day generation of tree-ring proxy-based reconstructions exhibit high correlations with seasonal hemispheric summer temperatures and display relatively better skill in tracking year-to-year climatic variabilities and decadal fluctuations than former proxy networks, as found by Wilson et al., (2016) and

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

1.9 degree in latitude (https://www.cesm.ucar.edu/projects/community-projects/LME/). The CESM-CAM5-LME simulation extends from 850 CE to 2006 CE using CMIP5 climate forcing reconstructions (Schmidt et al. 2011) and reconstructed forcing for the transient evolution of aerosols, solar irradiance, land use conditions, greenhouse gases, orbital parameters, and volcanic emissions. The atmosphere model employed in CESM1 is CAM5 (Hurrell et al., 2013), which is a significant advancement of CAM4 (Neale et al., 2013), whereas CCSM4 uses CAM4 as its atmospheric component. The CESM1-CAM5 uses the same ocean, land and sea ice models as CCSM4 (Hurrell et al., 2013) does. We use one simulation form the Last Millennium Ensemble (LME).

The third climate model is the Community Climate System Model CCSM4 model (Gent et al., 2011), also from NCAR, which uses the land (CLM4/ Lawrence et al., 2012), the ocean model(POP2/Smith et al., 2010)the atmosphere (CAM4/Neale et al., 2013) and the sea ice model (CICE4/Hunke et al., 2008) components. CCSM4 is modelled with a spatial resolution of 1.25 degree in longitude and 0.9 degree in latitude. Here, we use the simulations labelled past1000 and historical and labelled r1i1p1 from the CMIP5 data set. The past1000 simulation spans the period from 850 CE to 1849 CE. The historical simulations covers the period 1850 CE to 2005 CE. We concatenate both simulations together for this study. The boundary conditions and forcings follow the PMIP3 (Paleoclimate Modelling Intercomparison Project Phase III) protocols (Schmidt et al. 2011). The Lean et al. (2005) reconstruction of the total solar irradiance (TSI) combined with the Vieira et al. (2011) is employed to describe the TSI. The volcanic forcing is prescribed by using the ice-core-based index of Gao et al. (2008). The Pongratz et al. (2008) reconstruction of land use change merged with that of Hurtt et al. (2009) is used to describe seamlessly land use changes.

2.2 Methods

2.2.1 Construction of pseudo-proxies

To test the statistical reconstruction methods in the virtual laboratories of climate model simulations, we need records that mimic the statistical properties of real proxy records. The most important properties are their correlation to the local temperature and their location in a proxy network. A third important characteristic is the network size and temporal coverage. The usual method to produce pseudo-proxy records in climate simulations is to sample the simulated temperature at the grid-cell co-located with the real proxy record and contaminate the simulated temperature with added statistical noise, so that the correlations between the original temperature and the contaminated temperature resembles the typical temperature-proxy correlations. The real correlation is of the order of 0.5 or above for good proxy records. This parameter can be modulated in the pseudo-proxy record by the amount of noise added to the simulated temperature, and different proxy networks will help us to reveal how and to what extent degradations and effects would be expressed amongst methods and amongst climate models. Ideal pseudo-proxies contain only the temperature signal subsampled from the climate model. We then perturb the ideal pseudo-proxies with Gaussian white noise. Gaussian white noise with signal-to-noise ratio (SNR) values of 0.25, 0.5 and 1 is in general widely employed for contaminating the ideal pseudoproxy dataset (Smerdon, 2012; Wang et al., 2014). The noise level can be defined using various criteria including SNR, variance of pure white noise (NVAR), and percent noise by variance

Anchukaitis et al., (2017). Thus, we test NH summer temperature CFRs employing a pseudo-proxy continental network that is the result of blending two networks: the PAGES2k Consortium (Emile-Geay et al., 2017) multiproxy network, and the climate-tree-ring network of St. George (2014).

In the oceanic realm in the North Atlantic, additional marine proxy records based on mollusc shell bands (Pyrina et al., 2017) have been also used for climate reconstructions. These records, similarly to the dendroclimatological records, are based on annual growth bands, are annually resolved, and usually represent surface or subsurface water temperature. Therefore, they are technically rather similar to dendroclimatological records. Compelling evidence has already been provided by earlier studies that Atlantic Ocean variability is an important driver of European summer climate variability (Jacobeit et al., 2003; Sutton and Hodson, 2005; Folland et al., 2009). Thus, we also employ an updated proxy network by combining the locations of marine proxies and tree ring proxies (Pyrina et al., 2017; Emile-Geay et al., 2017; Luterbacher et al., 2016) to test the NAE summer temperature reconstructions.

The choice of climate models to run pseudo-experiments will have an impact on the estimation of method skills (Smerdon et al., 2011, 2015; Parsons et al., 2021), since the spatial and temporal cross-correlations between climate variables are usually model dependent. Thus, it is advisable to use several 'numerical laboratories' and employ several comprehensive Earth System Models (ESMs) to evaluate reconstructions methods. Constructing PPEs based on different ESMs will highlight model-based impacts on the reconstructed magnitude and spatial patterns (Smerdon et al., 2011, Smerdon, 2012; Amrhein et al., 2020). Accordingly, in this study two different comprehensive Earth System Models are employed as 'surrogate climate database for setting up PPEs: the Max-Planck-Institute Earth System Model model MPI-ESM-P and the Community Earth System Model CESM1-CAM5.

2 Data and Methods

2.1 Data

2.1.1 Proxy data locations

The pseudoproxies are constructed from the simulated grid-cell summer mean temperature sampled from two climate model simulations over the past millennium (see following subsections). In this context, 11 real proxy locations in the North Atlantic-European region (Pyrina et al., 2017; Emile-Geay et al., 2017; Luterbacher et al., 2016) are selected for regional NAE (60°W-30°E, 0-88°N) PPEs and 48 proxy locations across the Northern Hemisphere are chosen from the PAGES 2k network. The original Northern Hemisphere PAGES network was trimmed down by removing proxies that may show a combined temperature-moisture response and by selecting only one proxy among those deemed to be too closely located (and thus redundant from the climate model perspective). Specifically, the 48 dendrochronology locations are selected according to Figure 4 of St. George, (2014) which shows the correlation coefficient between the dendroclimatological proxy

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

(PNV) and so on (Smerdon, 2012). We employ here the PNV to define the noise level convention: The PNV expresses the ratio between the added noise variance and the total variance of resulting the pseudo-proxy time series.

Although individual real proxies contain different amounts of noise (non-climatic variability), we assume here a uniform level of noise throughout the whole pseudo-proxy network. In addition, real proxy records contain temporal gaps, and not all records span the same period. For the sake of simplicity, we assume in our pseudo-proxies network that the data have no temporal gaps and all records cover the whole period of the simulations

The dataset employed herein for constructing PPEs database is split into a calibration period that spans 850-1425AD, and a validation period that spans 1426-2000 AD. All the validation statistics of the CFR results are derived against the reconstruction period of 1426-2000 AD. Note that here we just split the entire temporal interval into two equal parts, while, in reality, the instrumental period only covers the most recent 150 years. It means that in reality, only the most recent 150 years can be used for training of the statistical models. The reconstruction skill derived in this study will, all other factors being equal, be larger than the real reconstruction skill, since the data available for calibration are less and the calibration period contains different climate regimes, e.g. the strong anthropogenic warming signal which is not present in the preindustrial period. On the other hand, we exclude the period of strong anthropogenic warming from the calibration data, so that we present to the methods the clear challenge of reconstructing the temperature beyond the range of variations seen in the calibration period, expecting that the possible deficiencies of the reconstruction methods become more apparent with this set-up.

2.2.2 Principal component regression

Principal component analysis is employed to construct a few new variables that are a linear combination of the components of the original climate field, and that ideally describe a large part of the total variability. The linear combinations that define the new variables are the eigenvectors of the cross-covariance matrix of the field. Associated to each variable (eigenvector), a principal component time series (scores) describes its temporal variation. In the PCR, the predictands are those scores identified by PCA of the climate field (Hotelling, 1957; Luterbacher et al., 2004; Pyrina et al., 2017). This results in a reduction of dimensionality without losing too much information, and reduces the risk of over-fitting. In the present study, the retained PCs capture at least 90% of the cumulative temporal variance of climate field. After selecting the empirical orthogonal functions-EOFs and principal components-PCs based on the calibration dataset and establishing the desired linear regression relationships between the PCs and the proxy dataset (predictors), the PCs in the validation period are reconstructed using the estimated regression coefficients. The full climate field is then reconstructed by the linear combination of the reconstructed PCs and their corresponding EOFs. A given climate field x_t, at time step t can be decomposed as follows:

$$x_{m,t} = \sum_{n=1}^{k} PC_{n,t}\, EOF_{m,n} \quad (1)$$

where m is the grid index of the field, t is the time index, and k denotes the total numbers of retained PCs.

The linear relationship between proxies and targeted climate field is established by the regression equation:

$$PC_{n,t} = \sum_{m=1}^{J} \omega_{n,m} Proxy_{m,t} + \varepsilon \quad (2)$$

records and summer temperature. At most of the retained locations, the correlation between the dendroclimatological record and regional temperature is higher than 0.5.

2.1.2 Climate Models

One of the climate models utilized in our study is the Max-Planck-Institute Earth System model MPI-ESM-P with a spatial horizontal resolution of about 1.9 degree in longitude and 1.9 degree in latitude. The simulation covers the period from 100 BC to 2000 CE. The model MPI-ESM-P consists of the spectral atmospheric model ECHAM6 (Stevens et al., 2013), the ocean model MPI-OM (Jungclaus et al., 2013), the land model JSBACH (Reick et al., 2013) and the bio-geophysical model HAMOCC (Ilyina et al., 2013). The setup of our simulations corresponds to the MPI-ESM-P LR setup in the CMIP5 simulations suite. However, since the present simulations does not belong to the CMIP5 project, the forcings used in this simulation and additional technical details are shown in the Appendix A.

The second climate model is the Community Earth System Model CESM1 Paleoclimate model CAM5 from the National Centre for Atmospheric Research (NCAR) (Otto-Bliesner et al., 2016) with a spatial resolution of 2.5 degree in longitude and 1.9 degree in latitude (https://www.cesm.ucar.edu/projects/community-projects/LME/). The CESM-CAM5-LME simulation extends from 850 CE to 2006 CE using CMIP5 climate forcing reconstructions (Schmidt et al. 2011) and reconstructed forcing for the transient evolution of aerosols, solar irradiance, land use conditions, greenhouse gases, orbital parameters, and volcanic emissions. The atmosphere model employed in CESM1 is CAM5 (Hurrell et al., 2013), which is a significant advancement of CAM4 (Neale et al., 2013), whereas CCSM4 uses CAM4 as its atmospheric component. The CESM1-CAM5 uses the same ocean, land and sea ice models as CCSM4 (Hurrell et al., 2013) does. We use the last one ensemble simulation member 13 from the Last Millennium Ensemble (LME).

2.2 Methods

2.2.1 Construction of pseudo-proxies

To test the statistical reconstruction methods in the virtual laboratories of climate model simulations, we need records that mimic the statistical properties of real proxy records. The most important properties are their correlation to the local temperature and their location in a proxy network. A third important characteristic is the network size and temporal coverage.

The usual method to produce pseudo-proxy records in climate simulations is to sample the simulated temperature at the grid cell that contains the proxy location and contaminate the simulated temperature with added statistical noise, so that the correlations between the original temperature and the contaminated temperature resembles the typical temperature-proxy correlations. The real correlation is of the order of 0.5 or above for good proxy records. This parameter can be modulated in the pseudo-proxy record by the amount of noise added to the simulated temperature, and different proxy networks will help

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

where the index m runs over the proxies, j denotes the total numbers of proxies, ω is the linear function coefficient, and ε denotes a residual term. The ω parameters are estimated by Ordinary Least Squares. Here, it is assumed that climate sensitive proxies are linearly related with the climate PCs. Based on Eq. (3) using the PCR method, the PCs during the validation interval will be reconstructed assuming that the linear coefficients derived in Eq. (3) are constant in time:

$$\widehat{PC}_{n,t} = \sum_{m=1}^{j} \omega_{n,m} Proxy_{m,t} \quad (3)$$

The final reconstructed field $\hat{x}$ will be derived by the linear combination of the reconstructed $\widehat{PC}$ with the EOFs derived from the calibration dataset, thereby assuming that the EOF patterns remain constant in time (Gómez-Navarro et al., 2017; Pyrina et al., 2017).

2.2.3 Canonical correlation analysis

Canonical Correlation Analysis CCA is also an eigenvector method. Similarly to PCA, CCA decomposes the variance of the fields as a linear combination of spatial patterns and their corresponding amplitude time series. In contrast to PCA, where the target is to maximize the explained variance with a few new variables, CCA constructs pairs of predictor-predictand variables that maximize the temporal correlation of the corresponding amplitude time series. The pairs of variables are identified by solving an eigenvalue problem that requires the calculation of the inverse of the covariance matrices of each field. These matrices can be pseudo-degenerate (one eigenvalue much smaller than the largest eigenvalue) and therefore the calculation of their inverse is, without regularization, numerically unstable. This regularization can be introduced by first projecting the original fields onto their leading EOFs (Widmann, 2005; Pyrina et al., 2017). This also reduces the number of degrees of freedom - thus hindering overfitting - and eliminate potential noise variance. After the dimensional transformation, a small number of pairs of patterns with high temporal correlation will be retained. In the present study, the number of retained PCs capture at least 90% cumulative variance of predictand climate field. Then these retained PC time series will be used as input variables of CCA to calculate the canonical correlation patterns (CCPs) and canonical coefficients (CCs) time series for both proxy and temperature field. The reconstructed climate field can be calculated by a linear combination of the CCPs with CCs for each time step t:

$$\boldsymbol{x}_{m,t} = \sum_{n=1}^{l} CC_{n,t}^{field} \boldsymbol{CCP}_{m,n}^{field} \quad (4)$$

$$\boldsymbol{Proxy}_{m,t} = \sum_{n=1}^{l} CC_{n,t}^{proxy} \boldsymbol{CCP}_{m,n}^{proxy} \quad (5)$$

Proxy denotes the reconstructed proxy field, and l is the number of CCA pairs. The correlation between each pair CC (proxy, field) are the canonical correlations, which are the square root of the CCA-eigenvalues. Therefore, once each $CC^{proxy}(t)$ is calculated from the proxy data through the validation period, the corresponding $CC^{field}(t)$ can be easily estimated as proportional to $CC^{proxy}(t)$, since the correlation between the different $CC_x^{proxy}(t)$ is zero. The final reconstruction of target climate field will be derived by linear combination of $CCP^{field}(t)$ and $CC^{field}(t)$, assuming again that the dominant canonical correlation patterns of climate variability are stationary in time.

us to reveal how and to what extent degradations of reconstruction skill caused by the amount of non-climatic signals present in the pseudo-proxies.

Ideal pseudo-proxies contain only the temperature signal subsampled from the climate model. We then perturb the ideal pseudo-proxies with Gaussian white noise, and also with red noise for a more realistic noise contamination experiment. We generate two types of pseudoproxies by adding Gaussian white noise and red noise (refer to Pyrina et al., 2017) to the subsampled summer-temperature time series at the tree ring proxy-based locations.

The noise level can be defined using various criteria including signal to noise ratio (SNR), variance of noise (NVAR), and percent of noise by variance (PNV) (Smerdon, 2012; Wang et al., 2014). We employ here the PNV to define the noise level convention. The PNV expresses the ratio between the added noise variance and the total variance of resulting the pseudo-proxy time series. Without loss of generalization we assume that the ideal proxy has unit variance, and thus

$$PNV = NVAR / |1 + NVAR| \quad (1)$$

Red noise for a specific PNV could be defined by:

$$Red_t = \alpha_1 Red_{t-1} + White_t \quad (2)$$

where Red_t represents red noise time series, α_1 indicates the damping coefficient, here in our study it is equal to 0.4 (Larsen and MacDonald, 1995; Büntgen et al., 2010; Pyrina et al., 2017), and $White_t$ is a random white noise time series correspondingly.

Although individual real proxies contain different amounts of noise (non-climatic variability), we assume here an uniform level of noise throughout the whole pseudo-proxy network. In addition, real proxy records contain temporal gaps, and not all records span the same period. For the sake of simplicity, we assume in our pseudo-proxies network that the data have no temporal gaps and all records cover the whole period of the simulations.

The dataset employed here for constructing the according PPEs database is split into a calibration period that spans 1900-1999AD, and a validation period that spans 850-1899 AD. This calibration period would represent the typical period of calibration of real proxy records. All the validation statistics of the CFR results are derived against the reconstruction period of 850-1899 AD.

2.2.2 Principal component regression

Principal component analysis is employed to construct a few new variables that are a linear combination of the components of the original climate field, and that ideally describe a large part of the total variability. The linear combinations that define the new variables are the eigenvectors of the cross-covariance matrix of the field. Associated to each variable (eigenvector), a principal component time series (scores) describes its temporal variation. In the PCR, the predictands are those scores identified by PCA of the climate field (Hotelling, 1957; Luterbacher et al., 2004; Pyrina et al., 2017). This results in a reduction of dimensionality without losing too much information, and reduces the risk of over-fitting. In the present study,

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

2.2.4 Bidirectional Long Short-term memory neural network

As a non-linear machine learning method, we test here a Bidirectional Long short-term memory neural network (Bi-LSTM). The LSTM networks, in contrast to the more traditional neural networks, also capture the information of the serial co-variability present in the data, and therefore are suitable to tackle data with a temporal structure. They are usually applied to the analysis of speech and of time series. The rationale of using these type of networks for climate reconstructions is the aforementioned underestimation of the past climate variations by most linear methods ('regression to the mean'). In principle, a LSTM network could exploit the temporal autocorrelation present in the time series to ameliorate this underestimation and perhaps also provide more realistic spectral properties of the reconstructed time series.

The structure of LSTM network is more complicated than the structure of a traditional neural network. The LSTM estimates a hidden variable $h(t)$ that encapsulates the state of the system at time t. The computation of the new system state at time $t+1$, $h(t+1)$, depends on the value of the predictors at $t+1$ but also on the value of the hidden state at time t, $h(t)$. The training of the LSTM can be accomplished sequentially by assimilating the information present in the training data from time steps in the past of the present time step. In some loose sense, a LSTM network would be the machine-learning equivalent of a linear auto-regressive process.

The training of the network is accomplished by feeding it with sequential data iteratively, forwards towards the future and backwards towards the past. Both forward and backward assimilations are processed by two separated LSTM neural layers, which are connected to the same output layer. Figure 1 illustrates the bidirectional structure of the Bi-LSTM network. Given a set of predictor-predictand variables ($\boldsymbol{X}_t$, $\boldsymbol{Y}_t$), our goal is to train a nonlinear function:

$$\widetilde{\boldsymbol{Y}}_t = F(\boldsymbol{X}) \tag{6}$$

The structure of this complex non-linear function F is defined as follows:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + B_f\right) \tag{7}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + B_i) \tag{8}$$

$$A_t = tanh(W_A[h_{t-1}, x_t] + B_A) \tag{9}$$

$$C_t = f_t C_{t-1} + i_t A_t \tag{10}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + B_o) \tag{11}$$

$$h_t = o_t tanh(C_t) \tag{12}$$

where W_f, W_i, W_A and W_o represent several weight matrices and B_f, B_i B_A and B_o represent different bias matrices. σ is the gate activation function, here we utilize the Rectified Linear Unit function-ReLU (Ramachandran et al., 2017) .

At time step $t-1$, the hidden state of LSTM cell's hidden layer is preserved as h_{t-1} , and this vector is combined with the vector of current input variables $\boldsymbol{X}_t$ to obtain the state of the forget gate, f_t (equation 7) , the input gate i_t (equation 8) and the state of

the retained PCs capture at least 90% of the cumulative temporal variance of climate field. After selecting the empirical orthogonal functions-EOFs and principal components-PCs based on the calibration dataset and establishing the desired linear regression relationships between the PCs and the proxy dataset (predictors), the PCs in the validation period are reconstructed using the estimated regression coefficients. The full climate field is then reconstructed by the linear combination of the reconstructed PCs and their corresponding EOFs. A given climate field x_t, at time step t can be decomposed as follows:

$$x_{m,t} = \sum_{n=1}^{k} PC_{n,t}\, EOF_{m,n} \tag{3}$$

where m is the grid index of the field, t is the time index, and k denotes the total numbers of retained PCs.

The linear relationship between proxies and targeted climate field is established by the regression equation:

$$PC_{n,t} = \sum_{m=1}^{j} \omega_{n,m} Proxy_{m,t} + \varepsilon \tag{4}$$

where the index m runs over the proxies, j denotes the total numbers of proxies, ω is the linear function coefficient, and ε denotes a residual term. The residual could be an unobserved random variable that adds noise to the linear relationship between the dependent variable (PC) and the targeted regressors (proxy or pseudoproxy) and includes all effects on the targeted regressors not related to the dependent variable (Christiansen, 2011).

The ω parameters are estimated by Ordinary Least Squares. Here, it is assumed that climate sensitive proxies are linearly related with the climate PCs. Based on Eq. (5) using the PCR method, the PCs during the validation interval will be reconstructed assuming that the linear coefficients derived in Eq. (5) are constant in time:

$$\widehat{PC}_{n,t} = \sum_{m=1}^{j} \omega_{n,m} Proxy_{m,t} \tag{5}$$

The final reconstructed field $\hat{x}$ will be derived by the linear combination of the reconstructed $\widehat{PC}$ with the EOFs derived from the calibration dataset, thereby assuming that the EOF patterns remain constant in time (Gómez-Navarro et al., 2017; Pyrina et al., 2017).

2.2.3 Canonical correlation analysis

Canonical Correlation Analysis CCA is also an eigenvector method. Similarly to PCA, CCA decomposes the variance of the fields as a linear combination of spatial patterns and their corresponding amplitude time series. In contrast to PCA, where the target is to maximize the explained variance with a few new variables, CCA constructs pairs of predictor-predictand variables that maximize the temporal correlation of the corresponding amplitude time series. The pairs of variables are identified by solving an eigenvalue problem that requires the calculation of the inverse of the covariance matrices of each field. These matrices can be pseudo-degenerate (one eigenvalue much smaller than the largest eigenvalue) and therefore the

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

memory cell A_t (equation 9). This memory cell state A_t is linearly combined with the previous state of the cell output C_{t-1} to update the value of its state. The weights of this linear combinations are the states of the forget gate f_t and of the input gate i_t (equation 10). The state of the output gate o_t is calculated from the previous hidden state and the current input variables (equation 11). This output is used to compute the updated hidden state h_t using the state of the cell output C_t (equation 12) (Huang et al., 2020; Chattopadhyay et al., 2020).

In the present application to climate reconstructions, we have a set of input pseudoproxy data $\boldsymbol{X}_t^m = [x_{t-n}, \ldots, x_{t-1}]$ and an output target temperature time series $\boldsymbol{Y}_t^m = [y_{t-n}, \ldots, y_{t-1}]$. The forward LSTM hidden state sequence $\overrightarrow{\boldsymbol{h}_t}$ (note the arrow direction) is calculated employing inputs information in a positive direction from time t-1 to time t-n iteratively, and for backward LSTM cell, the hidden state sequence $\overleftarrow{\boldsymbol{h}_t}$ is computed using the input within a reverse direction from time t-n to time t-1 iteratively. The final outputs from the forward and backward LSTM cells are calculated utilizing the calculation equation (Cui et al., 2018, Jahangir et al., 2020):

$$\overline{\boldsymbol{Y}}_t = concat\left(\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}\right) \tag{13}$$

where *concat* is the function used to concatenate the two output sequences $\overrightarrow{\boldsymbol{h}}$ and $\overleftarrow{\boldsymbol{h}}$ (Cui et al., 2018, Jahangir et al., 2020). During training process, the calibration dataset are fed into LSTM cell, and it will map the potential latent relationships (both linear and nonlinear) between input and output variables by updating its weight and threshold matrices. The object function for Bi-LSTM to be minimized during training is the Huber loss that expresses the mismatch between the reconstructed climate field and the 'real' climate field from model simulations. Object loss is optimized by gradient descent (Goodfellow et al., 2016). Huber loss has a key advantage of being less sensitive to outlier values:

$$L_\delta\left(\boldsymbol{Y}, f(\boldsymbol{X})\right) = \begin{cases} \frac{1}{2}\left(\boldsymbol{Y} - f(\boldsymbol{X})\right)^2 \\ \delta|\boldsymbol{Y} - f(\boldsymbol{X})| - \frac{1}{2}\delta^2 \end{cases} \tag{14}$$

where f denotes the neural network and the brackets denote the Euclidean norm. The Huber loss function changes from a quadratic to linear when δ (a positive real number) varies from small to big (Meyer, 2020). Huber loss will approach L2 loss when δ tends to be 0, and approach L1 when δ tends to be positive infinity, here we test its value and finally set δ 1.35. L2 is the square root of the sum of squared deviations and L1 is the sum of absolute deviations.

The main mechanism of LSTM is that the LSTM block manages to develop a regulated information flow by controlling which proportion of data information should be 'remembered' or should be 'forgotten. By controlling the regulation of the information flow, LSTM will manage to learn and preserve temporal characteristics and dependencies of the specific time series.

In this particular application, we have used two hidden Bi-LSTM layers with 700 neurons each (some additional tests on selecting the number of neurons in the hidden layer are shown in Appendix B). The optimization function is Adam with the learning rate of 0.0001 (Kingma and Ba, 2014). After training the model, the optimal network hyper-parameters, including

calculation of their inverse is, without regularization, numerically unstable. This regularization can be introduced by first projecting the original fields onto their leading EOFs (Widmann, 2005; Pyrina et al., 2017). This also reduces the number of degrees of freedom - thus hindering overfitting - and eliminate potential noise variance. After the dimensional transformation, a small number of pairs of patterns with high temporal correlation will be retained. In the present study, the number of retained PCs capture at least 90% cumulative variance of predictand climate field. Then these retained PC time series will be used as input variables of CCA to calculate the canonical correlation patterns (CCPs) and canonical coefficients (CCs) time series for both proxy and temperature field. The reconstructed climate field can be calculated by a linear combination of the CCPs with CCs for each time step t:

$$x_{m,t} = \sum_{n=1}^{l} CC_{n,t}^{field}\, CCP_{m,n}^{field} \tag{6}$$

$$Proxy_{m,t} = \sum_{n=1}^{l} CC_{n,t}^{proxy}\, CCP_{m,n}^{proxy} \tag{7}$$

Proxy denotes the reconstructed proxy field, and l is the number of CCA pairs. The correlation between each pair CC (proxy, field) are the canonical correlations, which are the square root of the CCA-eigenvalues. Therefore, once each $CC^{proxy}(t)$ is calculated from the proxy data through the validation period, the corresponding $CC^{field}(t)$ can be easily estimated as proportional to $CC^{proxy}(t)$, since the correlation between the different $CC_n^{proxy}(t)$ is zero. The final reconstruction of target climate field will be derived by linear combination of $CCP^{field}(t)$ and $CC^{field}(t)$, assuming again that the dominant canonical correlation patterns of climate variability are stationary in time.

The CCA method maximizes the correlation that can be attained with a linear change of variables, i.e. with a linear combination of the grid-cell series in each of the two fields. In the following, admittedly artificial, example, the resulting canonical correlation can be very high and yet the reconstruction skill in general can remain low. If one grid cell in each of the two fields are very highly correlated to each other (and assuming here no PCA pre-filtering), CCA will pick those two cells as the first CCA pair (i.e., a pattern in each field with very high loadings only on those cells). The rest of the cells will not contribute to the CCA pattern. The reconstruction skill will therefore generally be very low in all those cells, despite the canonical correlation being very high. In general, the reconstruction skill will be a monotonic function of the canonical correlation coefficient and the variance explained by the canonical predictand pattern. If the latter is low, the reconstruction skill will be low in large areas of the predictand field, even when the canonical correlation may be high.

2.2.4 Bidirectional Long Short-term memory neural network

As a non-linear machine learning method, we test here a Bidirectional Long short-term memory neural network (Bi-LSTM). The LSTM networks, in contrast to the more traditional neural networks, also capture the information of the serial co-variability present in the data, and therefore are suitable to tackle data with a temporal structure. These methods are usually applied to the analysis of sequential data, such as speech and time series. The rationale of using these type of networks for

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

weights and threshold matrices, are estimated. The validation pseudoproxy dataset is then used to reconstruct the target temperature field in the validation directly.

3 Results

We evaluate the reconstruction skill of the different methods based on the Pearson correlation coefficient (cc) between each target series and the corresponding reconstructed series, and their Standard deviation ratio (SD ratio, SD ratio = $SD_{reconstruction}/SD_{model}$). All the evaluation metrics are calculated in the validation period from 1426-2000 AD. High values of derived cc indicate better temporal covariance between target and reconstructed results, a high SD ratio denotes that more variance is preserved in the reconstructions. Usually, the reconstructed variance is smaller than the target variance and thus their ratio is usually smaller than unity.

3.1 North Atlantic-Europe CFRs

Fig. 2 illustrates the CFR results for the North Atlantic-European region employing the 11 ideal-noise-free pseudoproxies based on the three CFR methodologies and the three climate model simulations. When comparing the reconstruction skills across these three CFR methods derived with the same climate model (for example, MPI, CAM5 or CCSM correspondingly), all the spatial cc patterns exhibit similarities. This indicates that all three CFR methods show generally reasonable spatial reconstruction skills (mean cc's over the entire NAE are bigger than 0.45). In addition, cc maps show higher values over regions where more pseudoproxies are located. This confirms the well-documented tendency amongst different multivariate linear based regression methods to better reconstruction skill in the sub-regions with denser pseudoproxy sampling than in regions with sparser networks (Steiger et al., 2014; Evans et al., 2014; Wang et al., 2014). The cc pattern of the nonlinear method Bi-LSTM is very similar to that of the linear methods, even though the structure of the statistical models is very different. This shows that the nonlinear method employed herein has the similar tendency as linear models to obtain better reconstruction skill over denser proxy sampling regions.

The picture that emerges from the SD ratio is also very similar for the three methods (Fig. 2). In the regions with a high pseudo proxy density, the SD ratio is high, but outside of the densely sampled areas, all three CFR methods experience a similar degree of variance loss.

More realistic pseudo-proxies are those containing 80% Gaussian white noise contamination. This amount of noise is constructed and added to the ideal temperature signal of the 11 pseudoproxies subsampled from MPI, CAM5 and CCSM models. The corresponding spatial cc and SD ratio patterns are displayed in Fig. 3. Compared to reconstructions with ideal pseudo proxies (Fig. 2), a strong degradation of reconstruction skill amongst all CFR methods occurs over the entire NAE. The degradation is especially profound in the regions where denser pseudoproxies exist (the mean cc is reduced from 0.8 in the ideal PPEs to approximately 0.3 in the noisy PPEs). Weak reconstruction skill (approximately mean cc is 0.3, and SD ratio

climate reconstructions is the hypothesis that a better representation of the serial correlation could ameliorate the aforementioned underestimation of the past climate variations by most data-driven methods ('regression to the mean', Smerdon, 2012).

The structure of LSTM network is more complex than the structure of a traditional neural network. The LSTM estimates a hidden variable $h(t)$ that encapsulates the state of the system at time t. The computation of the new system state at time $t+1$, $h(t+1)$, depends on the value of the predictors at $t+1$ but also on the value of the hidden state at time t, $h(t)$. The training of the LSTM can be accomplished sequentially by assimilating the information present in the training data from time steps in the past of the present time step. In some loose sense, a LSTM network would be the machine-learning equivalent of a linear auto-regressive process.

A Bi-LSTM network, the training of the network is accomplished by feeding it with sequential data iteratively, forwards towards the future and backwards towards the past. Both forward and backward assimilations are processed by two separated LSTM neural layers, which are connected to the same output layer. Figure 1 illustrates the bidirectional structure of the Bi-LSTM network. Given a set of predictor-predictand variables ($\boldsymbol{X}_t$, $\boldsymbol{Y}_t$), our goal is to train a nonlinear function:

$$\widetilde{Y}_t = F(X) \tag{8}$$

where, $\hat{\boldsymbol{Y}}_t = F(\boldsymbol{X}_t)$ is a close as possible to $\boldsymbol{Y}_t$. The similarity between $\hat{\boldsymbol{Y}}_t$ *and* $\boldsymbol{Y}_t$ is defined by a cost function. The structure of this complex non-linear function F is defined as follows:

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + B_f\right) \tag{9}$$

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + B_i\right) \tag{10}$$

$$A_t = \tanh\left(W_A\left[h_{t-1}, x_t\right] + B_A\right) \tag{11}$$

$$C_t = f_t C_{t-1} + i_t A_t \tag{12}$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + B_o\right) \tag{13}$$

$$h_t = o_t \tanh\left(C_t\right) \tag{14}$$

where W_f, W_i, W_A and W_o represent several weight matrices and B_f, B_i B_A and B_o represent different bias matrices. σ is the gate activation function, here we utilize the Rectified Linear Unit function-ReLU (Ramachandran et al., 2017) .

At time step t-1, the hidden state of LSTM cell's hidden layer is preserved as h_{t-1}, and this vector is combined with the vector of current input variables $\boldsymbol{X}_t$ to obtain the state of the forget gate, f_t (equation 9) , the input gate i_t (equation 10) and the state of memory cell A_t (equation 11). This memory cell state A_t is linearly combined with the previous state of the cell output C_{t-1} to update the value of its state. The weights of this linear combinations are the states of the forget gate f_t and of the input gate

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

is 0.6) exists over regions where proxies are available, together with a small surrounding region. These noise contamination results shown in Fig. 3 demonstrate again that the nonlinear method exhibit CFR similarities to the linear methods.

3.2 Northern Hemisphere CFRs

NH summer temperature anomalies reconstructions based on PPEs using three CFR methodologies and the three climate models are displayed in Fig. 4-5.

Table 1. Skill reconstruction statistics for the North Hemisphere mean temperature in the verification period for ideal PPEs. The table shows the result for three CFR methods (PCR, CCA and Bi-LSTM) and three climate models (MPI, CAM and CCSM). The numbers in parenthesis indicate the skill statistics of 80% noise contaminated PPEs.

Method	SD Ratio			cc		
	MPI	CAM	CCSM	MPI	CAM	CCSM
PCR	0.642(0.427)	0.713(0.486)	0.732(0.483)	0.594(0.333)	0.636(0.372)	0.709(0.467)
CCA	0.557(0.388)	0.612(0.457)	0.604(0.443)	0.565(0.333)	0.587(0.370)	0.648(0.457)
Bi-LSTM	0.604(0.375)	0.729(0.425)	0.719(0.412)	0.556(0.319)	0.611(0.371)	0.660(0.456)

The spatial cc maps and SD ratio distributions for the ideal PPEs in NH are shown in Fig. 4. Again, all three CFR methodologies yield relatively similar spatial patterns of skill for each of the climate models employed here. Skilful reconstructions are again achieved over regions with a denser pseudoproxy network (over North American and Eurasia regions). In addition, relatively large SD ratio and high cc values also occur in tropical regions. These regions barely contain any pseudoproxies, indicating that the climate teleconnections between tropics and mid-latitude regions could be responsible for the reconstruction skill in tropical regions.

All derived CFRs suffer from variance losses as shown in Fig. 4 and in Table 1. The spatial distributions of the SD ratio vary between climate models and CFR methodologies. They also are spatially heterogeneous. The most important factor affecting the ratio of standard deviations seem to be again the chosen climate model. For instance, comparing the spatial patterns of SD ratio in Fig. 4 across the three models, more variance in the high-density pseudoproxy sampling and the tropical regions are preserved in the high-resolution CCSM model across all CFR methods, whereas, CAM5 and MPI model suffer from relatively more variance loss over these areas. In addition, the highest mean cc is obtained by all three methods in the CCSM simulation (See Table 1).

The Bi-LSTM and PCR methods exhibit relatively consistent patterns with similar SD ratios. The CCA methodology seems to suffer more strongly from variance losses (see Table 1) over the entire NH compared to PCR and Bi-LSTM.

The NH results obtained from the 80% noise contaminated PPEs are displayed in Fig. 5. The performance deterioration is expected and again significant. The reduction of reconstruction skill also occurs over regions where dense pseudoproxies are located. The nonlinear method Bi-LSTM seems to suffer more strongly from variance losses over the NH (see Table 1),

i_t (equation 12). The state of the output gate o_t is calculated from the previous hidden state and the current input variables (equation 13). This output is used to compute the updated hidden state h_t using the state of the cell output C_t (equation 14) (Huang et al., 2020; Chattopadhyay et al., 2020).

In the present application to climate reconstructions, we have a set of input pseudoproxy data $X_1^n = [x_{t-n}, \ldots, x_{t-1}]$ and an output target temperature time series $Y_1^m = [y_{t-n}, \ldots, y_{t-1}]$. The forward LSTM hidden state sequence $\overrightarrow{h_t}$ (note the arrow direction) is calculated employing input information in a positive direction from time t-n to time t-1 iteratively, and for backward LSTM cell, the hidden state sequence $\overleftarrow{h_t}$ is computed using the input within a reverse direction from time t-1 to time t-n iteratively. The final outputs from the forward and backward LSTM cells are calculated utilizing the calculation equation (Cui et al., 2018, Jahangir et al., 2020):

$$\widetilde{Y}_t = concat\left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{13}$$

where *concat* is the function used to concatenate the two output sequences $\overrightarrow{h}$ and $\overleftarrow{h}$ (Cui et al., 2018, Jahangir et al., 2020).

During training process, the calibration dataset are fed into LSTM cell, and it will map the potential latent relationships (both linear and nonlinear) between input and output variables by updating its weight and threshold matrices. The objective cost function for Bi-LSTM to be minimized during training is the Huber loss that expresses the mismatch between the reconstructed climate field and the 'real' climate field from model simulations. We minimize the loss with gradient descent (Goodfellow et al., 2016). Huber loss has a key advantage of being less sensitive to outlier values:

$$L_\delta\left[Y, f(X)\right] = \begin{cases} \frac{1}{2}\left[Y - f(X)\right]^2 \\ \delta\left|Y - f(X)\right| - \frac{1}{2}\delta^2 \end{cases} \tag{14}$$

where f denotes the neural network and the brackets denote the Euclidean norm. The Huber loss function changes from a quadratic to linear when δ (a positive real number) varies from small to big (Meyer, 2020). Huber loss will approach L2 loss when δ tends to be 0, and approach L1 when δ tends to be positive infinity, here we test its value and finally set δ 1.35. L2 is the square root of the sum of squared deviations and L1 is the sum of absolute deviations.

The main mechanism of LSTM is that the LSTM block manages to develop a regulated information flow by controlling which proportion of information from the past should be 'remembered' or should be 'forgotten' as time advances. By controlling the regulation of the information flow, LSTM will manage to learn and preserve temporal characteristics and dependencies of the specific time series.

Neural network is generally composed of one input layer, several hidden layers and one output layer. Many hyper-parameters in the neural network usually need to be initialized and tuned for obtaining reasonable results within specific tasks, for instance, activation functions in each layer, objective function for minimizing the loss of the network model, and

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

especially over the North Atlantic region (Fig. 5) across all three models. This might indicate that the nonlinear method could be more sensitive to the noise contamination than the other methods.

In addition, a relatively better performance is derived, also in the case of noisy pseudoproxies, from the CCSM using all CFR methods (see the spatial cc patterns shown in Fig. 5 and mean cc values in Table 1).

Considering the general methodological skill, as indicated by the derived cc and SD ratio values in Table 1, the Bi-LSTM method presents relatively worse performance with lower mean cc and SD ratio. The methods PCR and Bi-LSTM generally outperform the CCA methodology with higher mean cc with ideal PPEs. Overall, PCR generally outperforms CCA and Bi-LSTM with highest mean cc and SD ratio values across all PPEs in all three model simulations.

3.3 Northern Hemisphere and AMV indices

The evolution of the decadal NH mean temperature anomalies reconstructed by the three CFR methodologies and using perfect pseudoproxies from three models is illustrated in Fig. 6 on the left panel. All indices have been smoothed using a Butterworth low-pass filter to remove temporal fluctuations shorter than 10 years. The reconstruction performance varies amongst different CFR methodologies.

Table 2. RMSE (K) and cc during the verification interval for decadal NH mean temperature derived from ideal PPEs. The numbers in parenthesis indicate the RMSE (K) and cc of 80% noise contaminated PPEs.

Method	RMSE			cc		
	MPI	CAM	CCSM	MPI	CAM	CCSM
PCR	0.102(0.217)	0.066(0.119)	0.062(0.186)	0.974(0.872)	0.913(0.691)	0.991(0.939)
CCA	0.124(0.226)	0.075(0.120)	0.108(0.200)	0.967(0.867)	0.897(0.669)	0.976(0.931)
Bi-LSTM	0.123(0.239)	0.067(0.116)	0.109(0.222)	0.941(0.797)	0.908(0.698)	0.971(0.877)

Table 3. The same as Table 2, but for decadal AMV index

Method	RMSE			cc		
	MPI	CAM	CCSM	MPI	CAM	CCSM
PCR	0.106(0.209)	0.073(0.114)	0.071(0.175)	0.959(0.847)	0.883(0.691)	0.979(0.933)
CCA	0.129(0.217)	0.078(0.118)	0.117(0.185)	0.946(0.844)	0.875(0.665)	0.959(0.929)
Bi-LSTM	0.123(0.231)	0.081(0.115)	0.098(0.204)	0.927(0.769)	0.843(0.682)	0.969(0.883)

The temporal evolution of the AMV index (the right panel in Fig. 6) differs among the simulations, reflecting the different forcings used in each simulation and the model specific contribution of internal variability to the index variations (Wagner and Zorita, 2005; Schmidt et al., 2011). Considering the methodological performance, all three methods generally achieve good

learning rate for controlling the convergence speed of the network model (Goodfellow et al., 2016). In our specific CFR experiments, we have explored a range of Bi-LSTM architectures, including different network depths, introducing dropout layers, using different learning rates, and employing different loss functions to provide a more comprehensive evaluation of the Bi-LSTM performance and effectiveness (these tests are shown in Appendix B). These hyper-parameters within Bi-LSTM are finally selected and employed based on our experimental tests (Knerr, et, al. 1990; Kingma and Ba, 2014; Yu, et, al. 2019).

3 Results

We evaluate the reconstruction skill of the different methods based on the Pearson correlation coefficient (*cc*) between each target series and the corresponding reconstructed series, and their Standard deviation ratio (SD ratio, SD ratio = $SD_{reconstruction}/SD_{model}$). All the evaluation metrics are calculated in the validation period from 850-1899 AD. High values of derived *cc* indicate better temporal covariance between target and reconstructed results, a high SD ratio denotes that more variance is preserved in the reconstructions.

3.1 North Atlantic-Europe CFRs

Fig. 2 illustrates the CFR results for the North Atlantic-European region employing the 11 ideal-noise-free pseudoproxies based on the three CFR methodologies and the two climate model simulations. When comparing the reconstruction skills across these three CFR methods derived with the same climate model (for example, MPI and CAM5 correspondingly), the spatial *cc* patterns calculated between targets and derived reconstructions amongst three CFR methods generally exhibit similarities. This indicates that all three CFR methods show generally reasonable spatial reconstruction skills (mean *cc*'s over the entire NAE are bigger than 0.4). In addition, *cc* maps in Fig. 2 show higher values over regions with a denser pseudo-proxy network. This confirms the well-documented tendency amongst different multivariate linear based regression methods for better reconstruction skill in the sub-regions with denser pseudoproxy sampling than in regions with sparser networks (Smerdon, 2010, 2011; Steiger et al., 2014; Evans et al., 2014; Wang et al., 2014). The *cc* pattern of the nonlinear method Bi-LSTM is very similar to that of the linear methods, even though the structure of the statistical models is very different. This shows that the nonlinear method employed herein has the similar tendency as linear models to obtain better reconstruction skill over regions with denser proxy sampling.

The picture that emerges from the SD ratio is also very similar for the three methods (Fig. 2). In the regions with a high pseudo proxy density, the SD ratio is high, but outside of the densely sampled areas, all three CFR methods experience a similar degree of interannual variance underestimation. Appendix C displays the ratio of SD after applying a 30-year filter to the reconstructions and the target fields. The underestimation of variance is larger at these time scale, but the overall conclusion for all three methods remains.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

AMV index reconstructions when using perfect pseudo-proxies, as shown in each subfigures of Fig. 6, and also refer the RMSE in Table 3.

The NH and AMV indices derived from more realistic 80% noise contaminated CFRs are shown in Fig. 7. The larger noise contamination results in substantial skill deterioration (RMSD and cc displayed within brackets in Table 2 and 3). All three methods fail to capture the warming trend over the most recent century, and the magnitude of strong cooling events is strongly underestimated.

3.4 Probability distributions of reconstructed variables

Table 4. Kolmogorov-Smirnov test statistic and p-value for quantifying the histogram distributions between model and reconstructed NH decadal means. Low values of the KS statistic indicate larger similarity between the two distributions. The numbers in parenthesis indicate the KS statistic and p-value of 80% noise contaminated PPEs.

Method	KS statistic			p-value		
	MPI	CAM	CCSM	MPI	CAM	CCSM
PCR	0.092(0.213)	0.111(0.246)	0.053(0.128)	1e-2(6e-12)	1e-3(8e-16)	3e-1(1e-3)
CCA	0.111(0.246)	0.142(0.252)	0.072(0.149)	2e-3(8e-16)	1e-5(1e-16)	9e-2(5e-6)
Bi-LSTM	0.135(0.269)	0.126(0.253)	0.125(0.229)	4e-5(8e-19)	1e-4(1e-16)	2e-3(1e-13)

Table 5. The same as Table 4, but for AMV index.

Method	KS statistic			p-value		
	MPI	CAM	CCSM	MPI	CAM	CCSM
PCR	0.116(0.224)	0.123(0.198)	0.046(0.126)	8e-4(4e-13)	3e-4(2-e10)	5e-1(2e-4)
CCA	0.116(0.248)	0.118(0.215)	0.076(0.145)	8e-4(5e-16)	6e-4(4e-12)	6e-2(9e-6)
Bi-LSTM	0.151(0.262)	0.125(0.212)	0.093(0.197)	3e-6(7e-18)	2e-4(12e-12)	1e-2(3e-10)

Even though the three reconstructions methods tend to underestimate the overall variability when using noisy pseudoproxies, an interesting question is their skill in reproducing the probability distributions of the climate indices. In particular, a relevant question is whether the methods are able to capture extreme phases of those indices.

Fig. 8 and 9 display the histogram for decadal NH mean and AMV index, respectively. Each subfigure represents the histograms of reconstructed temperature indices across the three methods, compared with the histograms of the target temperature index.

For perfect pseudoproxies, the PCR reconstruction seems to capture the overall target distribution best. It captures the lower tail better than CCA and the upper tail better than CCA and Bi-LSTM. The differences between the methods become smaller for the reconstructions with noisy pseudo proxies, with the PCR still being better than the other two methods (subfigures for

Gaussian white and red noise is constructed and added to the ideal temperature signal of the 11 pseudoproxies subsampled from the MPI and CAM5 models. The corresponding spatial *cc* and SD ratio patterns are displayed in Fig. 3 and 4 correspondingly. Compared to reconstructions with ideal pseudo proxies (Fig. 2), a strong degradation of reconstruction skill amongst all CFR methods occurs over the entire NAE. The reduction in skill is especially profound in the regions where the pseudo-proxy network is denser. Weak reconstruction skill exists over regions where proxies are available and in within their proximity. These noise contamination results shown in Fig. 3 and 4 demonstrate again that the nonlinear method exhibit CFR similarities to the linear methods, whereas, the Bi-LSTM show relatively worse reconstruction skills, with variance underestimation compared to the other two methods in CAM5 based PPEs (referring to the spatial SD ratio in Fig. 4).

The ratio of reconstructed to target variance after 30-year low-pass filtering is also larger than for the interannual variance, but otherwise the patterns share the same properties with the ratios of interannual SD (not shown for the sake of brevity).

In general, all three CFR methods exhibit similar reconstruction performance. Specifically, better skills over regions where denser pseudoproxies exist indicates that the spatial covariance patterns learned from the training data (in the 20th century) are stationary enough to represent the covariance during the reconstruction period over NAE domain.

3.2 Northern Hemisphere CFRs

NH summer temperature anomalies reconstructions based on PPEs using three CFR methodologies and the three climate models are displayed in Fig. 5-7.

Table 1. Skill reconstruction statistics for the Northern Hemisphere mean temperature in the verification period for ideal PPEs. The table shows the result for three CFR methods (PCR, CCA and Bi-LSTM) and two climate models (MPI and CAM). The numbers in parenthesis indicate the skill statistics of white noise and red noise (italics) contaminated PPEs.

Method	SD Ratio		*cc*	
	MPI	CAM	MPI	CAM
PCR	0.878(0.904/*0.977*)	0.874(0.897/*0.913*)	0.401(0.169/*0.135*)	0.490(0.216/*0.206*)
CCA	0.603(0.706/*0.694*)	0.651(0.750/*0.778*)	0.406(0.165/*0.131*)	0.507(0.229/*0.218*)
Bi-LSTM	0.710(0.689/*0.669*)	0.770(0.714/*0.732*)	0.347(0.145/*0.125*)	0.462(0.210/*0.191*)

The spatial *cc* maps for the ideal PPEs in NH are shown in Fig. 5. Again, all three CFR methodologies yield relatively similar spatial *cc* patterns of skill for each of the climate models employed here. Skilful reconstructions are again achieved over regions with a denser pseudoproxy network (over North American and Eurasia regions). In addition, relatively high *cc* values also occur in tropical regions. A relatively high-reconstructed skill is achieved over regions with less or without pseudoproxies, indicating that climate teleconnections between tropics and mid-latitude regions could be responsible for the reconstruction skill in tropical regions.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

the contaminated PPEs in Fig. 8 and 9). The Bi-LSTM captures the lower tails of distribution much better than the upper tails, both for the NH mean and the AMV index.

We also quantify the distribution similarity between reconstructed and target distributions for both NH and AMV indices using the two-sample Kolmogorov-Smirnov test as a metric (Hodges, 1958) (see Table 4-5). The smallest KS statistic is achieved by the PCR method (see Table 4-5), confirming the impression that the PCR outperforms the other two methods in both the ideal and noise contaminated PPEs.

4 Discussions

4.1 Nonlinear method performance

As the Bi-LSTM method is the most complex of the three tested in this study, it is reasonable to expect a better reconstruction skill. However, this is not the case in our pseudoproxy experiments. For the spatially resolved NAE and NH fields, the nonlinear Bi-LSTM method achieves a similar skill as the linear PCR and CCA methods. Whereas the spatial SD ratio and cc present CFR are generally very similar, the PCR is generally the best method among the three methods, with the nonlinear Bi-LSTM as second best method (see mean skill statistics in Table 1).

For the area-mean indices, all three methods exhibit again similar skill. The Bi-LSTM is able to capture periods of extreme cooling better than the other two methods but strongly underestimates the recent warming trend. The inability to capture the warming trend indicates that the Bi-LSTM is not good at extrapolating to temperature ranges beyond the training set. Nonetheless, even though the traditional PCR seems to display the overall best performance, the fact that the Bi-LSTM is able to capture some extremes better is encouraging. This indicates that there may not be one sole reconstruction method which captures the mean and the extremes equally well. In addition, Guillot et al. (2015) constructed different CFR methods for exploring the disagreement between climate field reconstructions and area mean index reconstructions. They demonstrated that the skill for the regionally averaged time series is a relatively poor indicator of the spatial performance.

Nonlinear methods are usually capable of mapping complex systems with high nonlinearity. Here, we employ one nonlinear neural network method Bi-LSTM to test its performance on CFR reconstructions. Compared with linear methods PCR and CCA, neural network model did not show clear advantages, except for capturing extremes in the lower tail of the distribution. It is possible that the performance of the Bi-LSTM could be further improved by optimizing the architecture and parameters of the network, including the type of object function, type of neural activation function, network optimization function, number of hidden layers, the model-learning rate etc. At this point, it would be quite natural to consider whether the selection/settings of these hyper-parameters in our study is optimal, and also to what extent the reconstruction skill is sensitive to changes in the hyper-parameters. Nadiga (2020) pointed out that the skill of some machine learning-methods are strongly dependent on these hyper-parameters. Machine learning methods include an extensive range of complexity, and therefore it remains an open issue as to which ML techniques are most or relatively suitable for paleoclimate. It is not clear how the structure of the machine-learning methods can be systematically optimized. At the moment, there is still a considerably amount of 'trial and error' in

All derived CFRs suffer from underestimation of interannual variance, as shown in Fig. 5 and in Table 1, except that the PCR method presents a clearly interannual variance overestimation referring to the specific spatial SD ratio map in Fig. 5. This overestimation may be impacted by overfitting, since the number of predictors is 47 pseudo-proxies and the calibration period spans 100 time steps. The spatial distributions of the SD ratio also vary between climate models and CFR methodologies. They also are spatially heterogeneous. The CCA method and Bi-LSTM generally preserve more variance over regions with denser pseudoproxies in both CAM and MPI model, and a relatively higher SD ratio appeared in tropical regions within Bi-LSTM based PPEs shown in Fig. 5.

The CCA methodology seems to suffer more strongly from variance losses (see Table 1) over the entire NH compared to PCR and Bi-LSTM.

Considering the general methodological skill, as indicated by the derived spatial mean *cc* and SD ratio values in Table 1, the Bi-LSTM method presents relatively worse performance with lower mean *cc*. The methods PCR and Bi-LSTM generally outperform the CCA methodology with higher mean SD ratio within ideal PPEs.

3.3 Spatially variability patterns of the reconstructed fields

In this section, we test the skill of the CFR in replicating the leading spatial patterns of variability, conducting an EOF analysis of the reconstructed temperature fields and compare them with the patterns derived from the target climate simulations. This type of comparison follow the tests performed by (Yun et al, 2021). In this comparison, the PCA and CCA methods have a clear built-in advantage relative to the Bi-LSTM network, since these two methods operate by design in the space spanned by the leading EOFs of the temperature field. In the case of PCR, these reconstructed fields are a linear combination of the EOF patterns themselves. Therefore, in as much the reconstructed PC series remain uncorrelated, the EOFs of the reconstructed field will be exactly equal to the EOFs of the target climate simulations. Deviations from this behaviour may be caused by the lack of strict orthogonality between the reconstructed PC series caused by the relationship between proxy (predictors) and the PC series (predictands). However, it is reasonably to think that it would not be *a priori* surprising that the EOFs of the PCR-reconstructed fields would be similar to the original EOFs. The case for CCA is theoretically similar, but there are some potentially important points to bear in mind. The CCA patterns, which serve as a basis for the reconstructed field, are linear combinations of the original EOFs. These linear combinations may, for instance, not include the leading EOF of the original field, and thus, the EOFs of the reconstructed field will not replicate the original leading EOF, even if the CCA series can be perfectly reconstructed by the proxy series. The third method Bi-LSTM is in this sense at disadvantage relative to PCR and CCA, since the spatial covariance of the original field is not technically incorporated in its machinery. If the EOF patterns of the reconstructed field resemble the original EOF patterns, this would be an indication that the method itself is able to capture the main covariance patterns of the original field.

In order to have a deeper insight for the reconstruction performance of three CFR methods, we calculated the four leading EOF patterns, and their proportion of explained variance of the reconstructed field, derived from the three reconstruction methods using the CAM5 pseudo-proxies. The EOF patterns represented in Figure 8 confirm the suggestions that the

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

the design and connection of the neural layers. Here, we have tested the Bi-LSTM network with a relatively simple architecture of two separated hidden layers, and evaluated its performances on CFR experiments, which could be a preliminary try. It is, however possible that more complex models or architectures (Kadow et al., 2020) might achieve better or comparable skill in CFR experiments. Thus, more different theory-based and architecture-based machine learning methods might be worth exploring in future studies.

4.2 Model and pseudoproxy experiment dependency

The evaluation of the reconstruction skill seems to depend as much on the reconstruction method as on the underlying climate model simulation from which the pseudoproxies were generated. The differences in skill for the same method with different climate model data is of the same order as the differences in skill for the different methods with the same climate model data. The performance of the method does not seem to depend on the domain of the reconstruction. The reconstructions behave very similar for both the NAE and the NH test cases.

Considering the effects of noise contamination on the methodological performance, all three methods display similar skills in the ideal PPEs, but all methods suffer from variance underestimation and lower correlation coefficients in the more realistic PPEs (80% noise contaminated PPEs). The nonlinear Bi-LSTM is more strongly impacted by the noise contamination (Table 1).

From the perspective of the spatial coverage of the proxy network, the spatial cc and SD ratio patterns reveal reconstruction skill over the entire NH regions, although this skill is weaker in areas more poorly sampled by the pseudo-proxy network. Interestingly, the tropical regions do show some reconstruction skill, although almost not pseudo-proxies are located in the Tropics. In addition, the reconstruction methods achieve better reconstruction skill when evaluated with the climate model with highest (most realistic) spatial resolution. This result indicates the climate teleconnections between tropics and mid-latitude regions could lead to some indirect skill, and that in the real world this indirect reconstruction skill could be larger than that obtained in our pseudo-proxy experiments. However, the proxy networks and noise scenarios constructed in the context are certainly not able to mimic/simulate the full range of characteristics completely for climatic proxies in the real world.

5 Conclusions

A nonlinear Bi-LSTM neural network method to reconstruct North Atlantic-Europe and Northern Hemisphere temperature fields was tested with climate surrogate data generated by simulations with three different climate models. Compared to the more classical methods of linear Principal Components Regression and Canonical Correlation Analysis, the NAE and NH summer temperature field could be reasonably reconstructed using both linear and nonlinear methodologies. All three methods show skill similarities in both NAE and NH PPEs.

temperature reconstructed by the PCR and CCA methods (two lower rows in Figure 8) replicate very closely the three leading patterns. The fourth EOF pattern displays some divergences from the original fourth pattern, but as we will show later, the variance explained by this fourth EOF is already rather low, so that the spatial pattern may be subject to statistical noise. More importantly, the Bi-LSTM method (second row) does produce EOF patterns than closely resemble the ones derived from the original field. This supports the idea that the method is able to replicate the spatial cross-covariance of the temperature field.

The corresponding spectrum of explained variance is displayed in Figure 9. Here, the percentage of explained variance is calculated relative to the total variance of the original field. We deviate here from the analysis of Yun et al. (2021), who according to their methodological description calculate the portion of variance relative to the total variance of the respective field (targetor reconstructed). The reason for our choice is to avoid statistical artefacts. For instance, when using the PCR regression method, we would choose to reconstruct only the leading EOF pattern. This pattern will explain 100% of the reconstructed variance, but this result would be obviously not informative.

3.4 An alternative pseudo-proxy network

In this section, we summarize a few additional experiments using the original locations of the PAGES network (Emile-Geay et al., 2017) instead of the filtered network used in previous experiments. In this section, we show only one model test-bed, for ideal, white-noise and red-noise pseudo-proxies. The results obtained with the MPI-ESM-M model are similar and are, omitted here for the sake of brevity.

The reconstruction skill measured by *cc* and SD ratio display similar spatial patterns as those obtained with network pre-selected according to the criteria of St. George (2014). As shown in Fig. 10, the derived correlations are generally higher over regions where denser pseudoproxy exits across both ideal and noisy PPEs, and weakly reconstructed correlations appeared over pseudoproxies-free regions. The PCR method presents a distinct interannual variance overestimation as shown in the specific spatial SD ratio map in Fig. 10 amongst ideal and noisy PPEs, while a clearly interannual variance overestimation also occurs in CCA-based CFRs in the noisy PPEs. A relatively reasonable SD ratio is revealed in tropical regions within Bi-LSTM based PPEs shown in Fig. 10. In general, high reconstruction skills remain over regions where denser pseudoproxy exists based on this additional PAGES 2k pseudoproxy network.

3.5 Northern Hemisphere and AMV indices

The evolution of the decadal NH mean temperature anomalies reconstructed by the three CFR methodologies and using pseudoproxies from two models is illustrated in Fig. 8. All indices have been smoothed using a Butter worth low-pass filter to remove temporal fluctuations shorter than 10 years. The reconstruction performance varies amongst different the CFR methodologies. We will employ the correlation coefficient-*cc*, standard deviation-SD and root mean square error-RMSE as evaluation metrics for NH and AMV indices.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

In general, all three methods display similar skills when using ideal (noise-free) pseudoproxies, while in the more realistic PPEs (80% noise contaminated PPEs), the Nonlinear Bi-LSTM seems to suffer more strongly from variance losses. This might indicate that the nonlinear method would be more sensitive to the noise contamination than the other methods.

The pseudoproxy networks used in this study were mostly located in the extratropical regions with only three proxies in the tropical area. All CFR methodologies produce generally good reconstructions in regions where dense pseudoproxy networks are available. Moreover, teleconnections are explored by these CFR methodologies, leading to some weak spatial reconstruction skills outside of the proxy-sampled regions, for instance the tropical region.

The classical linear-based PCR method generally outperforms the Bi-LSTM and CCA method in both spatial and index reconstructions. However, the Bi-LSTM seems to be able to capture extreme cooling events better than the other two methods, while failing to capture the warm tails of the temperature. In particular, this is reflected in the inability of the non-linear method to replicate the 20th century warming trend.

Here, we could draw a general conclusion that nonlinear artificial neural network method Bi-LSTM employed herein is not superior for CFR reconstructions, at least in our PPEs. In general, Bi-LSTM show worse skill in spatial and temporal CFRs than PCR and CCA, also in capturing extremes. Yet, it is essential to employ a large range of nonlinear CFR methods to evaluate different model structures, and further test their performance on CFRs. For example, additional nonlinear-based regression methodologies, convolutional neural network and one of the widely implemented Reservoir Computing methods-Echo State Network, could be techniques with powerful non-linear regression capability for paleoclimate field reconstructions.

Appendix A

The simulation with the model MPI-ESM-P is not part of the standard CMIP5 simulation suite. In the following, we include additional technical details on this simulation. The MPI simulation was started from the year of 100 BC with restart files from a 500-year spin-down simulation experiments forced with constant external conditions representing the year 100 of BC. After 100 BC, variation in volcanic, solar, orbital, and GHG concentrations are implemented. Land usage was held constant until 850 AD with conditions representing those for year 850 AD. The variation of orbital parameters are calculated after the PMIP3-protocol (Schmidt et al. 2011). The solar activity has been rebuilt on the basis of the reconstruction of Vieira et al. 2011 employing the algorithm and scaling outlined in Schmidt et al. 2011 which corresponds to a difference in short-wave top of the atmosphere insolation of 1.25 Wm-2 (– 0.1%) between the 2nd half of the 20th century (1950 – 2000) and the Maunder Minimum (1645 – 1715). Variations in greenhouse gas concentrations related to CO2, N2O and CH4 are after the reconstruction of the PMIP3 protocol – The concentrations were held constant to the values of year 1 AD between 100 BC and 1 AD because the law Dome records does not extend beyond year 1 AD. After 1850 AD also a reconstructed aerosol loading after Stine et al. 2018 were employed to account for transient anthropogenic aerosol emissions. The extension and reconstruction of the volcanic forcing is related to a rescaling of the newly available Sigl et al. (2015) dataset to the reconstruction of Crowley and Unterman (2013). The large volcanoes for different latitudinal bands are rescaled according to

Table 2. *cc*, SD and RMSE (K) during the verification interval for decadal NH mean temperature derived from ideal PPEs. The numbers in parenthesis indicate skill statistics of white and red (italics) noise contaminated PPEs.

Method	*cc*		SD		RMSE	
	MPI	CAM	MPI	CAM	MPI	CAM
PCR	0.880 (0.632/*0.302*)	0.871 (0.532/*0.435*)	0.821 (0.806/*0.883*)	0.763 (0.502/*0.688*)	0.086 (0.143/*0.202*)	0.072 (0.122/*0.135*)
CCA	0.882 (0.664/*0.203*)	0.853 (0.536/*0.262*)	0.704 (0.647/*0.711*)	0.560 (0.464/*0.660*)	0.091 (0.135/*0.187*)	0.086 (0.122/*0.141*)
Bi-lstm	0.873 (0.593/*0.351*)	0.901 (0.559/*0.394*)	0.561 (0.513/*0.540*)	0.597 (0.398/*0.470*)	0.104 (0.146/*0.173*)	0.076 (0.122/*0.133*)

Table 3. The same as Table 2, but for decadal AMV index

Method	*cc*		SD		RMSE	
	MPI	CAM	MPI	CAM	MPI	CAM
PCR	0.819 (0.577/*0.336*)	0.758 (0.354/*0.429*)	0.831 (0.826/*0.961*)	0.753 (0.602/*0.837*)	0.108 (0.161/*0.213*)	0.091 (0.135/*0.139*)
CCA	0.822 (0.631/*0.288*)	0.777 (0.457/*0.424*)	0.689 (0.669/*0.744*)	0.591 (0.541/*0.766*)	0.110 (0.146/*0.200*)	0.092 (0.125/*0.136*)
Bi-lstm	0.846 (0.573/*0.344*)	0.829 (0.435/*0.450*)	0.623 (0.539/*0.576*)	0.600 (0.440/*0.536*)	0.108 (0.154/*0.182*)	0.084 (0.126/*0.125*)

The temporal evolution of the original AMV indices (Fig. 12) differs among the simulations, reflecting the different forcings used in each simulation and the model specific contribution of internal variability to the index variations (Wagner and Zorita, 2005; Schmidt et al., 2011). Considering the methodological performance, all three methods generally achieve good AMV index reconstructions when using perfect pseudo-proxies, as shown in each subfigures of Fig. 11 and in Table 3.

The NH and AMV indices derived from more realistic noise contaminated CFRs are shown in Fig. 11 and Fig. 12 correspondingly. The larger noise contamination results in substantial skill deterioration (*cc*, SD and RMSD displayed within brackets in Table 2 and 3). All three methods generally fail to capture the complete variance of the target indices, and the magnitude of strong cooling phases is strongly underestimated.

Fig. 13 illustrates the comparison of Northern Hemisphere indices power spectral density for both, ideal and noise-contaminated PPEs between reconstructions and target models. As indicated in Fig. 13, all three methods generally underestimate the power density, whereas this underestimation is more significant for the noise-contaminated derived PPE.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

sulfate concentrations and eventually the Crowley algorithm was applied to yield aerosol optical depths and effective radius for four latitudinal bands separated by 30°.

Appendix B

Table B1. Skill reconstruction statistics for the North Hemisphere mean temperature in the verification period for noise-contaminated PPEs based on CCSM using different neuron numbers in hidden layers of Bi-LSTM method.

Number of neurons	SD Ratio	cc
50 neurons	0.289	0.443
200 neurons	0.315	0.461
500 neurons	0.339	0.467
700 neurons	0.412	0.456

In order to check whether the number of neurons in the hidden layer will substantial impact the CFR skills. We also conducted additional tests for our Bi-LSTM model structures with different neurons in the hidden layer based on one of the settings for PPEs. Table 1 indicates that, when we fix the rest of the hyper-parameters, etc. optimization function and learning rate, of the Bi-LSTM structure, just tuning the number of neurons in the hidden layer does not yield a too strong impact on the mean correlation coefficients, but it results in an obvious impact on the SD ratio. Considering the variability capturing performance, it seems that the hidden layer with 700 neurons structure outperforms the hidden layer with less neurons structure, whereas the temporal covariance between target and reconstructed results does not change too much amongst these four different hidden layer structures. In our PPE tests on paleo CFRs, it seems that there is no specific neural network structure could universally outperform another one.

Data availability

The CCSM4 and MPI-ESM-P model output that was employed for this study is publicly accessible and can be downloaded by the Earth System Grid Federation (ESGF); https://esgf-data.dkrz.de/projects/esgf-dkrz/. The CESM1 model data can be downloaded: https://www.cesm.ucar.edu/projects/community-projects/LME/.

Author contributions

The analysis was performed by ZZ with the consultation of SW, MK and EZ. ZZ prepared the paper with contributions from all co-authors.

3.6 Probability distributions of reconstructed variables

Even though the three reconstructions methods tend to underestimate the overall variability when using noisy pseudoproxies, an interesting question is their skill in reproducing the probability distributions of the climate indices. In particular, a relevant question is whether the methods are able to capture extreme phases of those indices.

Fig. 14 and 15 display the histogram for the decadal NH mean and AMV indices, respectively. Each subfigure represents the histograms of reconstructed temperature indices across the three methods, compared with the histograms of the target temperature index.

For perfect pseudoproxies, the PCR reconstruction seems to capture the overall target distribution best. It captures the lower tail better than CCA and the upper tail better than CCA and Bi-LSTM. The differences between the methods become smaller for the reconstructions with noisy pseudo proxies, with the PCR still being better than the other two methods (subfigures for the contaminated PPEs in Fig. 14 and 15). The Bi-LSTM performs worst in capturing the lower and upper tails of distribution amongst the three methods, both for the NH mean and the AMV index.

Table 4. Kolmogorov-Smirnov test statistic and p-value for quantifying the histogram distributions between model and reconstructed NH decadal means. Low values of the KS statistic indicate larger similarity between the two distributions. The numbers in parenthesis indicate the KS statistic and p-value of white and red (italics) noise contaminated PPEs.

Method	KS statistic		p-value	
	MPI	CAM	MPI	CAM
PCR	0.043(0.074/*0.093*)	0.009(0.193/*0.111*)	2e-1(6e-3/*2e-4*)	3e-4(1e-17/*4e-6*)
CCA	0.068(0.081/*0.073*)	0.171(0.197/*0.130*)	1e-2(1e-3/*7e-3*)	6e-14(2e-18/*3e-8*)
Bi-lstm	0.120(0.142/*0.112*)	0.178(0.241/*0.200*)	5e-7(9e-10/*3e-6*)	5e-15(2e-27/*5e-19*)

Table 5. The same as Table 4, but for AMV index.

Method	KS statistic		p-value	
	MPI	CAM	MPI	CAM
PCR	0.052(0.050/*0.086*)	0.101(0.143/*0.085*)	1e-2(1e-1/*7e-4*)	3e-5(6e-10/*8e-4*)
CCA	0.082(0.088/*0.083*)	0.159(0.163/*0.103*)	1e-3(5e-4/*1e-3*)	5e-12(1e-12/*2e-5*)
Bi-lstm	0.117(0.154/*0.129*)	0.172(0.224/*0.191*)	1e-6(2e-11/*4e-8*)	4e-14(1e-23/*3e-17*)

We also quantify the distribution similarity between reconstructed and target distributions for both NH and AMV indices using the two-sample Kolmogorov-Smirnov test as a metric (Hodges, 1958) (see Table 4-5). A smaller value of the KS statistic indicates a stronger overall similarity between the two probability distributions. The smallest KS statistic is achieved

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors thank the CCSM4 and MPI-ESM modelling groups participating in the CMIP5 initiative for providing their data and the CESM1 modelling group for making their data available. This work is funded by the China Scholarship Council (no. 201806570017), and is part of the project Reduced Complexity Models (Redmod), funded by the Helmholtz Association through its Incubator program.

References

Amrhein, D. E., Hakim, G. J., and Parsons, L. A.: Quantifying structural uncertainty in paleoclimate data assimilation with an application to the Last Millennium, Geophys. Res. Lett., 47, e2020GL090485. https://doi.org/10.1029/2020GL090485, 2020.

Anchukaitis, K. J., Wilson, R., Briffa, K. R., Büntgen, U., Cook, E. R., D'Arrigo, R., Davi, N., Esper, J., Frank, D., Gunnarson, B. E., Hegerl, G., Helama, S., Klesse, S., Krusic, P. J., Linderholm, H. W., Myglan, V., Osborn, T. J., Zhang, P., Rydval, M., Schneider, L., Schurer, A., Wiles, G., and Zorita, E.: Last millennium Northern Hemisphere summer temperatures from tree rings: Part II, spatially resolved reconstructions, Quaternary Sci. Rev., 163, 1–22, https://doi.org/10.1016/j.quascirev.2017.02.020, 2017.

Biswas, S., Sinha, M.: Performances of deep learning models for Indian Ocean wind speed prediction, Model. Earth Syst. Environ., 7, 809–831,https://doi.org/10.1007/s40808-020-00974-9, 2021.

Biswas, K., Kumar, S., and Pandey, A. K.: Intensity Prediction of Tropical Cyclones using Long Short-Term Memory Network. arXiv [preprint], https://arxiv.org/abs/2107.03187, 2021.

Christiansen, B. and Ljungqvist, F. C.: Challenges and perspectives for large-scale temperature reconstructions of the past two millennia, Rev. Geophys., 55, 40–96, https://doi.org/10.1002/2016RG000521, 2016.

Chattopadhyay, A., Hassanzadeh, P., and Subramanian, D.: Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network, Nonlin. Processes Geophys., 27, 373–389, https://doi.org/10.5194/npg-27-373-2020, 2020.

Crowley, T. J. and Unterman, M. B.: Technical details concerning development of a 1200 yr proxy index for global volcanism, Earth Syst. Sci. Data, 5, 187–197, https://doi.org/10.5194/essd-5-187-2013, 2013.

Cui, Z.; Ke, R.; Pu, Z.; Wang, Y. Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. arXiv [preprint], https://arxiv.org/abs/1801.02143, 2018.

Dueben, P. D. and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geosci. Model Dev., 11, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018, 2018.

by the PCR method (see Table 4-5), confirming the impression that the PCR outperforms the other two methods for indices reconstructions in both the ideal and noise contaminated PPEs.

3.7 Alternative architectures of the Bi-LSTM method

Although the design of machine-learning methods may be guided by the physical considerations, machine-learning methods are still to a large extent a matter of trial and error. The same complexity of the method hinders the disentangling of the causes as to why the methods behave in a certain way. Here, we explore alternative architectures of the Bi-LSTM method to assess the resoluteness of the conclusions drawn from the basic design. We have explored varying network depths (number of layers), different learning rates, and different cost-functions to optimize the network parameters, among others. A summary of the results is included in the Appendix B.

We could not recognize systematic effects in the skill in this set of different networks designs. The skill varies rather randomly, and probably the identification of optimal network architectures for this specific reconstruction question may not be extrapolated to other applications in paleoclimate. We settled for this application, on a heuristic basis, on an architecture with 2 hidden layers, 4000 hidden nodes, with a learning rate of 10^{-3}, with the activation function *leaky relu*, a batchsize of 20 and the Huber loss function.

4 Discussion

4.1 Nonlinear method performance

Our initial hypothesis was that a more sophisticated model might be able to better capture relationships that are more complex. For instance, a linear model cannot capture non-linear links outside a narrow range of variations. Artificial neural network is a subset of machine learning method that can be understood as a universal approximator, which can map and approximate any kind of functions by selecting a suitable set of connecting weights and transfer functions (Hornik et al., 1989). Thus, it is reasonable to assume that a better representation of the links between proxy series and climate fields, and thus a better reconstruction performance, might be achieved.

The Bi-LSTM method is the most complex of the three tested in this study. Among them, it is also the one that aims at capturing the serial dependencies. Our hypothesis was that better reconstruction skill could be achieved by the Bi-LSTM method. However, this is not the case in our pseudoproxy experiments. For the spatially resolved NAE fields, the nonlinear Bi-LSTM method achieves a similar skill as the linear PCR and CCA methods, both with ideal and noisy PPEs (see Fig. 2-4).

For spatially resolved NH field, the PCR overestimates the variabilities both in ideal and noisy PPEs (see spatial SD ratio maps in Fig. 5-7 and mean statistics skills Table 1), and the CCA method shows relatively lower overestimation in noisy PPEs, the Bi-LSTM presents relatively reasonable reconstructions without clearly overestimation both in ideal and noisy

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

Evans, M., Smerdon, J. E., Kaplan, A., Tolwinski-Ward, S., and González-Rouco, J. F.: Climate field reconstruction uncertainty arising from multivariate and nonlinear properties of predictors, Geophys. Res. Lett., 41, 9127–9134, https://doi.org/10.1002/2014gl062063, 2014.

Folland, C. K., Knight, J., Linderholm, H. W., Fereday, D., Ineson, S., and Hurrell, J. W.: The Summer North Atlantic Oscillation: Past, present, and future, J. Climate, 22(5), 1082–1103, doi:10.1175/2008JCLI2459.1, 2009.

Gao, C., Robock, A., and Ammann, C.: Volcanic forcing of climate over the past 1500 years: An improved ice core-based index for climate models, J. Geophys. Res.-Atmos., 113, D23111, https://doi.org/10.1029/2008JD010239, 2008.

Gómez-Navarro, J. J., Zorita, E., Raible, C. C., and Neukom, R.: Pseudo-proxy tests of the analogue method to reconstruct spatially resolved global temperature during the Common Era, Clim. Past, 13, 629–648, https://doi.org/10.5194/cp-13-629-2017, 2017.

Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural and other neural network architectures, Neural Netw 18(5), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042, 2005.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., and Vertenstein, M.: The community climate system model version 4, J. Clim., 24, 4973–4991, doi:10.1175/2011JCLI4083.1, 2011.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning, MIT Press, 2016.

Guillot, D., Rajaratnam, B., and Emile-Geay, J.: Statistical paleoclimate reconstructions via Markov Random Fields, Ann. Appl. Stat., 9, 324–352, https://doi.org/10.1214/14-aoas794, 2015.

Hernández, A., Martin-Puertas, C., Moffa-Sánchez, P., Moreno-Chamarro, E., Ortega, P., Blockley, S., Cobb, K. M., Comas-Bru, L., Giralt, S., Goosse, H., Luterbacher, J., Martrat, B., Muscheler, R., Parnell, A., Pla-Rabes, S., Sjolte, J., Scaife, A. A., Swingedouw, D., Wise, E., and Xu, G.: Modes of climate variability: Synthesis and review of proxy-based reconstructions through the Holocene, Earth Sci. Rev., 271, 103286, https://doi.org/10.1016/j.earscirev.2020.103286, 2020.

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., and Yang, H.: Machine learning and artificial intelligence to aid climate change research and preparedness, Environ. Res. Lett., 14, 124007, https://doi.org/10.1088/1748-9326/ab4e55, 2019.

Huang, Y., Yang, L., and Fu, Z.: Reconstructing coupled time series in climate systems using three kinds of machine-learning methods, Earth Syst. Dynam., 11, 835–853, https://doi.org/10.5194/esd-11-835-2020, 2020.

Hunke, E., Lipscomb, W., Turner, A., Jeffery, N., and Elliott, S.: CICE: the Los Alamos sea ice model, documentation and software, version 4.0. Los Alamos National Laboratory, Tech. Rep, LA-CC-06-012, 2008. Los Alamos National Laboratory, Los Alamos, NM.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J. F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, B. Am. Meteorol. Soc., 94, 1339–1360, https://doi.org/10.1175/bams-d-12-00121.1, 2013.

PPEs (see Fig. 5-7 and Table 1). Amongst ideal PPEs across two models, the PCR is generally the best method among the three methods, and the nonlinear Bi-LSTM is second best method with higher SD ratio and worse *cc* than CCA method (see Fig. 5-7 and mean skill statistics in Table 1). Whereas, both PCR and CCA exhibit overestimated reconstructions within noisy PPEs, the Bi-LSTM presents relatively robust reconstructions especially without overestimations in noisy PPEs (see Fig. 5-7 and mean skill statistics in Table 1).

For the area-mean indices, all three methods exhibit again generally similar skill. Nevertheless, the Bi-LSTM more strongly underestimates the amplitude of variabilities, and especially over some extreme cooling phases than PCR and CCA. This underestimation is also generally model dependent (see different reconstructed performances in Fig 11-12). In general, the PCR methods achieved the best performance both in extreme cooling signal capture and indices reconstructions across two models and amongst three methods. The power spectral density plots in Fig. 13 provide a deep insight about these different reconstruction performances in NH temperature indices.

The general inability to capture the cooling extreme signals prior to 20th century indicates that the Bi-LSTM is not good at extrapolating to temperature ranges beyond the training set – a phenomenon that is intrinsic to most ML-based methods. Therefore, compared with linear methods PCR and CCA, neural network model did not show clear advantages. The performance of the Bi-LSTM might be further improved by optimizing the architecture and parameters of the network, including the type of objective function, type of neural activation function, network optimization function, number of hidden layers, the model-learning rate etc. At this point, it would be quite natural to consider whether the selection/settings of these hyper-parameters in our study is optimal, and also to what extent the reconstruction skill is sensitive to changes in the hyper-parameters. Nadiga (2020) pointed out that the skill of some machine learning-methods are strongly dependent on these hyper-parameters. Machine learning methods include an extensive range of complexity, and therefore it remains an open issue as to which ML techniques are most or relatively suitable for paleoclimate. It is not clear how the structure of the machine-learning methods can be systematically optimized. At the moment, there is still a considerably amount of 'trial and error' in the design and connection of the neural layers. Here, we have tested the Bi-LSTM network with several different architecture settings, and finally decided a relatively optimal architecture with two separated hidden layers, and evaluated its performances on CFR experiments, which could be a preliminary try. Our first implementation of the more complex Bi-LSTM does not show superiority in CFRs, at least in our specific experiments, compared to traditional CFR methods, so we would like to draw an assumption that more complicated architecture might not be helpful for CFRs. In addition, a degradation of out-of-sample performance may well be expected when a limited dataset is used to train a neural network model (Najafabadi et al., 2015). Nevertheless, we would like to point out to other methods, such as an Echo State Network (ESN, Lukosevicius and Jaeger, 2009; Nadiga, 2020) for paleo climate research. Both ESN and LSTM belong to the family of RNN, yet ESN is much simpler than LSTM (Lukosevicius and Jaeger 2009), and has outperformed the RNN methods in other applications (Chattopadhyay et al., 2019; Nadiga, B. 2020).

Another reason to consider machine-learning methods is the non-linearity of the link between proxies and climate fields. In this particular application with pseudoproxies, the implied link is probably close to linear. However, these can be different

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

Hurtt, G. C., Chini, L. P., Frolking, S., Betts, R., Feddema, J., Fischer, G., Goldewijk, K. K., Hibbard, K., Janetos, A., and Jones, C.: Harmonisation of global land-use scenarios for the period 1500–2100 for IPCC-AR5, ILEAPS Newsletter, No. 7, 6–8, 2009.

Hotelling, H.: The relations of the newer multivariate statistical methods to factor analysis, Brit. J. Statist. Psych., 10, 69–76, https://doi.org/10.1111/j.2044-8317.1957.tb00179.x, 1957.

Hodges, J. L.: The significance probability of the Smirnov two-sample test, Arkiv för Matematik, 3, 469–486, 1958.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., and Núñez-Riboni, I.: Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations, J. Adv. Model. Earth Syst., 5, 287–315, 2013.

Jacobeit, J., Wanner, H., Luterbacher, J., Beck, C., Philipp, A., and Sturm, K.: Atmospheric circulation variability in the NorthAtlantic-European area since the mid-seventeenth century, Clim. Dynam., 20, 341–352, https://doi.org/10.1007/s00382-002-0278-0, 2003.

Jahangir, H., Tayarani, H., Gougheri, S.S., Golkar, M.A., Ahmadian, A., Elkamel, A.: Deep Learning-based Forecasting Approach in Smart Grids with Micro-Clustering and Bi-directional LSTM Network. IEEE Trans. Ind. Electron., 68, 8298–8309, doi: 10.1109/TIE.2020.3009604, 2020.

Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., Mikolajewicz, U., Notz, D., and vonStorch, J. S.: Characteristics of the ocean simulations in MPIOM, the ocean component of the MPI-Earth system model, J. Adv. Model. Earth Syst., 5, 422–446, doi:10.1002/jame.20023, 2013.

Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv [preprint], https://arxiv.org/abs/1412.6980, 2014.

Knight, J. R., Folland, C. K., and Scaife, A. A.: Climate impacts of the Atlantic Multidecadal Oscillation, Geophys. Res. Lett., 33, L17706, doi:10.1029/2006GL026242, 2006.

Kadow, C., Hall, D. M., and Ulbrich, U.: Artificial intelligence reconstructs missing climate information, Nat. Geosci., 13, 408–413, https://doi.org/10.1038/s41561-020-0582-5, 2020.

Lean, J., Rottman, G., Harder, J., and Kopp, G.: SORCE contributions to new understanding of global change and solar variability, in: The Solar Radiation and Climate Experiment (SORCE), Springer, 2005.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H.: European seasonal and annual temperature variability, trends, and extremes since 1500, Science, 303, 1499–1503, DOI: 10.1126/science.1093877, 2004.

Luterbacher J., Werner, J. P., Smerdon, J. E., Fernández-Donado, L., González-Rouco, F. J., Barriopedro, D., Ljungqvist, F. C., Büntgen, U., Zorita, E., Wagner, S., Esper, J., McCarroll, D., Toreti, A., Frank, D., Jungclaus, J. H., Barriendos, M., Bertolin, C., Bothe, O., Brázdil, R., Camuffo, D., Dobrovolný, P., Gagen, M., García-Bustamante, E., Ge, Q., Gómez-Navarro, J. J., Guiot, J., Hao, Z., Hegerl, G. C., Holmgren, K., Klimenko, V. V., MartínChivelet, J., Pfister, C., Roberts, N., Schindler, A., Schurer, A., Solomina, O., von Gunten, L., Wahl, E., Wanner, H., Wetter, O., Xoplaki, E., Yuan, N., Zanchettin, D., Zhang, H., and Zerefos, C.: European summer temperatures since Roman times, Environ. Res. Lett., 11, 024001, doi:10.1088/1748-9326/11/2/024001, 2016.

on other cases. And it might be the case for more complex problems, (i.e. the reconstruction of proxy-precipiation fields or other modes of natural variability such as the NAO or ENSO). As such, ML methods should not a-priori be excluded from the portfolio of CFR methods leading to more skilful reconstructions of climate.

4.2 Model and pseudoproxy network dependency

The evaluation of the reconstruction skill seems to depend as much on the reconstruction method as on the underlying climate model simulation from which the pseudoproxies were generated. The differences in skill for the same method with different climate model data is of the same order as the differences in skill for the different methods with the same climate model data. The performance of the method does not seem to depend on the domain of the reconstruction. The reconstructions behave generally similar for the NAE, nevertheless, show some differences in the NH test cases, especially in the derived SD ratio patterns.

Considering the effects of noise contamination on the methodological performance, both PCR and CCA method exhibit overestimation (see SD ratio patters in Fig. 9-10 and mean skills in Table 1). However, all methods suffer from lower correlation coefficients in the more realistic PPEs (white and red noise contaminated PPEs). The nonlinear Bi-LSTM is more strongly impacted by the noise contamination (Table 1).

From the perspective of the spatial coverage of the proxy network, the spatial *cc* and SD ratio patterns (except PCR method) reveal reconstruction skill over the entire NH regions, although this skill is weaker in areas more poorly sampled by the pseudo-proxy network (spatial *cc* patterns in Fig. 5-7). Interestingly, the tropical regions do show some reconstruction skill especially in the derived reconstructions based on Bi-LSTM (spatial SD ratio patterns in Fig. 5-7), although almost no pseudo-proxies are located in the Tropics. This result indicates the climate teleconnections between tropics and mid-latitude regions could lead to some indirect skill. However, the proxy networks and noise scenarios constructed in the context are certainly not able to mimic/simulate the full range of characteristics completely for climatic proxies in the real world.

5 Conclusions

A nonlinear Bi-LSTM neural network method to reconstruct North Atlantic-Europe and Northern Hemisphere temperature fields was tested with climate surrogate data generated by simulations with two different climate models. Compared to the more classical methods of linear Principal Components Regression and Canonical Correlation Analysis, the NAE and NH summer temperature field could be reasonably reconstructed using both linear and nonlinear methodologies referring to spatial *cc* metric. In the relatively larger spatial region-NH temperature field, more discrepancies of reconstructions appeared amongst different climate models and methods based on the derived spatial SD ratio metric. The conclusions drawn from this study can be summarized as follows:

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

Lindgren, A., Lu, Z., Zhang, Q., and Hugelius, G.: Reconstructing past global vegetation with random forest machine learning, sacrificing the dynamic response for robust results J. Adv. Model. Earth. Syst., 13, p.e2020MS002200. https://doi.org/10.1029/2020MS002200, 2021.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Fletcher, C. G., Lawrence, P. J., Levis, S., Swenson, S. C., and Bonan, G. B.: The CCSM4 land simulation, 1850–2005: Assessment of surface climate and new capabilities, J. Clim., 25, 2240–2260, doi:10.1175/JCLI-D-11-00103.1, 2012.

Li, B. and Smerdon, J. E.: Defining spatial comparison metrics for evaluation of paleoclimatic field reconstructions of the Common Era, Environmetrics, 23, 394–406, doi:10.1002/env.2142, 2012.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779–787, https://doi.org/10.1038/33859, 1998.

Mann, M. E. and Rutherford, S.: Climate reconstruction using "Pseudoproxies", Geophys. Res. Lett., 29, 1501, https://doi.org/10.1029/2001GL014554, 2002.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Testing the fidelity of methods used in proxy-based reconstructions of past climate, J. Clim., 18, 4097–4107, https://doi.org/10.1175/JCLI3564.1, 2005.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Robustness of proxy-based climate field reconstruction methods, J. Geophys. Res.-Atmos., 112, D12109, doi:10.1029/2006JD008272, 2007.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, P. Natl. A. Sci., 105(36), 13252–13257, https://doi.org/10.1073/pnas.0805721105 , 2008.

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F.: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly, Science, 326, 1256–1260, DOI: 10.1126/science.1177303, 2009a.

Mann, M. E., Woodruff, J. D., Donnelly, J. P., and Zhang, Z.: Atlantic hurricanes and climate over the past 1,500 yr, Nature, 460, 880–883, doi:10.1038/nature08219, 2009b.

Meyer, G.P.: An Alternative Probabilistic Interpretation of the Huber Loss. arXiv [preprint], https://arxiv.org/abs/1911.02088, 2020.

Michel, S., Swingedouw, D., Chavent, M., Ortega, P., Mignot, J., and Khodri, M.: Reconstructing climatic modes of variability from proxy records using ClimIndRec version 1.0, Geosci. Model Dev., 13, 841–858, https://doi.org/10.5194/gmd-13-841-2020, 2020.

Nadiga, B.: Reservoir Computing as a Tool for Climate Predictability Studies, J. Adv. Model. Earth. Syst., p. e2020MS002290, https://doi.org/10.1029/2020MS002290, 2020.

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., and Zhang, M.: The mean climate of the Community Atmosphere Model (CAM4) in forced SST and fully coupled experiments, J. Clim., 26, 5150–5168, doi:10.1175/JCLI-D-12-00236.1, 2013.

1) In general, all three methods display similar skills when using ideal (noise-free) pseudoproxies, while in the more realistic PPEs (noise contaminated PPEs), both PCR and CCA method exhibit an overestimation on temperature variance preservation, in contrast to the nonlinear Bi-LSTM.
2) The pseudoproxy networks used in this study were mostly located in the extratropical regions with only three proxies in the tropical area. All CFR methodologies produce generally good reconstructions in regions where dense pseudoproxy networks are available. Moreover, teleconnections are explored by these CFR methodologies, leading to some weak spatial reconstruction skills outside of the proxy-sampled regions, for instance the tropical region. The classical linear-based PCR method generally outperforms the Bi-LSTM and CCA method in both spatial and index reconstructions.
3) Here, we could draw a general conclusion that nonlinear artificial neural network method Bi-LSTM employed herein is not superior for CFR reconstructions, at least in our PPEs. In general, Bi-LSTM show worse skill in spatial and temporal CFRs than PCR and CCA, also in capturing extremes. Yet, it is advisable to employ a larger set of nonlinear CFR methods to evaluate different model structures, and further test their performance on CFRs.

Appendix A

The simulation with the model MPI-ESM-P is not part of the standard CMIP5 simulation suite. In the following, we include additional technical details on this simulation. The MPI simulation was started from the year of 100 BC with restart files from a 500-year spin-down simulation experiments forced with constant external conditions representing the year 100 of BC. After 100 BC, variation in volcanic, solar, orbital, and GHG concentrations are implemented. Land usage was held constant until 850 AD with conditions representing those for year 850 AD. The variation of orbital parameters are calculated after the PMIP3-protocol (Schmidt et al. 2011). The solar activity has been rebuilt on the basis of the reconstruction of Vieira et al. 2011 employing the algorithm and scaling outlined in Schmidt et al. 2011 which corresponds to a difference in short-wave top of the atmosphere insolation of 1.25 Wm-2 (– 0.1%) between the 2nd half of the 20th century (1950 – 2000) and the Maunder Minimum (1645 – 1715). Variations in greenhouse gas concentrations related to CO2, N2O and CH4 are after the reconstruction of the PMIP3 protocol – The concentrations were held constant to the values of year 1 AD between 100 BC and 1 AD because the law Dome records does not extend beyond year 1 AD. After 1850 AD also a reconstructed aerosol loading after Stine et al. 2018 were employed to account for transient anthropogenic aerosol emissions. The extension and reconstruction of the volcanic forcing is related to a rescaling of the newly available Sigl et al. (2015) dataset to the reconstruction of Crowley and Unterman (2013). The large volcanoes for different latitudinal bands are rescaled according to sulfate concentrations and eventually the Crowley algorithm was applied to yield aerosol optical depths and effective radius for four latitudinal bands separated by 30°.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

Otto-Bliesner, B. L., Brady, E. C., Fasullo, J., Jahn, A., Landrum, L., Stevenson, S., Rosenbloom, N., Mai, A., and Strand, G.: CLIMATE VARIABILITY AND CHANGE SINCE 850 CE An Ensemble Approach with the Community Earth System Model, B. Am. Meteorol. Soc., 97, 735–754, https://doi.org/10.1175/bamsd-14-00233.1, 2016.

PAGES 2k Consortium: Continental-scale temperature variability during the last two millennia, Nat. Geosci., 6, 339–346, doi:10.1038/ngeo1797, 2013.

PAGES 2k Consortium: A global multiproxy database for temperature reconstructions of the Common Era, Sci. Data, 4, 170088, https://doi.org/10.1038/sdata.2017.88, 2017.

PAGES 2k Consortium: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nat. Geosci., 536, 411, https://doi.org/10.1038/s41561-019-0400-0, 2019.

PAGES Hydro2k Consortium: Comparing proxy and model estimates of hydroclimate variability and change over the Common Era, Clim. Past, 13, 1851–1900, https://doi.org/10.5194/cp-13-1851-2017, 2017.

Pongratz, J., Reick, C., Raddatz, T., and Claussen, M.: A reconstruction of global agricultural areas and land cover for the last millennium, Global Biogeochem. Cy., 22, GB3018, https://doi.org/10.1029/2007GB003153, 2008.

Po-Chedley, S., Santer, B. D., Fueglistaler, S., Zelinka, M., Cameron-Smith, P., Painter, J., and Fu, Q.: Natural variability contributes to model-satellite differences in tropical tropospheric warming. Proc. Natl Acad. Sci., 118(13), e2020962118, https://doi.org/10.1073/pnas.2020962118, 2020.

Ramachandran P., Zoph B., and Le QV.: Searching for activation functions. arXiv [preprint], https://arxiv.org/abs/1710.05941, 2017.

Pyrina, M., Wagner, S., and Zorita, E.: Pseudo-proxy evaluation of climate field reconstruction methods of North Atlantic climate based on an annually resolved marine proxy network, Clim. Past, 13, 1339–1354, https://doi.org/10.5194/cp-13-1339-2017, 2017.

Qasmi, S., Cassou, C., and Boé, J.: Teleconnection Between Atlantic Multidecadal Variability and European Temperature: Diversity and Evaluation of the Coupled Model Intercomparison Project Phase 5 Models, Geophys. Res. Lett., 44, 11–140, https://doi.org/10.1002/2017GL074886, 2017.

Rasp, S. and Lerch, S.: Neural Networks for Postprocessing Ensemble Weather Forecasts, Mon. Weather Rev., 146, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1, 2018.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y.: Tackling climate change with machine learning, arXiv [preprint], arXiv:1906.05433. https://arxiv.org/abs/1906.05433, 2019.

Reick, C. H., Raddatz, T., Brovkin, V., and Gayler, V.: Representation of natural and anthropogenic land cover change in MPIESM, J. Adv. Model. Earth Syst., 5, 459– 482, doi:10.1002/jame.20022, 2013.

Schmidt, G. A.: Enhancing the relevance of paleoclimatic model/data comparisons for assessments of future climate change, J Quaternary Sci., 25, 79–87, doi:10.1002/jqs.1314, 2010.

Appendix B

We have explored a range of Bi-LSTM architectures, including employing different network depths, introducing dropout layers, using different learning rates, and employing different loss functions to provide a more comprehensive evaluation of the Bi-LSTM performance and effectiveness. Table 1B-6B present reconstruction statistics skill for the spatial North Hemisphere mean temperature in the verification period for ideal PPEs based on CAM5 using different architecture settings of Bi-LSTM method. In our PPE tests on paleo CFRs, it seems that in this case we could not univocally identify optimal neural network structure that could universally outperform all others. And the final Bi-LSTM architecture employed in our CFR experiments was finally determined with 2 hidden layers with 4000 hidden nodes, learning rate is 10^{-3}, activation function is leaky relu, batchsize is 20 and Huber loss function.

Table 1B. Different loss function conditioned on other parameters fixed (2 hidden layers with 4000 hidden nodes, learning rate is 10^{-3}, activation function is leaky relu, batchsize is 20)

Loss functions	*cc*	SD Ratio
MAE	0.483	0.670
MAPE	0.124	0.050
MSE	0.465	0.759
Huber	0.462	0.770

MAE: mean absolutely error, MAPE: mean absolutely percentage error, MSE: mean square error, Huber: Huber loss

Table 2B. Different learning rate using Huber loss, and with the rest parameters fixed as in Table 1B

Learning rates	*cc*	SD Ratio
1e-1	-7e-3	1e7
1e-4	0.462	0.770
1e-6	0.462	0.675
1e-8	0.012	0.271

Table 3B. Different activation functions with the rest parameters fixed as in Table 1B

Activation function	*cc*	SD Ratio
ReLU	0.505	0.566
Leaky ReLU	0.462	0.770
ELU	0.529	0.617

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

Schmidt, G. A., Jungclaus, J. H., Ammann, C. M., Bard, E., Braconnot, P., Crowley, T. J., Delaygue, G., Joos, F., Krivova, N. A., Muscheler, R., Otto-Bliesner, B. L., Pongratz, J., Shindell, D. T., Solanki, S. K., Steinhilber, F., and Vieira, L. E. A.: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0), Geosci. Model Dev., 4, 33–45, https://doi.org/10.5194/gmd-4-33-2011, 2011.

Smerdon, J. E.: Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments, Wiley Interdisciplinary Reviews, Clim. Change., 3, 63–77, https://doi.org/10.1002/wcc.149, 2012.

Smerdon, J. E., Kaplan, A., Zorita, E., Gonzalez-Rouco, J. F., and Evans, M. N.: Spatial performance of four climate field reconstruction methods targeting the Common Era, Geophys. Res. Lett., 38, L11705, doi:10.1029/2011GL047372, 2011.

Smerdon, J. E., Coats, S., and Ault, T. R.: Model-dependent spatial skill in pseudoproxy experiments testing climate field reconstruction methods for the Common Era, Clim. Dynam., 46, 1921–1942, https://doi.org/10.1007/s00382-015-2684-0, 2016.

Smith, R., Jones, P., Briegleb, B., Bryan, F., Danabasoglu, G., Dennis,J., Dukowicz, J., Eden, C., Fox-Kemper, B., and Gent, P.: The Parallel Ocean Program (POP) Reference Manual Ocean Component of the Community Climate System Model (CCSM) and Community Earth System Model (CESM), Rep. LAUR-01853, 1–141, 2010.

Sutton, R. T. and Hodson, D. L. R.: Atlantic ocean forcing of North American and European summer climate, Science, 309, 115– 118, doi:10.1126/science.1109496, 2005.

St. George, S.: An overview of tree-ring width records across the Northern Hemisphere, Quaternary Sci. Rev., 95, 132–150, https://doi.org/10.1016/j.quascirev.2014.04.029, 2014.

St. George, S. and Esper, J.: Concord and discord among Northern Hemisphere paleotemperature reconstructions from tree rings, Quat. Sci. Rev., 203, 278–281, https://doi.org/10.1016/j.quascirev.2018.11.013, 2019.

Schneider, T., Lan, S., Stuart, A., and Teixeira, J.: Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations, Geophys. Res. Lett., 44, 12396–12417, https://doi.org/10.1002/2017GL076101, 2018.

Su, H., Zhang, T., Lin, M., Lu, W., Yan, X. H.: Predicting subsurface thermohaline structure from remote sensing data based on long short-term memory neural networks, Remote Sens. Environ., 260, 112465 10.1016/j.rse.2021.112465, 2021.

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., and Roe, G. H.: Assimilation of time-averaged pseudoproxies for climate reconstruction, J. Clim., 27, 426–441, https://doi.org/10.1175/JCLI-D-12-00693.1, 2014.

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E.: Atmospheric component of the MPI-M Earth System Model: ECHAM6-HAM2, J. Adv. Model. Earth Syst., 5, 146–172, doi:10.1002/jame.20015, 2013.

Sigl, M., Winstrup, M., McConnell, J. R., Welten, K C., Plunkett, G., Ludlow, F., Büntgen, U., Caffee, M., Chellman, N., Dahl-Jensen, D., Fischer, H., Kipfstuhl, S., Kostick, C., Maselli, O. J., Mekhaldi, F., Mulvaney, R., Muscheler, R., Pasteris,

PReLU	0.509	0.544

Table 4B. Different hidden layer number with the rest parameters fixed as in Table 1B

Number of layers	*cc*	SD Ratio
1	0.508	0.733
2	0.462	0.770
4	0.442	0.603
6	0.335	0.411

Table 5B. Different hidden node numbers in each layer with the rest parameters fixed as in Table 1B

Number of hidden nodes	*cc*	SD Ratio
200	0.479	0.620
1000	0.502	0.692
2000	0.503	0.711
4000	0.462	0.770

Table 6B. With and without dropout layers conditioned on the rest parameters are fixed as in Table 1B

Dropout	*cc*	SD Ratio
Dropout	0.462	0.770
Non-dropout	0.467	0.760

Appendix C

Appendix C displays the SD ratios for ideal pseudo-proxies after filtering the reconstructed and target fields with a 30-year low pass filter. At these time scale, the SD ratio is again lower than for the interannual variance.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

D. R., Pilcher, J. R., Salzer, M., Schüpbach, S., Steffensen, J. P., Vinther, B. M., and Woodruff, T. E.: Timing and climate forcing of volcanic eruptions for the past 2,500 years, Nature, 523, 543–549, doi:10.1038/nature14565, 2015.

von Storch, H., Zorita, E., Jones, J. M., Dimitriev, Y., González-Rouco, F., and Tett, S. F.: Reconstructing past climate from noisy data, Science, 306, 679–682, DOI: 10.1126/science.1096109, 2004.

Vieira, L. E. A., Solanki, S. K., Krivova, N. A., and Usoskin, I.: Evolution of the solar irradiance during the Holocene, Astron. Astrophys., 531, A6, doi:10.1051/0004-6361/201015843, 2011.

Wang, J., Emile-Geay, J., Guillot, D., Smerdon, J. E., and Rajaratnam, B. : Evaluating climate field reconstruction techniques using improved emulations of real-world conditions, Clim. Past, 10, 1–19, https://doi.org/10.5194/cp-10-1-2014, 2014.

Wang, J., Yang, B., Ljungqvist, F. C., Luterbacher, J., Osborn, T. J., Briffa, K. R., and Zorita, E.: Internal and external forcing of multidecadal Atlantic climate variability over the past 1,200 years, Nat. Geosci., 10, 512–517, https://doi.org/10.1038/ngeo2962, 2017.

Wilson, R., Anchukaitis, K., Briffa, K. R., Buentgen, U., Cook, E., D'Arrigo, R., Davi, N., Esper, J., Frank, D., Gunnarson, B., Hegerl, G., Helama, S., Klesse, S., Krusic, P. J., Linderholm, H. W., Myglan, V., Osborn, T. J., Rydval, M., Schneider, L., Schurer, A., Wiles, G., Zhang, P., and Zorita, E.: Last millennium Northern Hemisphere summer temperatures from tree rings: Part I: The long term context, Quaternary Sci. Rev., 134, 1–18, https://doi.org/10.1016/j.quascirev.2015.12.005, 2016.

Widmann, M.: One-Dimensional CCA and SVD, and Their Relationship to Regression Maps, J. Climate, 18, 2785–2792, https://doi.org/10.1175/jcli3424.1, 2005.

Wagner, S. and Zorita, E.: The influence of volcanic, solar and CO2 forcing on the temperatures in the Dalton Minimum (1790–1830): a model study, Clim. Dynam., 25, 205–218, doi:10.1007/s00382-005-0029-0, 2005.

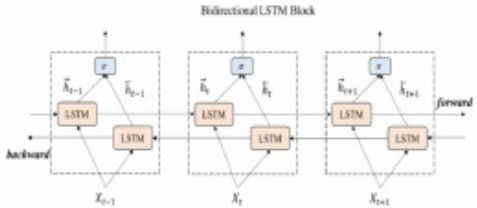

Figure 1: the bidirectional structure of the Bi-LSTM network.

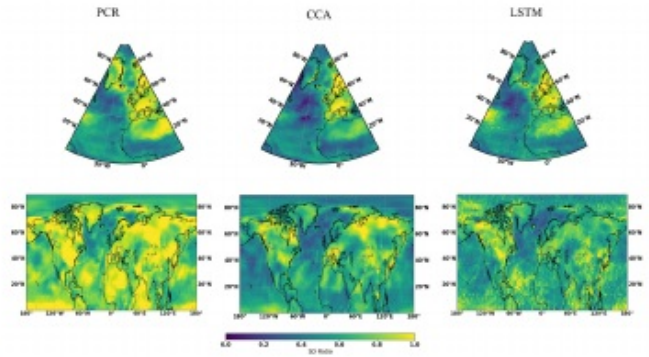

Figure 1C: 30year filtered SD ratio pattern using Ideal-PPEs based on MPI model over validation period 850-1899 for NAE (upper row) and NH (lower row)

Data availability

The MPI-ESM-P model output that was employed for this study is available upon request from the authors S.W or E.Z. And from the Paleoclimatology data repository of the National and Oceanic and Atmospheric Administration in the U.S.A (https://www.ncei.noaa.gov/products/paleoclimatology [Note for the reviewers and editor: these data will be uploaded after publication]). The CESM1 model data can be downloaded from https://www.cesm.ucar.edu/projects/community-projects/LME/.

Author contributions

The analysis was performed by ZZ with the consultation of SW, MK and EZ. ZZ prepared the paper with contributions from all co-authors.

Competing interests

The authors declare that they have no conflict of interest.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

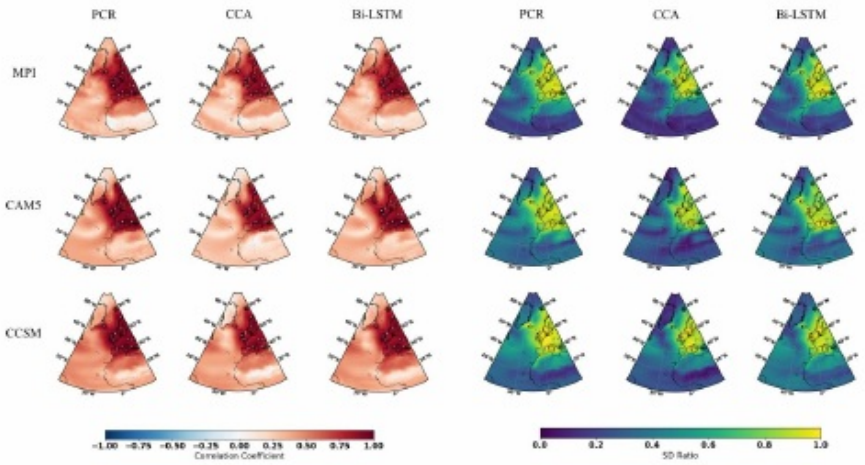

Figure 2: NAE Reconstruction results of CFR methods (including PCR, CCA, Bi-LTSM : Bidirectional long short term memory neural networks) using MPI, CESM1-CAM5 and CCSM numerical simulation as target temperature field, all the CFR methods employ the same proxy network with full 11 ideal pseudoproxies which span the same reconstruction period from 1426-2000 AD. The employed pseudoproxies geolocations are show in white circles in all the sub-figures ; CC is Correlation Coefficient and SD represents Standard Deviation Ratio. The employed pseudoproxies' geolocation is shown as white circles in all the sub-figures.

Acknowledgements

The authors thank the MPI-ESM modelling groups participating in the CMIP5 initiative for providing their data and the CESM1 modelling group for making their data available. This work is funded by the China Scholarship Council (no. 201806570017), and is part of the project Reduced Complexity Models (Redmod), funded by the Helmholtz Association through its Incubator program.

References

Amrhein, D. E., Hakim, G. J., and Parsons, L. A.: Quantifying structural uncertainty in paleoclimate data assimilation with an application to the Last Millennium, Geophys. Res. Lett., 47, e2020GL090485. https://doi.org/10.1029/2020GL090485, 2020.

Anchukaitis, K., Breitenmoser, P., Briffa K., Buchwal, A., Büntgen, U., Cook E., D'Arrigo, R., Esper, J., Evans, M., Frank, D., Grudd ,H., Gunnarson, B., Hughes, M., Kirdyanov ,A., Körner, C., Krusic, P., Luckman, B., Melvin, T., Salzer, M., Shashkin, A., Timmreck, C., Vaganov, E., and Wilson, R.: Tree-rings and volcanic cooling, Nat. Geosci., 5, 836–837, doi:10.1038/ngeo1645, 2012.

Anchukaitis, K. J., Wilson, R., Briffa, K. R., Büntgen, U., Cook, E. R., D'Arrigo, R., Davi, N., Esper, J., Frank, D., Gunnarson, B. E., Hegerl, G., Helama, S., Klesse, S., Krusic, P. J., Linderholm, H. W., Myglan, V., Osborn, T. J., Zhang, P., Rydval, M., Schneider, L., Schurer, A., Wiles, G., and Zorita, E.: Last millennium Northern Hemisphere summer temperatures from tree rings: Part II, spatially resolved reconstructions, Quaternary Sci. Rev., 163, 1–22, https://doi.org/10.1016/j.quascirev.2017.02.020, 2017.

Bengio, Y., Simard, P., and Frasconi, P.: Learning long-term dependencies with gradient descent is difficult, IEEE T. Neural Networ., 5, 157–166, 1994.

Biswas, S., Sinha, M.: Performances of deep learning models for Indian Ocean wind speed prediction, Model. Earth Syst. Environ., 7, 809–831,https://doi.org/10.1007/s40808-020-00974-9, 2021.

Biswas, K., Kumar, S., and Pandey, A. K.: Intensity Prediction of Tropical Cyclones using Long Short-Term Memory Network. arXiv [preprint], https://arxiv.org/abs/2107.03187, 2021.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data Assimilation in the Geosciences: An overview on methods, issues, and perspectives, WIREs Climate Change, 9, e535, https://doi.org/10.1002/wcc.535, 2018.

Chattopadhyay, A., Hassanzadeh, P., and Subramanian, D.: Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network, Nonlin. Processes Geophys., 27, 373–389, https://doi.org/10.5194/npg-27-373-2020, 2020.

Christiansen, B., Schmith, T., and Thejll, P.: A surrogate ensemble study of climate reconstruction methods: stochasticity and robustness, J. Climate, 22, 951–976 doi:10.1175/2008JCLI2301.1, 2009.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

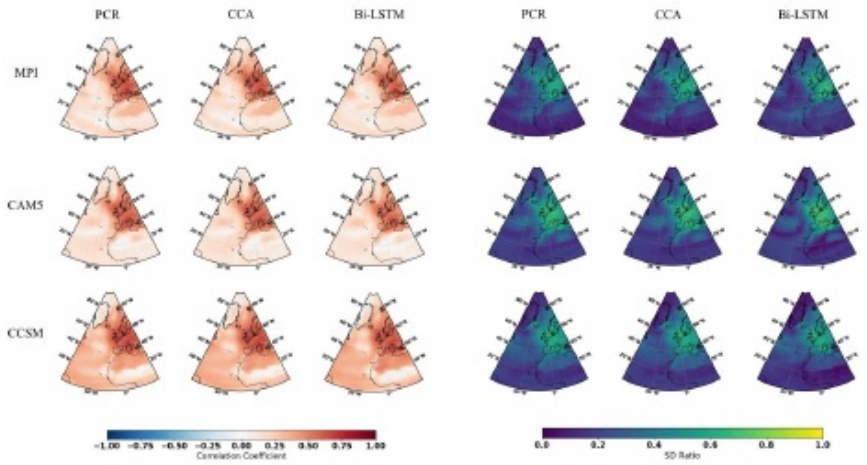

Figure 3: the same as Figure 2, but for employing the full 11 pseudoproxies network with 80% noise contamination.

Christiansen, B.: Reconstructing the NH mean temperature: can underestimation of trends and variability be avoided?, J. Clim., 24, 674–692, 2011.

Christiansen, B. and Ljungqvist, F. C.: The extra-tropical Northern Hemisphere temperature in the last two millennia: reconstructions of low-frequency variability, Clim. Past, 8, 765–786, doi:10.5194/cp-8-765-2012, 2012.

Christiansen, B. and Ljungqvist, F. C.: Challenges and perspectives for large-scale temperature reconstructions of the past two millennia, Rev. Geophys., 55, 40–96, https://doi.org/10.1002/2016RG000521, 2017.

Chattopadhyay, A., Hassanzadeh, P., and Subramanian, D.: Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network, Nonlin. Processes Geophys., 27, 373–389, https://doi.org/10.5194/npg-27-373-2020, 2020.

Coats, S., Smerdon, J. E., Cook, B. I., and Seager, R.: Stationarity of the tropical pacific teleconnection to North America in CMIP5/PMIP3 model simulations, Geophys. Res. Lett., 40, 4927–4932, https://doi.org/10.1002/grl.50938, 2013.

Crowley, T. J. and Unterman, M. B.: Technical details concerning development of a 1200 yr proxy index for global volcanism, Earth Syst. Sci. Data, 5, 187–197, https://doi.org/10.5194/essd-5-187-2013, 2013.

Cui, Z.; Ke, R.; Pu, Z.; Wang, Y. Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. arXiv [preprint], https://arxiv.org/abs/1801.02143, 2018.

Dueben, P. D. and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geosci. Model Dev., 11, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018, 2018.

Emile-Geay, J., McKay, N. P., Kaufman, D. S., Von Gunten, L., Wang, J., Anchukaitis, K. J., and Henley, B. J.: A global multiproxy database for temperature reconstructions of the Common Era, Sci. Data, 4, 170088, https://doi.org/10.1038/sdata.2017.88, 2017.

Evans, M., Smerdon, J. E., Kaplan, A., Tolwinski-Ward, S., and González-Rouco, J. F.: Climate field reconstruction uncertainty arising from multivariate and nonlinear properties of predictors, Geophys. Res. Lett., 41, 9127–9134, https://doi.org/10.1002/2014gl062063, 2014.

Frank, D., Esper, J., Zorita, E., and Wilson, R.: A noodle, hockey stick, and spaghetti plate: a perspective on high-resolution paleoclimatology, Wiley Interdisciplinary Reviews: Climate Change, 1, 507–516, 2010.

Folland, C. K., Knight, J., Linderholm, H. W., Fereday, D., Ineson, S., and Hurrell, J. W.: The Summer North Atlantic Oscillation: Past, present, and future, J. Climate, 22(5), 1082–1103, doi:10.1175/2008JCLI2459.1, 2009.

Gao, C., Robock, A., and Ammann, C.: Volcanic forcing of climate over the past 1500 years: An improved ice core-based index for climate models, J. Geophys. Res.-Atmos., 113, D23111, https://doi.org/10.1029/2008JD010239, 2008.

Gómez-Navarro, J. J., Zorita, E., Raible, C. C., and Neukom, R.: Pseudo-proxy tests of the analogue method to reconstruct spatially resolved global temperature during the Common Era, Clim. Past, 13, 629–648, https://doi.org/10.5194/cp-13-629-2017, 2017.

Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural and other neural network architectures, Neural Netw 18(5), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042, 2005.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

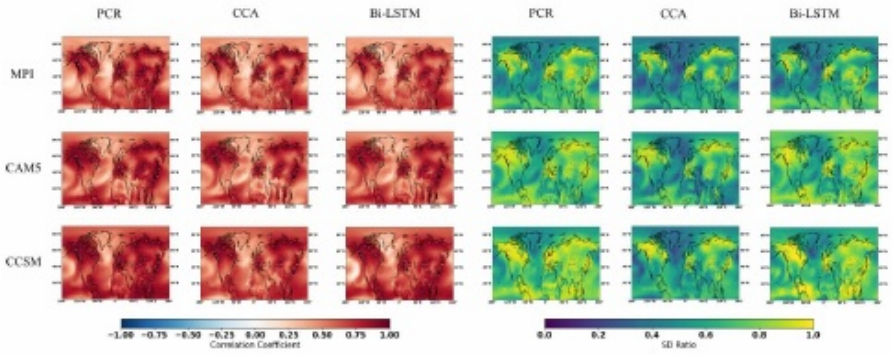

Figure 4: NH Reconstruction results of CFR methods (including PCR, CCA, Bi-LSTM: Bidirectional long short term memory neural networks) using MPI, CESM1-CAM5 and CCSM numerical simulation as target temperature field, all the CFR methods employ the same proxy network with full 48 ideal pseudoproxies which span the same reconstruction period from 1426-2000 AD. The employed pseudoproxies geolocations based on TRW are shown in white circles in all the sub-figures; CC is Correlation Coefficient and SD represents Standard Deviation Ratio. The employed pseudoproxies' geolocation is shown as white circles in all the sub-figures .

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., and Vertenstein, M.: The community climate system model version 4, J. Clim., 24, 4973–4991, doi:10.1175/2011JCLI4083.1, 2011.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning, MIT Press, 2016.

Harlim, J.: Model error in data assimilation, in: Nonlinear and stochastic climate dynamics, edited by: Franzke, C. L. E. and O'Kane, T. J., Cambridge University Press, 276–317, https://doi.org/10.1017/9781316339251.011, 2017.

Hochreiter, S., and Schmidhuber, J. Long short-term memory, Neural Comput., 9, 1735–1780, DOI: 10.1162/neco.1997.9.8.1735, 1997.

Hegerl, G. C., Crowley, T. J., Hyde, W. T., and Frame, D. J.: Climate sensitivity constrained by temperature reconstructions over the past seven centuries, Nature, 440, 1029–1032, https://doi.org/10.1038/nature04679, 2006.

Hegerl, G., Crowley, T., Hyde, W. T., and Frame, D. J.: Uncertainty in climate-sensitivity estimates (Reply), Nature, 446, E2, https://doi.org/10.1038/nature05708, 2007.

Hernández, A., Martin-Puertas, C., Moffa-Sánchez, P., Moreno-Chamarro, E., Ortega, P., Blockley, S., Cobb, K. M., Comas-Bru, L., Giralt, S., Goosse, H., Luterbacher, J., Martrat, B., Muscheler, R., Parnell, A., Pla-Rabes, S., Sjolte, J., Scaife, A. A., Swingedouw, D., Wise, E., and Xu, G.: Modes of climate variability: Synthesis and review of proxy-based reconstructions through the Holocene, Earth Sci. Rev., 271, 103286, https://doi.org/10.1016/j.earscirev.2020.103286, 2020.

Hornik, K., Stinchcombe, M., and White, H.: Multilayer feedforward networks are universal approximators, Neural Netw, 2, 359–366, https://doi.org/10.1016/0893-6080(89)90020-8, 1989.

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., and Yang, H.: Machine learning and artificial intelligence to aid climate change research and preparedness, Environ. Res. Lett., 14, 124007, https://doi.org/10.1088/1748-9326/ab4e55, 2019.

Huang, Y., Yang, L., and Fu, Z.: Reconstructing coupled time series in climate systems using three kinds of machine-learning methods, Earth Syst. Dynam., 11, 835–853, https://doi.org/10.5194/esd-11-835-2020, 2020.

Hunke, E., Lipscomb, W., Turner, A., Jeffery, N., and Elliott, S.: CICE: the Los Alamos sea ice model, documentation and software, version 4.0. Los Alamos National Laboratory, Tech. Rep, LA-CC-06-012, 2008. Los Alamos National Laboratory, Los Alamos, NM.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J. F., Large, W. G., Lawrence, D.,
Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, B. Am. Meteorol. Soc., 94, 1339–1360, https://doi.org/10.1175/bams-d-12-00121.1, 2013.

Hurtt, G. C., Chini, L. P., Frolking, S., Betts, R., Feddema, J., Fischer, G., Goldewijk, K. K., Hibbard, K., Janetos, A., and Jones, C.: Harmonisation of global land-use scenarios for the period 1500–2100 for IPCC-AR5, ILEAPS Newsletter, No. 7, 6–8, 2009.

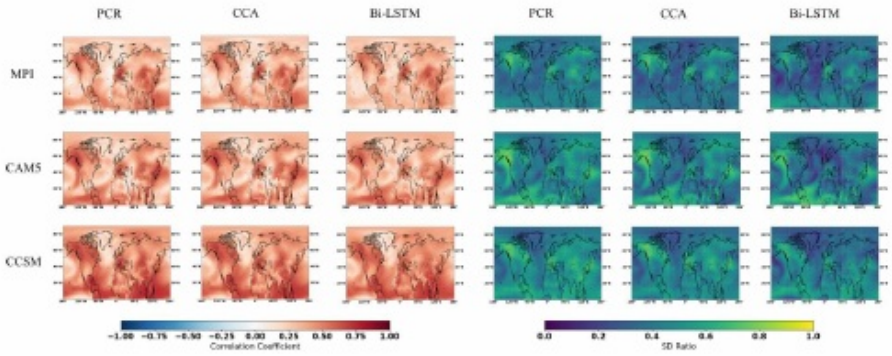

Figure 5: the same as Figure 4, but for employing the full 48 pseudoproxies network with 80% noise contamination.

Hotelling, H.: The relations of the newer multivariate statistical methods to factor analysis, Brit. J. Statist. Psych., 10, 69–76, https://doi.org/10.1111/j.2044-8317.1957.tb00179.x, 1957.

Hodges, J. L.: The significance probability of the Smirnov two-sample test, Arkiv för Matematik, 3, 469–486, 1958.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., and Núñez-Riboni, I.: Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations, J. Adv. Model. Earth Syst., 5, 287–315, 2013.

Jacobeit, J., Wanner, H., Luterbacher, J., Beck, C., Philipp, A., and Sturm, K.: Atmospheric circulation variability in the NorthAtlantic-European area since the mid-seventeenth century, Clim. Dynam., 20, 341–352, https://doi.org/10.1007/s00382-002-0278-0, 2003.

Jahangir, H., Tayarani, H., Gougheri, S.S., Golkar, M.A., Ahmadian, A., Elkamel, A.: Deep Learning-based Forecasting Approach in Smart Grids with Micro-Clustering and Bi-directional LSTM Network. IEEE Trans. Ind. Electron., 68, 8298–8309, doi: 10.1109/TIE.2020.3009604, 2020.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A., and Weston, P.: On the representation error in data assimilation, Q. J. Roy. Meteor. Soc., 144, 1257–1278, https://doi.org/10.1002/qj.3130, 2018.

Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vinther, B. M., Luterbacher, J., Wahl, E. R., Zwiers,
F. W., Mann, M. E., Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Goosse, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., Küttel, M., Mosley-Thompson, E., Overpeck, J. T., Riedwyl, N., Schulz, M., Tudhope, A. W., Villalba, R., Wanner, H., Wolff, E., and Xoplaki, E.: Highresolution palaeoclimatology of the last millennium: A review of current status and future prospects, Holocene, 19, 3–49, doi:10.1177/0959683608098952, 2009.

Jones, P. D. and Mann, M. E.: Climate over past millennia, Rev. Geophys., 42, RG2002, doi:2010.1025/2003RG000143, 2004.

Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., Mikolajewicz, U., Notz, D., and vonStorch, J. S.: Characteristics of the ocean simulations in MPIOM, the ocean component of the MPI-Earth system model, J. Adv. Model. Earth Syst., 5, 422–446, doi:10.1002/jame.20023, 2013.

Knerr, S., Lé, P., and Dreyfus, G.: Single-layer learning revisited: a stepwise procedure for building and training a neural network, in: Neurocomputing, Springer, Berlin, Heidelberg, 41–50, DOI: 10.1007/978-3-642-76153-9_5, 1990.

Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv [preprint], https://arxiv.org/abs/1412.6980, 2014.

Knight, J. R., Folland, C. K., and Scaife, A. A.: Climate impacts of the Atlantic Multidecadal Oscillation, Geophys. Res. Lett., 33, L17706, doi:10.1029/2006GL026242, 2006.

Kadow, C., Hall, D. M., and Ulbrich, U.: Artificial intelligence reconstructs missing climate information, Nat. Geosci., 13, 408–413, https://doi.org/10.1038/s41561-020-0582-5, 2020.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

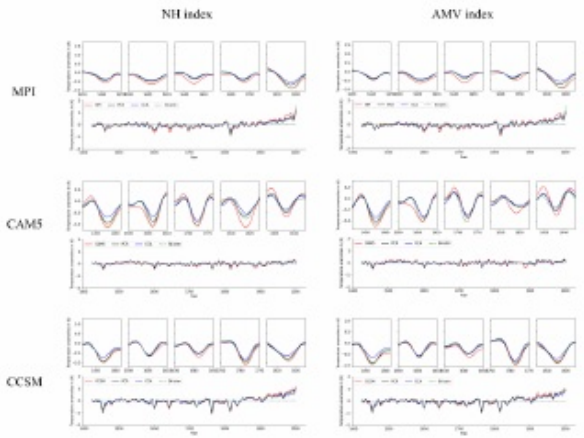

Figure 6: mean time series evolution of the validated reconstructions for NH summer temperature anomaly and Atlantic Multidecadal Variability (AMV) index using full 48 ideal-pseudoproxies based on PCR, CCA, Bi-LSTM CFR methods. All time series have been smoothed using a Butterworth low-pass filter to remove temporal fluctuations less than 10 years. MPI, CAM5 and CCSM represent MPI/CAM5/CCSM model simulated 'true' climatology. We selected several reconstructed extreme cooling period with a shorter interval (each 10 years are selected before and after the specific extreme cooling year) and plotted them above each entire reconstruction means amongst models and methods.

Lean, J., Rottman, G., Harder, J., and Kopp, G.: SORCE contributions to new understanding of global change and solar variability, in: The Solar Radiation and Climate Experiment (SORCE), Springer, 2005.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H.: European seasonal and annual temperature variability, trends, and extremes since 1500, Science, 303, 1499–1503, DOI: 10.1126/science.1093877, 2004.

Luterbacher J., Werner, J. P., Smerdon, J. E., Fernández-Donado, L., González-Rouco, F. J., Barriopedro, D., Ljungqvist, F. C., Büntgen, U., Zorita, E., Wagner, S., Esper, J., McCarroll, D., Toreti, A., Frank, D., Jungclaus, J. H., Barriendos, M., Bertolin, C., Bothe, O., Brázdil, R., Camuffo, D., Dobrovolný, P., Gagen, M., García-Bustamante, E., Ge, Q., Gómez-Navarro, J. J., Guiot, J., Hao, Z., Hegerl, G. C., Holmgren, K., Klimenko, V. V., MartínChivelet, J., Pfister, C., Roberts, N., Schindler, A., Schurer, A., Solomina, O., von Gunten, L., Wahl, E., Wanner, H., Wetter, O., Xoplaki, E., Yuan, N., Zanchettin, D., Zhang, H., and Zerefos, C.: European summer temperatures since Roman times, Environ. Res. Lett., 11, 024001, doi:10.1088/1748-9326/11/2/024001, 2016.

Lindgren, A., Lu, Z., Zhang, Q., and Hugelius, G.: Reconstructing past global vegetation with random forest machine learning, sacrificing the dynamic response for robust results J. Adv. Model. Earth. Syst., 13, p.e2020MS002200. https://doi.org/10.1029/2020MS002200, 2021.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Fletcher, C. G., Lawrence, P. J., Levis, S., Swenson, S. C., and Bonan, G. B.: The CCSM4 land simulation, 1850–2005: Assessment of surface climate and new capabilities, J. Clim., 25, 2240–2260, doi:10.1175/JCLI-D-11-00103.1, 2012.

Li, B. and Smerdon, J. E.: Defining spatial comparison metrics for evaluation of paleoclimatic field reconstructions of the Common Era, Environmetrics, 23, 394–406, doi:10.1002/env.2142, 2012.

Lukoševičius, M. and Jaeger, H.: Reservoir computing approaches to recurrent neural network training, Comput. Sci. Rev., 3, 127–149, https://doi.org/10.1016/j.cosrev.2009.03.005, 2009.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779–787, https://doi.org/10.1038/33859, 1998.

Mann, M. E. and Rutherford, S.: Climate reconstruction using "Pseudoproxies", Geophys. Res. Lett., 29, 1501, https://doi.org/10.1029/2001GL014554, 2002.

Mann, M. E. and Jones, P. D.: Global surface temperatures over the past two millennia, Geophys. Res. Lett., 30, 1820, https://doi.org/10.1029/2003GL017814, 2003.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Testing the fidelity of methods used in proxy-based reconstructions of past climate, J. Clim., 18, 4097–4107, https://doi.org/10.1175/JCLI3564.1, 2005.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Robustness of proxy-based climate field reconstruction methods, J. Geophys. Res.-Atmos., 112, D12109, doi:10.1029/2006JD008272, 2007.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, P. Natl. A. Sci., 105(36), 13252–13257, https://doi.org/10.1073/pnas.0805721105 , 2008.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

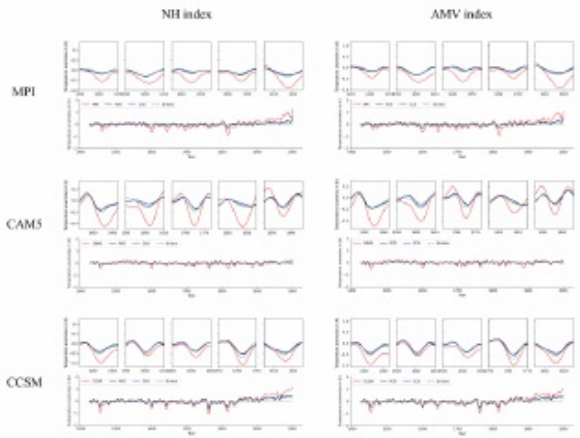

Figure 7: The same as Figure 6, but derived from 80% noise contaminated PPEs.

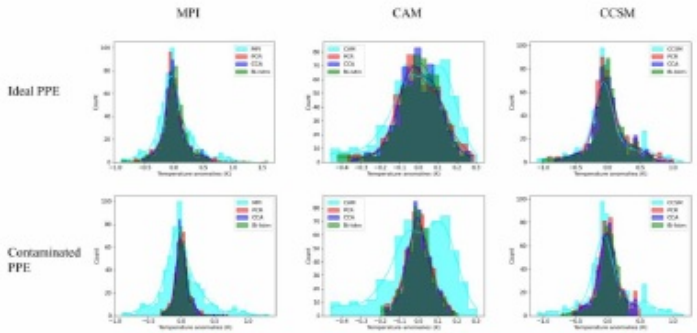

Figure 8: Histogram for decadal filtered NH mean index. The x axis denotes temperature anomaly values, and y axis is the number of data in each bin. Totally 30 bins are selected to plot each of the histogram.

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F.: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly, Science, 326, 1256–1260, DOI: 10.1126/science.1177303, 2009a.

Mann, M. E., Woodruff, J. D., Donnelly, J. P., and Zhang, Z.: Atlantic hurricanes and climate over the past 1,500 yr, Nature, 460, 880–883, doi:10.1038/nature08219, 2009b.

Meyer, G.P.: An Alternative Probabilistic Interpretation of the Huber Loss. arXiv [preprint], https://arxiv.org/abs/1911.02088, 2020.

Michel, S., Swingedouw, D., Chavent, M., Ortega, P., Mignot, J., and Khodri, M.: Reconstructing climatic modes of variability from proxy records using ClimIndRec version 1.0, Geosci. Model Dev., 13, 841–858, https://doi.org/10.5194/gmd-13-841-2020, 2020.

Nadiga, B.: Reservoir Computing as a Tool for Climate Predictability Studies, J. Adv. Model. Earth. Syst., p. e2020MS002290, https://doi.org/10.1029/2020MS002290, 2020.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, M., Wald, R., and Muharemagic, E.: Deep learning applications and challenges in big data analytics, J. Big Data, 2, 1, https://doi.org/10.1186/s40537-014-0007-7, 2015.

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., and Zhang, M.: The mean climate of the Community Atmosphere Model (CAM4) in forced SST and fully coupled experiments, J. Clim., 26, 5150–5168, doi:10.1175/JCLI-D-12-00236.1, 2013.

Otto-Bliesner, B. L., Brady, E. C., Fasullo, J., Jahn, A., Landrum, L., Stevenson, S., Rosenbloom, N., Mai, A., and Strand, G.: CLIMATE VARIABILITY AND CHANGE SINCE 850 CE An Ensemble Approach with the Community Earth System Model, B. Am. Meteorol. Soc., 97, 735–754, https://doi.org/10.1175/bamsd-14-00233.1, 2016.

Parsons, L. A., Amrhein, D. E., Sanchez, S. C., Tardif, R., Brennan, M. K., & Hakim, G. J.: Do multi-model ensembles improve reconstruction skill in paleoclimate data assimilation? Earth and Space Science. https://doi.org/10.1029/2020ea001467, 2021.

Pongratz, J., Reick, C., Raddatz, T., and Claussen, M.: A reconstruction of global agricultural areas and land cover for the last millennium, Global Biogeochem. Cy., 22, GB3018, https://doi.org/10.1029/2007GB003153, 2008.

Po-Chedley, S., Santer, B. D., Fueglistaler, S., Zelinka, M., Cameron-Smith, P., Painter, J., and Fu, Q.: Natural variability contributes to model-satellite differences in tropical tropospheric warming. Proc. Natl Acad. Sci., 118(13), e2020962118, https://doi.org/10.1073/pnas.2020962118, 2020.

Ramachandran P., Zoph B., and Le QV.: Searching for activation functions. arXiv [preprint], https://arxiv.org/abs/1710.05941, 2017.

Pyrina, M., Wagner, S., and Zorita, E.: Pseudo-proxy evaluation of climate field reconstruction methods of North Atlantic climate based on an annually resolved marine proxy network, Clim. Past, 13, 1339–1354, https://doi.org/10.5194/cp-13-1339-2017, 2017.

DiffPDF • /home/zorita/MANUSCRIPTS/ZEGUO/CP_manu_ZZ_original_version.pdf mit /home/zorita/MANUSCRIPTS/ZEGUO/CP_revision3_ZG.pdf • 2022-07-29

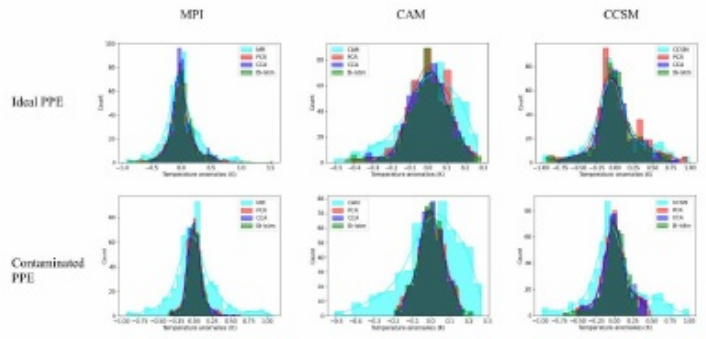

Figure 9: The same as Figure 8, but for decadal filtered AMV index.

Qasmi, S., Cassou, C., and Boé, J.: Teleconnection Between Atlantic Multidecadal Variability and European Temperature: Diversity and Evaluation of the Coupled Model Intercomparison Project Phase 5 Models, Geophys. Res. Lett., 44, 11–140, https://doi.org/10.1002/2017GL074886, 2017.

Rasp, S. and Lerch, S.: Neural Networks for Postprocessing Ensemble Weather Forecasts, Mon. Weather Rev., 146, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1, 2018.

Riley, P.: Three pitfalls to avoid in machine learning, Nature 572, 27–29, https://doi.org/10.1038/d41586-019-02307-y, 2019.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y.: Tackling climate change with machine learning, arXiv [preprint], arXiv:1906.05433. https://arxiv.org/abs/1906.05433, 2019.

Reick, C. H., Raddatz, T., Brovkin, V., and Gayler, V.: Representation of natural and anthropogenic land cover change in MPIESM, J. Adv. Model. Earth Syst., 5, 459– 482, doi:10.1002/jame.20022, 2013.

Schmidt, G. A.: Enhancing the relevance of paleoclimatic model/data comparisons for assessments of future climate change, J Quaternary Sci., 25, 79–87, doi:10.1002/jqs.1314, 2010.

Schmidt, G. A., Jungclaus, J. H., Ammann, C. M., Bard, E., Braconnot, P., Crowley, T. J., Delaygue, G., Joos, F., Krivova, N. A., Muscheler, R., Otto-Bliesner, B. L., Pongratz, J., Shindell, D. T., Solanki, S. K., Steinhilber, F., and Vieira, L. E. A.: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0), Geosci. Model Dev., 4, 33–45, https://doi.org/10.5194/gmd-4-33-2011, 2011.

Schurer, A. P., Hegerl, G. C., Mann, M. E., Tett, S. F. B., and Phipps, S. J.: Separating forced from chaotic climate variability over the past millennium, J. Climate, 26, 6954-6973, doi:10.1175/JCLID-12-00826.1, 2013.

Schurer, A. P., Tett, S. F. B., and Hegerl, G. C.: Small influence of solar variability on climate over the past millennium, Nat. Geosci., 7, 104–108, doi:10.1038/ngeo2040, 2014.

Smerdon, J. E., Kaplan, A., Chang, D., and Evans, M. N.: A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium, J. Climate, 23, DOI: https://doi.org/10.1175/2010JCLI3328.1, 4856–4880, 2010.

Smerdon, J. E.: Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments, Wiley Interdisciplinary Reviews, Clim. Change., 3, 63–77, https://doi.org/10.1002/wcc.149, 2012.

Smerdon, J. E., Kaplan, A., Zorita, E., Gonzalez-Rouco, J. F., and Evans, M. N.: Spatial performance of four climate field reconstruction methods targeting the Common Era, Geophys. Res. Lett., 38, L11705, doi:10.1029/2011GL047372, 2011.

Smerdon, J. E., Cook, B. I., Cook, E. R., and Seager, R.: Bridging past and future climate across paleoclimatic reconstructions,
observations, and models: a hydroclimate case study, J. Climate, 28, 3212–3231, https://doi.org/10.1175/jcli-d-14-00417.1, 2015.