Reconstruction of climate fields using machine-learning methods

Dissertation
with the aim of achieving a doctoral degree
at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Earth Sciences
at Universität Hamburg

submitted by Zeguo Zhang
From Shandong, China

Hamburg, 2022

Department of Earth Sciences

Date of Oral Defense:	**20.02.2023**
Reviewers:	**Prof. Dr. Corinna Schrum**
	Dr. Eduardo Zorita
Members of the examination commission: Chair	**Prof. Dr. Corinna Schrum**
	Dr. Eduardo Zorita
	Prof. Dr. Lars Kutzbach
	PD Dr. Richard Blender
	Dr. Leonard Borchert
Chair of the Subject Doctoral Committee Earth System Sciences:	**Prof. Dr. Hermann Held**
Dean of Faculty MIN:	**Prof. Dr.-Ing. Norbert Ritter**

Abstract

Abstract

Climate research is often constrained by a limited amount of available data, for instance sparse and incomplete point observations. Yet it aims to understand and predict climate variability at large temporal and spatial scales. In those cases, the information provided by those limited and incomplete data sets needs to be interpreted, interpolated and extrapolated by the application of suitable mathematical methods, which are also required to maintain the physical consistency of the data.

The question of *Climate Reconstructions* arises, for example, in the research area of the past climates. Reconstructing past climate variability, can provide us with a better perspective to better understand climate dynamics, for instance, extreme heat waves and sea level rise dynamics. The information about past climates is encoded in indirect indicators, such as tree-rings or ice cores, which are sparse and incomplete records. Other important area is the extrapolation and extension of a network of point observations into a complete spatial field to provide a more accurate picture of particular events. Another area of application of climate field reconstructions (CFRs) is the extrapolation of station data to construct a spatially resolved, complete, and physically consistent, field, to reveal the general spatial and temporal variability of particular climate variables. Climate field reconstruction is usually obtained by linearly mapping the relationship between the local information - proxy records or station data to targeted climate variables based on regression theory. Yet, climate dynamics is nonlinear which compromises the application of a simple linear relationship between the climate variability at different locations. Most standard methods based on linear regression theory used so far might not be able to capture these nonlinearities very well. One usual deficiency of classical linear reconstruction methods is the underestimation of temporal and spatial variability. This deficiency originates in the tendency of all linear methods ‘to regress to the mean’ when the information available is sparse or uncertain: the mean value is the less risky estimation when information is insufficient.

The main objective in this thesis is to test whether the newly emerging machine learning methods, for example, artificial neural networks, could be helpful for capturing a higher degree of underlying linear and nonlinear relationships between target climate variables and proxy, thereby providing better climate field reconstructions. Moreover, their advantage of feature extraction and selection of relevant predictors might be helpful for selecting spatial-temporal features of climate variables to achieve better reconstructions, qualifying machine-learning methods as superior characteristics in mitigating these shortcomings of traditional reconstruction methods. This dissertation, therefore, has a mostly methodological character, aiming at the design of modern machine learning methods to the classical problem of climate field reconstructions.

Three different machine-learning methods for climate field reconstructions were tested in our study. Two of these methods were applied for the reconstruction of the temperature of the past centuries based on proxy data. Each of these two methods - Long-Short-Term Memory Network and Echo State Network - implement machine-learning algorithms that capture the serial correlation structure in the data. This characteristic sets them apart of most other climate reconstruction methods. The third method, Generative Adversarial Networks, was applied to the reconstruction of the spatial sea-level field in the North Sea based on point coastal observations.

Our results indicate in general that when we choose machine-learning methods with appropriate structures and hyper-parameters, comparable or better climate reconstructions can be achieved compared to traditional Climate field reconstruction (CFR) methods. However, they do not lead to a clear-cut improvement of some of the deficiencies of classical reconstruction methods, such as the underestimation of variability.

Zusammenfassung

Zusammenfassung

Die Klimaforschung wird oft durch eine begrenzte Menge verfügbarer Daten eingeschränkt, beispielsweise durch spärliche und unvollständige Punktbeobachtungen. Dennoch zielt sie darauf ab, die Klimavariabilität auf großen zeitlichen und räumlichen Skalen zu verstehen und vorherzusagen. Die Informationen dieser begrenzten und unvollständigen Datensätze müssen durch die Anwendung geeigneter mathematischer Methoden interpretiert, interpoliert und extrapoliert werden, was auch erforderlich ist, um die physikalische Konsistenz der Daten zu erhalten.

Die Frage nach Klimarekonstruktionen stellt sich beispielsweise im Forschungsbereich der vergangenen Klimate. Die Rekonstruktion vergangener Klimaschwankungen kann uns eine bessere Perspektive bieten, um die Klimadynamik besser zu verstehen, beispielsweise extreme Hitzewellen und die Dynamik des Meeresspiegelanstiegs. Die Informationen über vergangene Klimazonen sind in indirekten Indikatoren wie Baumringen oder Eisbohrkernen kodiert, die spärliche und unvollständige Aufzeichnungen sind. Ein weiterer wichtiger Bereich ist die Extrapolation und Erweiterung eines Netzwerks von Punktbeobachtungen in ein vollständiges räumliches Feld, um ein genaueres Bild bestimmter Ereignisse zu liefern. Ein weiteres Anwendungsgebiet von Klimafeldrekonstruktionen (CFRs) ist die Extrapolation von Stationsdaten zur Konstruktion eines räumlich aufgelösten, vollständigen und physikalisch konsistenten Feldes, um die allgemeine räumliche und zeitliche Variabilität bestimmter Klimavariablen aufzudecken. Im Allgemeinen wird die Klimafeldrekonstruktion normalerweise durch lineares Abbilden der Beziehung zwischen den lokalen Informationen – Proxy-Aufzeichnungen oder Stationsdaten – auf Zielklimavariablen basierend auf der Regressionstheorie erhalten. Die Klimadynamik ist jedoch chaotisch, was die Anwendung einer einfachen linearen Beziehung zwischen der Klimavariabilität an verschiedenen Orten beeinträchtigt. Die meisten bisher verwendeten Standardmethoden, die auf der linearen Regressionstheorie basieren, sind möglicherweise nicht in der Lage, diese Nichtlinearitäten sehr gut zu erfassen. Ein üblicher Mangel klassischer linearer Rekonstruktionsverfahren ist die Unterschätzung der zeitlichen und räumlichen Variabilität. Dieser Mangel ergibt sich aus der Tendenz aller linearen Methoden, bei spärlichen oder unsicheren Informationen „auf den Mittelwert zu regressieren“: Der Mittelwert ist die weniger riskante Schätzung, wenn die Informationen unzureichend sind.

Das Hauptziel dieser Arbeit ist es zu testen, ob die neu aufkommenden maschinellen Lernmethoden, zum Beispiel künstliche neuronale Netze, hilfreich sein könnten, um ein höheres Maß an zugrunde liegenden linearen und nichtlinearen Beziehungen zwischen Zielklimavariablen und Proxy zu erfassen und dadurch ein besseres Klimafeld bereitzustellen Rekonstruktionen. Darüber hinaus könnte ihr Vorteil der Merkmalsextraktion und Auswahl relevanter Prädiktoren hilfreich sein, um räumlich-zeitliche Merkmale von Klimavariablen auszuwählen, um bessere Rekonstruktionen zu erzielen, und maschinelle Lernmethoden als überlegene Eigenschaften bei der Minderung dieser Mängel traditioneller Rekonstruktionsmethoden zu qualifizieren. Diese Dissertation hat daher einen überwiegend methodischen Charakter und zielt auf die Gestaltung moderner maschineller Lernmethoden auf das klassische Problem der Klimafeldrekonstruktion ab.

In unserer Studie wurden drei verschiedene maschinelle Lernverfahren für Klimafeldrekonstruktionen getestet. Zwei dieser Methoden wurden für die Rekonstruktion der Temperatur vergangener Jahrhunderte auf der Grundlage von Proxydaten angewendet. Jede dieser beiden Methoden – Long-Short-Term Memory Network und Echo State Network – implementiert maschinelle Lernalgorithmen, die die serielle Korrelationsstruktur in den Daten erfassen. Diese Eigenschaft unterscheidet sie von den meisten anderen Methoden zur Klimarekonstruktion. Die dritte Methode, Generative Adversarial Networks, wurde auf die Rekonstruktion des räumlichen Meeresspiegelfeldes in der Nordsee basierend auf punktuellen Küstenbeobachtungen angewendet.

Unsere Ergebnisse zeigen im Allgemeinen, dass bei der Wahl von maschinellen Lernmethoden mit geeigneten Hyperparametern vergleichbare oder bessere Klimarekonstruktionen im Vergleich zu herkömmlichen Methoden der Klimafeldrekonstruktion (CFR) erzielt werden können. Sie führen jedoch nicht zu einer eindeutigen Verbesserung einiger Mängel klassischer Rekonstruktionsmethoden, wie etwa der Unterschätzung der Variabilität.

Table of Contents

Chapter 1: Introduction

Introduction

Climate reconstructions, for example, surface temperature or sea surface high field reconstructions, can help us to better understand past climate variability and its drivers. Several networks and international projects have aimed at reconstructing the climate of the past centuries and learn from the past. With these inferences/knowledge about past climatic information we could potentially better project the future (The Millennium project, 2006; PAGES 2k Consortium, 2013, 2017, 2019; PAGES Hydro2k Consortium, 2017).

The reconstruction of past climates is based on the information provided by indirect local indicators of past environmental conditions (Mann et al., 2008; Sheldon and Tabor, 2009; Eiler, 2011; Luterbacher et al., 2016). A typical, and rather well known, example are tree-ring data (Fritts, 1976; St. George, 2014; St. George and Esper, 2019). Trees tend to grow wider or denser rings when the environmental conditions are more favorable and thinner and produce less dense rings when these conditions are adverse. The environmental drivers of tree growth are in many cases temperature and/or precipitation during particular seasons in the year and geographical location (Cook et al., 1999; D'Arrigo et al., 2008; St. George, 2014). Living trees or fossil trees can be sampled (non-destructively) and the width of their growth rings can be measured. Additionally the year in which they were formed can be very precisely determined. Other type of proxy records, for instance, those derived from stalagmites, can be geochemically analyzed with respect to their isotopic composition (Tan et al., 2006; Baker et al., 2008; Fairchild and Treble, 2009; Wong and Breeker, 2015). This isotopic composition can also yield information about the amount of precipitation in one particular year or periods in the past.

The way that this indirect information (tree-ring width) is translated into physical/meteorological variables is by statistically comparing the proxy record with a physical record of temperature (or precipitation) measured in a nearby station during a period of overlap, which is typically the 20th century (Esper et al., 2002; St. George, 2014; Wilson et al., 2016; Anchukaitis et al., 2017). Thus, a type of correlation or regression between both records is established. This analysis allows first to identify which physical variables mainly determines the variability of the proxy record (temperature or precipitation), and it allows then to infer the value of the physical variable for periods in the past for which only the proxy record exists. These statistical methods are not only linear (for instance simple correlations), but comprise a more sophisticated statistical machinery, which hopefully may be more adequate to represent a possibly non-linear relationship between proxy record and physical variable. For instance, the growth of trees may optimally occur in a range of temperatures, and be therefore smaller for colder but also for warmer temperatures beyond this range. In this case, the link between proxy-record and temperature is not a simple proportionality constant (Speer, 2010).

In addition, the variations of the proxy records (tree-ring growth) may be affected by other variables, not just the more common temperature and/or precipitation (Esper et al., 2002; Stoffel and Bollschweiler, 2008; Speer, 2010). The concentrations of nutrients in the soil, the impact of fire, of parasites and of the competition with other trees, all may influence the tree growth in a particular year or series of years. This impact is reflected in a deviation from the growth expected just from considering only the purely climatic drivers. It distorts a clean statistical identification of the signal of these climatic drivers and adds uncertainty to the estimation of past climate from the proxy records.

These confounding, non-climatic, factors and their influence on the proxy record is denoted as '*noise*'. It is easy to illustrate the impact of this noise in the estimation of past climate (von Storch et al., 2004; Christiansen and Ljungqvist, 2017) based, here as an example, on only one proxy record. A standard univariate linear regression model (Su et al., 2012) links the proxy record P as predictor with temperature T as predictand, widely used in dendroclimatology (Sheppard, 2010) is:

$$T(t) = a_0 + a_1 P(t) + noise(t) \tag{1.1}$$

where a_0 and a_1 are the regression coefficients and t is time. The value of these parameters can be estimated by minimizing the variance of the noise term (least squares regression, LSR; Su et al., 2012), using data in an overlapping period. The LSR model assumes that the noise term is not correlated with the predictor, and therefore the variance of T must be smaller than the variance of a_1P.

Once the values of the parameters a_0 and a_1 are estimated as $\widehat{a_0}$ and $\widehat{a_1}$, the temperature in past times t' can be reconstructed by:

$$T(t') = \widehat{a_0} + \widehat{a_1} P(t') \tag{1.2}$$

This reconstruction method then leads to an underestimation of the reconstructed temperature.

One problem with this simple linear regression model is that the conditions for LSR to applicable are not totally fulfilled, since in reality the noise term is indeed correlated with P. The noise term is actually a part of the proxy record P itself, as we have discussed before. A more correct statistical model, which is however, very seldom applied, consists of inverting the role of predictor and predictand:

$$P(t) = b_0 + b_1 T(t) + noise(t) \tag{1.3}$$

This model better reflects the physical process that link temperature and proxy record, as temperature is actually the driver of the variations of tree growth (Boisvenue and Running, 2006), and not the other way around. More relevant here is that the optimal estimated value of $\widehat{a_1}$ and $\widehat{b_1}$ can be shown not to be related by

$$\widehat{b_1} = 1/\widehat{a_1} \tag{1.4}$$

so that both regression models are indeed distinct and not just a simple reformulation.

These type of caveats permeate other more complex models, giving rise to different biases in the estimation of past climate and temperature, which has already been recognized that many classical reconstruction methods present some extent underestimation of low frequency climatic variability, especially in large-scale temperature reconstructions (von Storch et al., 2004; Bürger et al., 2006; Christiansen et al., 2009; Frank et al., 2010; Smerdon et al., 2011; Christiansen, 2011; Tingley et al., 2012; Wang et al., 2014; Evans et al., 2014). von Storch et al. (2004) and Christiansen (2011) demonstrated that some underestimation of climate reconstructions could also be attributed to the employed CFR methods, which indicates that the selection of methodologies might introduce additional uncertainties and biases.

For instance, seven different reconstruction methods, including Principal Components Regression (PCR), Canonical Correlation Analysis (CCA), Regularized Expectation Maximization - RegEM Ridge (Schneider, 2001), RegEM truncated total least squares - TTLS (Mann et al,. 2009a) and so on, were employed by Christiansen et al. (2009) for reconstruction studies, they found that a general underestimation tendency appeared in the reconstructed amplitude of the low-frequency variability amongst all methods. Specifically, all seven reconstruction approaches produce underestimation and large biases both in the amplitude and trend of low frequency variability. Christiansen and Ljungqvist (2017) also emphasized that most employed climate reconstruction methods until now have a general assumption that there is a linear relationship or response between proxy and target climate variable. However, nonlinearity may exist in nature, therefore this general assumption could lead to additional error or bias of the amplitude of reconstruction variability. They made a general conclusion that underestimation of low frequency climate variability is a serious issue or deficit for most reconstruction methodologies. These methods also present similar issues in climate reconstructions that are not in the range of calibration time interval. They revealed that the RegEM-TTLS and Local approach (reconstructions by indirect regression the local surface temperature at positions were proxies are available, Christiansen, 2011) could present superiority compared with most other previous and present reconstruction methods. However, the RegEM-TTLS method still produces underestimation of variability outside the range of calibration time interval, and the LOC method depends strongly on a strictly screening process and spatial and temporal averaging.

In addition, climatic variables, such as surface temperature field or sea surface height, usually present a complicated spatial covariance structure (Jones and Briffa, 1996) and these structures may even be nonstationary, i.e. it may change with time. Many reconstruction methods, for instance, PCR or CCA, have an assumption that some domain climate variability patterns are constant in time (Gómez-Navarro et al., 2017; Pyrina et al., 2017). Nevertheless, this assumption of stationarity between temperature and proxy is difficult to evaluate (Evans et al., 2013); these assumptions and potential uncertainties may lead to reconstruction bias.

Additional reconstruction errors or biases could be introduced if we do not choose a method appropriately, which means that we might need a comprehensive understanding of the relationship between proxy and target climate variable, and the spatial-temporal structure of target climate field. Tingley et al. (2012) attempted to involve specific covariance models into their reconstructions, but there is a challenge since spatial and temporal correlations of climate variables are rather entangled. Data assimilation (Steiger et al., 2014; Carrassi et al., 2018) methods may be helpful for this challenge. 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. On the other hand, they have the advantage that they make use of the physical relationships between climate variables as encapsulated in Earth System models. However, 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 previously mentioned shortcomings of previous and present climate reconstruction methods, we introduce the newly emerging machine learning techniques into the area of climate reconstructions. The machine learning methods, for example, Artificial Neural Network - ANN, have been demonstrated to be capable of capturing more potential nonlinearities from dynamical physical systems (Schneider et al., 2018; Rasp and Lerch, 2018; Rolnick et al., 2019; Chattopadhyay et al., 2020; Huang et al., 2020; Nadiga, 2020; Lindgren et al., 2021). Specifically, three machine leaning methods, including the Reservoir computing (RC), RNN-LSTM and a feed-forward neural network, are employed for reproducing the long-term statistics and short-term evolution of a multiscale spatial-temporal nonlinear Lorenz 96 system by Chattopadhyay et al. (2020). They revealed that these three methods could present some skillful predictions, while the Reservoir computing- Echo stet network (RC-ESN) could produce accurately forecasting of the chaotic trajectories and substantially outperform the other two methods. Rasp and Lerch (2018) demonstrated that neural network methods could incorporate potential nonlinear relationships between predictor climatic variables and prediction distribution parameters that can be captured automatically in a purely data-driven way instead of requiring pre-specified link functions. Gordon et al. (2021) demonstrated that ANN is able to employ oceanic patterns that have been previously connected to predictable Pacific Decadal Oscillation behavior, which could support the further utilization of ANN in climate and ocean studies.

Another advantage of nonlinear machine-learning is that the assumption of climate proxy networks are linear or nonlinear related with climate variables is unnecessary. A nonlinear method will map the relationships between the predictor and predictand automatically with necessary model parameters and updated weights (Goodfellow, et al 2016). In addition, they do not rely on statistical (for instance the PCA) methods for the purpose of raw data preprocessing. The inherent variabilities in calibration dataset is sequentially and dynamically adjusted and retained with the optimized hyperparameters of nonlinear methods, the nonlinear CFR method will take all the calibration/training dataset as model input information and extract principal features by its trained hyper-parameters automatically. Eventually, validation is directly derived using this trained model which weights and thresholds being relatively robust and fixed. Non-linear CFRs can also enhance our insight of complex multivariate connections in chaotic nonlinear systems compared to purely data-driven learning methods, yet avoiding to specify a process-based model (Rasp, S. et al, 2018; Huntingford, C, et al. 2019; Rolnick, D, et al. 2019). Figure 1.1 illustrates some potential machine learning (ML) and Artificial Intelligence (AI) based understanding enhances next generation of Earth System Models.

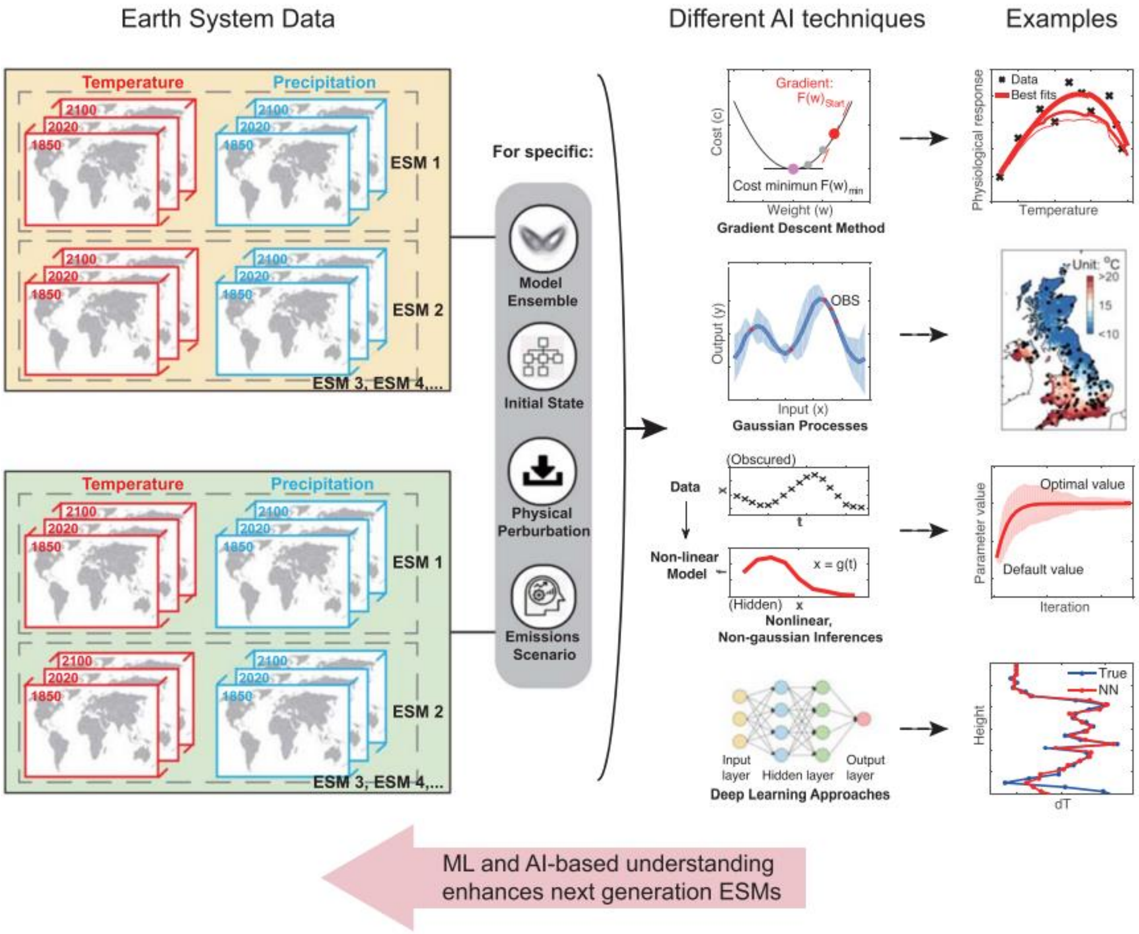

Figure 1.1: Schematic of different machine learning methods with potential applications in earth system science from Huntingford, C, et al. (2019)

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 is that the neural network methods employed in this thesis would be able to better capture episodes of larger deviations from the mean, especially those that stretch over several time steps. Also for the employed convolutional neural network (CNN) application in sea surface height field reconstructions, it has been proved that CNN were capable of restoring or filling missing climate information with reasonable accuracy (Dong, J. et al., 2019; Kadow et al., 2020; Barth et al., 2020 and 2022). Thus, we assume that the neural network methods employed and tested in this thesis could produce more accurate or reasonable reconstruction results in order to overcome or to a certain degree mitigate these previously mentioned problems or challenges of classic reconstruction methods. Figure 1.2 displays different neural network architectures for the application of specific regression/reconstruction tasks.

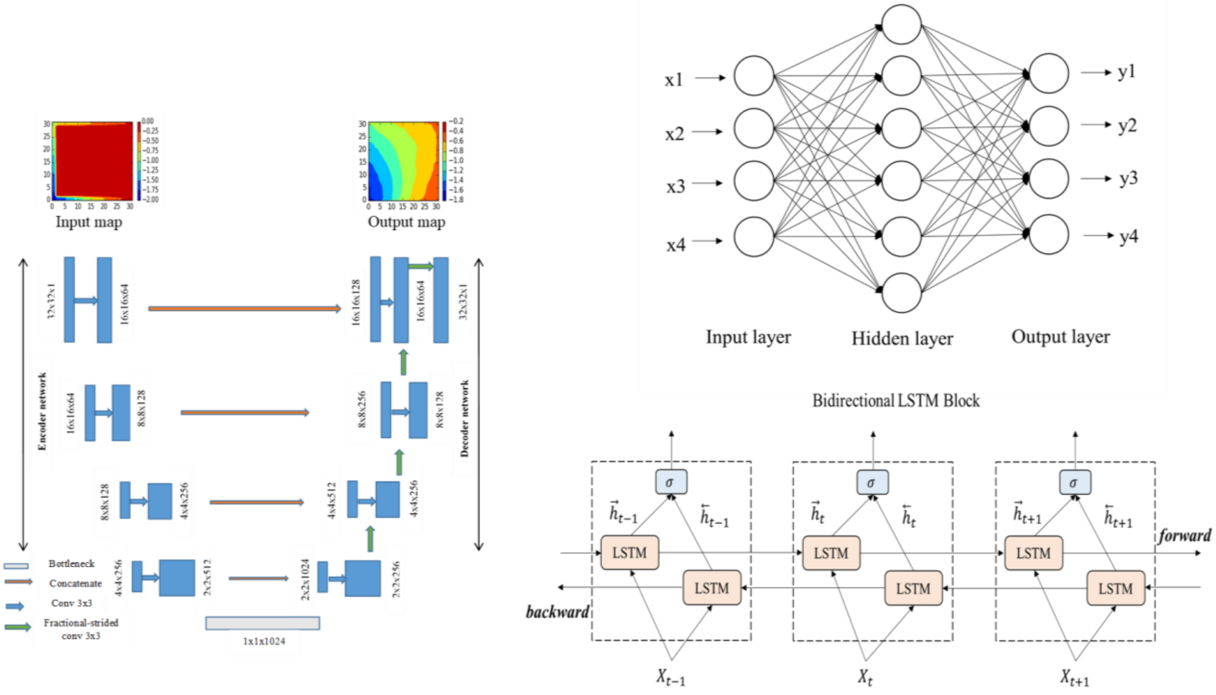

Figure 1.2: Machine learning methods mechanism

A challenge in setting up CFRs is that 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, and therefore offers little room to test the reconstructions with data that lie outside the range of the training data. As mentioned previously, 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 affect the quality of climate reconstructions (Christiansen and Ljungqvist, 2017), but this quality is difficult to test with independent data.

An alternative direction to estimate the quality of the reconstruction is to test reconstruction methods in the controlled/physically consistent conditions provided by climate simulations with state-of-the-art Earth System Models. These comprehensive climate 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; Mann, 2002;

Smerdon, 2012; Gómez-Navarro et al., 2017). The main idea is, therefore, to consider a climate simulation as a realization of a 'true' reality. We need only to construct pseudo-proxy records in this virtual reality (e.g. synthetic tree-ring records), and apply the reconstructions methods to those pseudo-proxy records. The outcome of this application can be then compared to the climate fields directly produced by the Earth System Model, thus allowing for an assessment of the reconstructions method.

PPEs are mainly motivated from the fact that real climate reconstructions are usually based on many different proxy networks, statistical methodologies and calibration period selections. Reconstruction uncertainties thus will occur due to the combination of these employed statistical methods, the spatial and temporal coverage of selected proxy networks, the stochastic noise between reconstructions and climatic proxy records, which makes it difficult to isolate the impact of one specific factor in reconstruction evaluations and comparisons (Christiansen and Ljungqvist, 2017). PPEs provide an experimental framework that can be systematically evaluated and altered, thus testing different statistical methods and their dependencies. In addition, PPEs lead to much longer, albeit synthetic, validation time intervals compared to instrumental records. Therefore, methodological assessments can extend to lower frequencies and longer time scales (Smerdon, 2012).

In this thesis, different traditional CFR methods are compared with three state-of-the-art machine leaning methods for surface temperature and sea surface high field reconstructions, respectively. The main question is that we want to test the reconstruction performance of these machine-learning methods and evaluate whether the nonlinear machine leaning method could achieve comparable or even better reconstruction skills compared to traditional CFR methods. The general workflow of climate field reconstructions in this thesis is shown in Figure 1.3 shows as below.

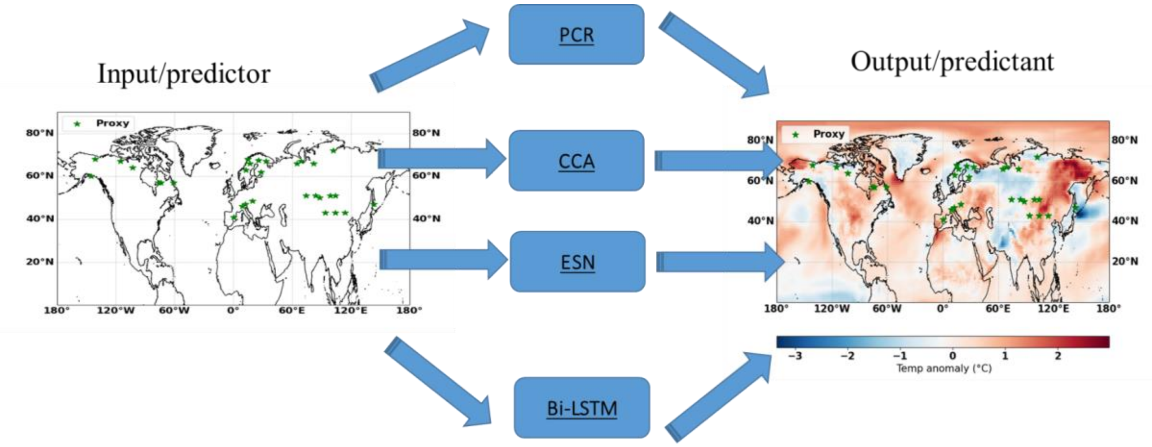

Figure 1.3: Workflow of climate field reconstructions using machine-learning methods

Specifically, following three main questions are explored in **3** separate manuscripts:

1). Can a recurrent neural network improve CFRs compared to traditional Multivariate linear methods?

2). Does reservoir computing improve CFRs skills compared to traditional Multivariate linear methods?

3). Does Convolutional neural network produce reasonable CFRs?

In **manuscript 1**, three different CFR methods are designed to reconstruct spatially resolved summer surface temperature field over the past millennium. These methods can be applied to real proxy records, but here they are tested in controlled conditions using *pseudoproxy experiments*. Two of these methods are traditional multivariate linear methods (PCR and 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 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 relatively more complex machine-learning method would provide better reconstructions for temperature fields was not confirmed.

Manuscript 2 expands the results obtained in manuscript 1. We employ and test a non-linear CFR method on the application of Hemisphere surface temperature field reconstructions, this approach belongs to the family of machine learning method, the Echo State Network (ESN). We compare the reconstruction performance of this ESN method with the Bi-LSTM that we have tested before in manuscript 1, and with PCR and CCA. One more import reason for employing ESN in manuscript 2 is that despite a number of successful applications of RNNs, it has been revealed that RNNs are difficult to train based on gradient descent. The whole model parameters of RNNs are gradually changed and optimized within training processes, of which the gradient information degenerates and might be ill defined, these deficits could lead to unguaranteed convergence. Besides, the model parameter updates could be computationally expensive during training processes, which also leads to unnecessary long training period. The ESN method was proposed for mitigating these previously mentioned shortcomings of RNNs, and it has been demonstrated that this ESN paradigm outperforms the classical RNNs in different regression and classification tasks. The experimental results in manuscript 2 demonstrate that ESN show a certain degree of superiority both in hemispheric surface temperature field and indices reconstructions compared to other three CFR methods. Thus, we confirm our working hypothesis that the relatively structural and training-simpler machine learning method - ESN could provide a better temperature field reconstruction.

In **manuscript 3**, we deviate from the reconstructions of past climate and instead aim at the reconstruction of the sea surface height field of the North Sea using a limited amount of data from tidal gauges (TGs). We apply a generative adversarial network (GANs), which belongs to the family of deep learning/machine learning method. In addition, we also compare another data assimilation method - Kalman filter approach with GANs about the reconstruction performance of sea surface high fields. Our objective in the manuscript 3 is to explore the performance of deep learning methodologies on reconstructing sea surface height fields in the specific North Sea using pseudo-observations and real observation data collected on coastal TGs stations. Individual reconstruction experiments using different combinations of training and target data during the training and validation process demonstrated similarities with data assimilation - Kalman filter method when errors in the data and model were not handled appropriately. The proposed method GANs demonstrated good reconstruction skills when analyzing both the full SSH field signal, as well as the low frequency SSH variability only. Thus, our hypothesis that deep learning method - GANs employed in the manuscript 3 were able to skillfully reconstruct SSH field in a specific regional sea and can achieve comparable reconstructions compared with Kalman filter data assimilation method was confirmed.

The thesis consists of five chapters: The first chapter is the introduction chapter; chapters 2 to 4 represent the three manuscripts. Finally, chapter 5 contains conclusions and provides some perspectives for potential future investigations.

Chapter 2: Evaluation of statistical climate reconstruction methods

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

Manuscript minor revisions in Climate of the Past (Zhang et al., 2022a)

Summary

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. Yet, a certain degree of reconstruction performance achieved by LSTM shows positive tests on using small simple with limited dataset.

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; 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 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 ca/pture 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 (see 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.

For the sake of completeness, we briefly mention here relevance for our study of the reconstruction methods that combine the assimilation of 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. These special characteristics make the comparison with purely data-driven methods more difficult and probably unfair, since data assimilations uses a much larger amount of information from climate simulations. In addition, this use of information from climate simulations compromises one of the main objectives of climate reconstructions, namely the validation of climate models in climate regimes outside the variations of the observational period. Therefore, the testing of purely data-driven reconstruction methods retains its relevance, despite the availability of more sophisticated data assimilation methods.

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 AMV is sometimes defined as the decadal variability of the North Atlantic sea-surface temperature, whereas the term Atlantic Multidecadal oscillation (AMO) is reserved for the decadal internal variations (excluding the externally forced variability). Here we focus on the total variability of the North Atlantic SST and define 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.

2 Data and Methods

2.1 Data

2.1.1 Proxy data locations

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 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 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 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

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 CESM.

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 2A.

The second climate model is the Community Earth System Model CESM Paleoclimate model 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 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 CESM 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 CESM 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 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

$$PNV = NVAR/(1 + NVAR) \tag{2.1}$$

Red noise for a specific PNV could be defined by:

$$Red_t = \alpha_1 Red_{t-1} + White_t \tag{2.2}$$

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 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 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, 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 $\boldsymbol{x}_{\mathrm{t}}$, at time step *t* can be decomposed as follows:

$$\boldsymbol{x}_{m,t} = \sum_{n=1}^{k} PC_{n,t}\, \boldsymbol{EOF}_{m,n} \tag{2.3}$$

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:

$$PC_{n,t} = \sum_{m=1}^{j} \omega_{n,m} Proxy_{m,t} + \varepsilon \tag{2.4}$$

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:

$$\widehat{PC}_{n,t} = \sum_{m=1}^{j} \omega_{n,m} Proxy_{m,t} \tag{2.5}$$

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 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*:

$$\boldsymbol{x}_{m,t} = \sum_{n=1}^{l} CC_{n,t}^{field}\,\boldsymbol{CCP}_{m,n}^{field} \tag{2.6}$$

$$\boldsymbol{Proxy}_{m,t} = \sum_{n=1}^{l} CC_{n,t}^{proxy}\,\boldsymbol{CCP}_{m,n}^{proxy} \tag{2.7}$$

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 is possibly 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 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).

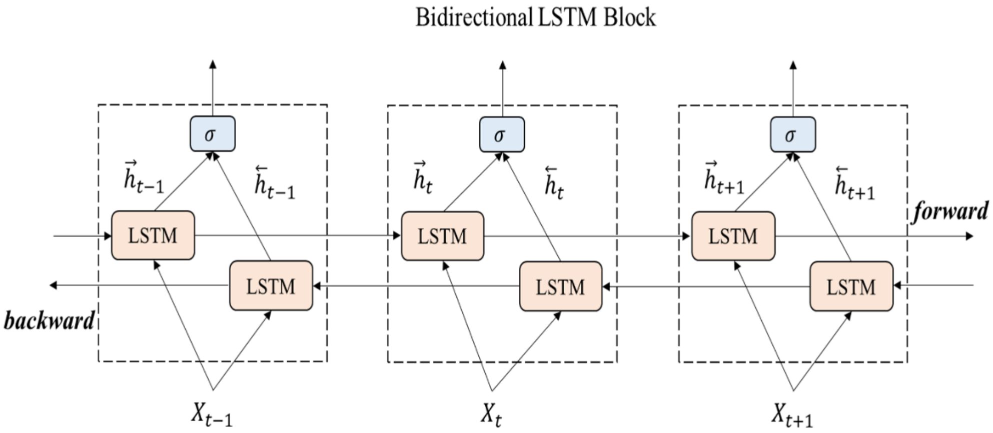

Figure 2.1: the bidirectional structure of the Bi-LSTM network.

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 2.1 illustrates the bidirectional structure of the Bi-LSTM network. Given a set of predictor-predictand variables ($\boldsymbol{X_t}$, $\boldsymbol{Y}_{\mathrm{t}}$), our goal is to train a nonlinear function:

$$\widetilde{\boldsymbol{Y_t}} = F(\boldsymbol{X}) \tag{2.8}$$

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

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + B_f\right) \tag{2.9}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + B_i) \tag{2.10}$$

$$A_t = tanh(W_A[h_{t-1}, x_t] + B_A) \tag{2.11}$$

$$C_t = f_t C_{t-1} + i_t A_t \tag{2.12}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + B_o) \tag{2.13}$$

$$h_t = o_t tanh(C_t) \tag{2.14}$$

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 2.9) , the input gate i_t (equation 2.10) and the state of memory cell A_t (equation 2.11). 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 i_t (equation 2.12). The state of the output gate o_t is calculated from the previous hidden state and the current input variables (equation 2.13). This output is used to compute the updated hidden state h_t using the state of the cell output C_t (equation 2.14) (Huang et al., 2020; Chattopadhyay et al., 2020).

In the present application to climate reconstructions, we have a set of input pseudoproxy data $\boldsymbol{X}_t^n$= [$x_{\mathrm{t-i}}$,…, $x_{\mathrm{t-1}}$]and an output target temperature time series $\boldsymbol{Y}_t^m$ = [$y_{\mathrm{t-i}}$,…, $y_{\mathrm{t-1}}$]. The forward LSTM hidden state sequence $\overrightarrow{\boldsymbol{h}_t}$ (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 $\overleftarrow{\boldsymbol{h}_t}$ 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):

$$\widetilde{\boldsymbol{Y}}_t = concat\left(\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}\right) \tag{2.15}$$

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 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:

$$L_{\delta}\big(\boldsymbol{Y}, f(\boldsymbol{X})\big) = \begin{cases} \quad \frac{1}{2}\big(\boldsymbol{Y} - f(\boldsymbol{X})\big)^2 \\ \delta|\boldsymbol{Y} - f(\boldsymbol{X})| - \frac{1}{2}\delta^2 \end{cases} \tag{2.16}$$

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 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 2B). 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. 2.2 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 CESM 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. 2.2 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.

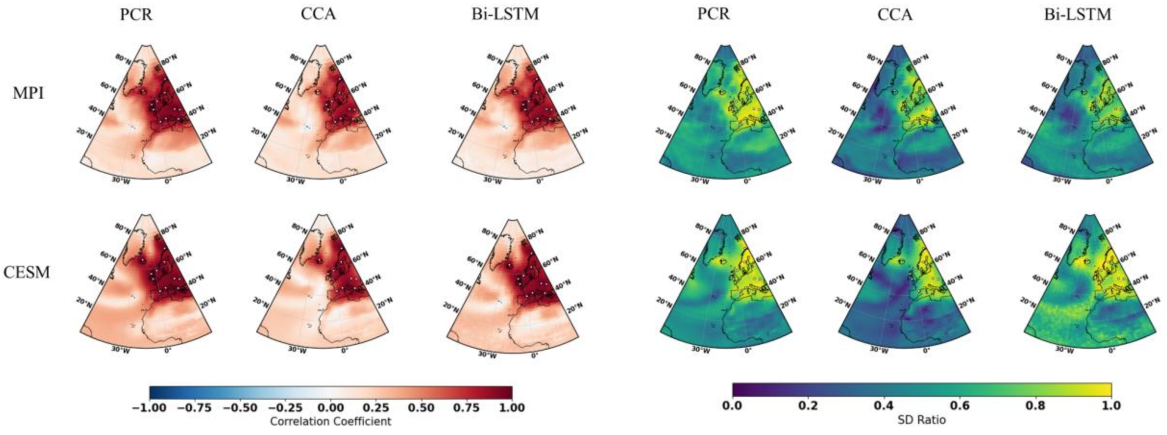

Figure 2.2: NAE Reconstruction results of CFR methods (including PCR, CCA, Bi-LTSM : Bidirectional long short term memory neural networks) using MPI and CESM numerical simulation as target temperature field, all the CFR methods employ the same proxy network with full 11 ideal pseudoproxies which span

the same reconstruction period from 850-1899 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.

The picture that emerges from the SD ratio is also very similar for the three methods (Fig. 2.2). 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 2C 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.

Gaussian white and red noise is constructed and added to the ideal temperature signal of the 11 pseudoproxies subsampled from the MPI and CESM models. The corresponding spatial *cc* and SD ratio patterns are displayed in Fig. 2.3 and 2.4 correspondingly. Compared to reconstructions with ideal pseudo proxies (Fig. 2.2), 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. 2.3 and 2.4 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 CESM based PPEs (referring to the spatial SD ratio in Fig. 2.4).

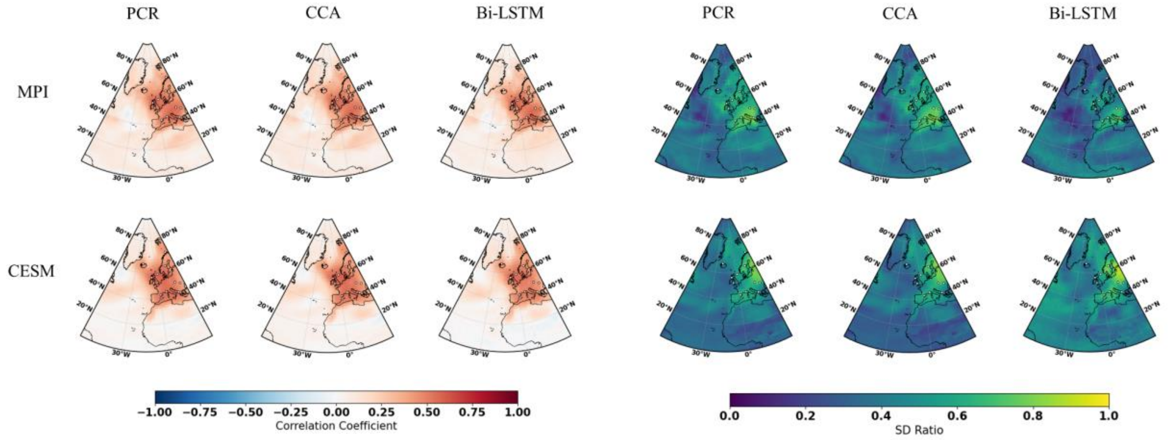

Figure 2.3: the same as Figure 2.2, but for employing the full 11 pseudoproxies network with white noise contamination.

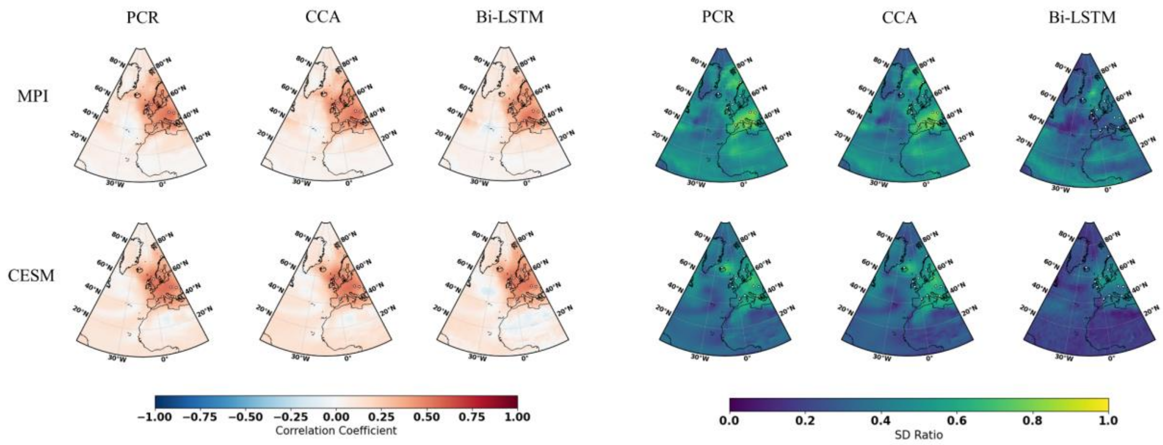

Figure 2.4: the same as Figure 2.2, but for employing the full 11 pseudoproxies network with red noise contamination.

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. 2.5-2.7.

Table 2.1 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 CESM). The numbers in parenthesis indicate the skill statistics of white noise and red noise (italics) contaminated PPEs.

Method	SD Ratio		*cc*	
	MPI	CESM	MPI	CESM
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. 2.5. 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.

All derived CFRs suffer from underestimation of interannual variance, as shown in Fig. 2.5 and in Table 2.1, except that the PCR method presents a clearly interannual variance overestimation referring to the specific spatial SD ratio map in Fig. 2.5. 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 CESM and MPI model, and a relatively higher SD ratio appeared in tropical regions within Bi-LSTM based PPEs shown in Fig. 2.5.

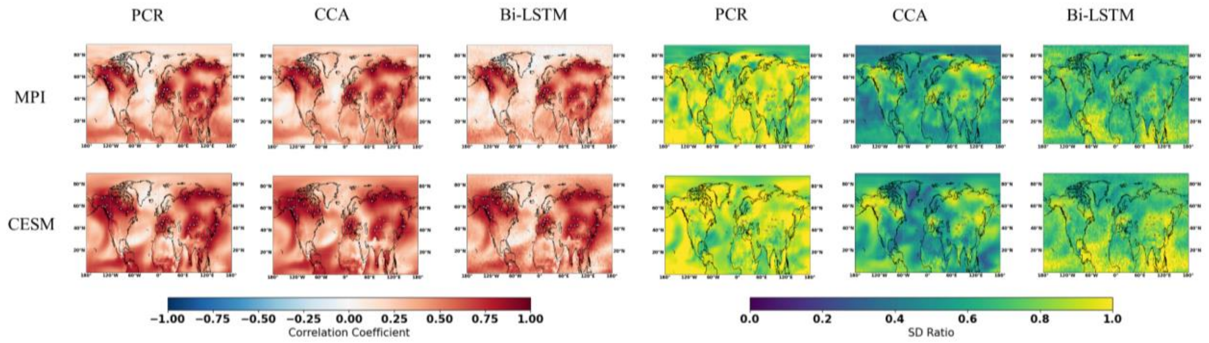

Figure 2.5: NH Reconstruction results of CFR methods (including PCR, CCA, Bi-LSTM: Bidirectional long short term memory neural networks) using MPI and CESM numerical simulation as target temperature field, all the CFR methods employ the same proxy network with full 48 ideal pseudoproxies which span the same reconstruction period from 850-1899AD. 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.

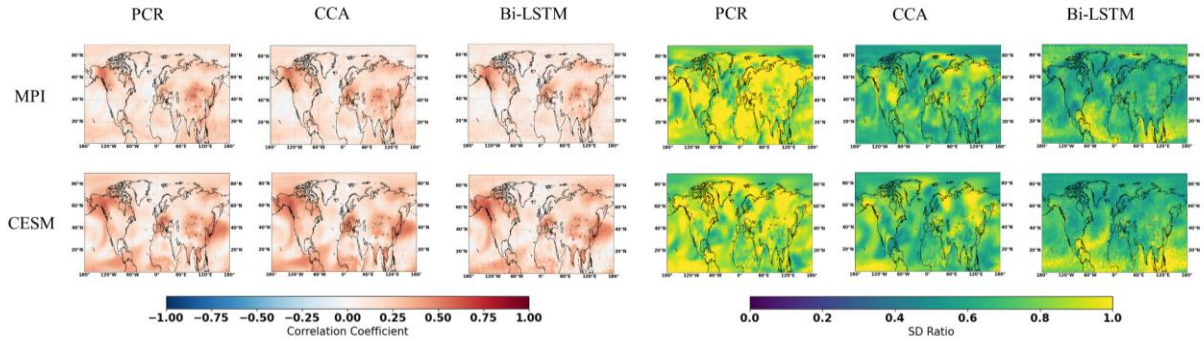

Figure 2.6: the same as Figure 2.5, but for employing the full 48 pseudoproxies network with white noise contamination.

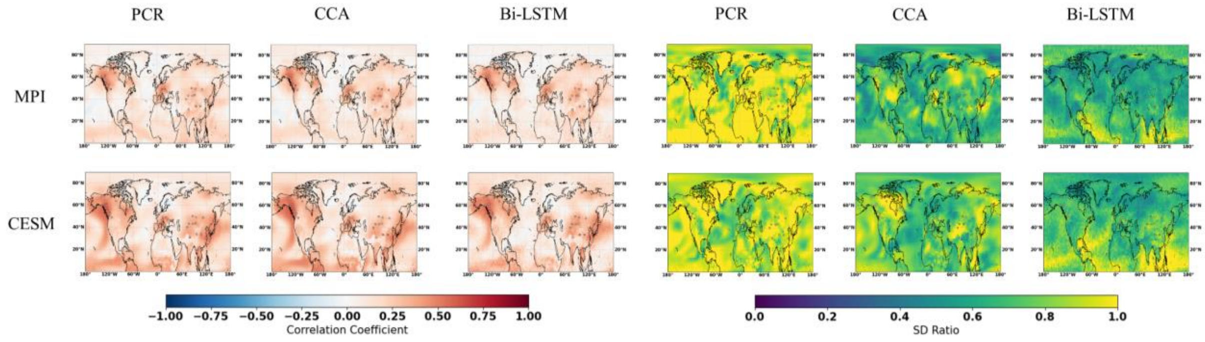

Figure 2.7: the same as Figure 2.5, but for employing the full 48 pseudoproxies network with red noise contamination.

The CCA methodology seems to suffer more strongly from variance losses (see Table 2.1) 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 2.1, 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 refers to 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 based on the results from the reconstruction interval, and their proportion of explained variance of the reconstructed field, derived from the three reconstruction methods using the CESM pseudo-proxies. The EOF patterns represented in Figure 2.8 confirm the suggestions that the temperature reconstructed by the PCR and CCA methods (two lower rows in Figure 2.8) 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.

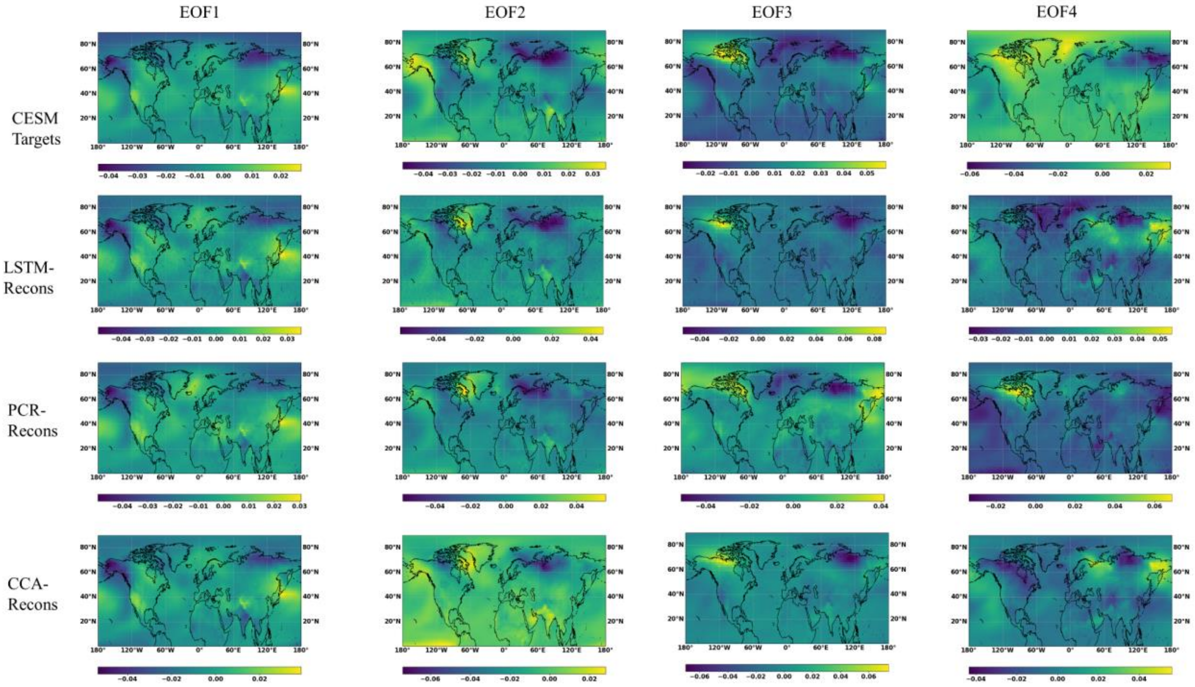

Figure 2.8: First four EOF patterns of the temperature field derived from for CESM target and derived from the temperature field reconstructed by the three methods-based ideal PPEs reconstructions

The corresponding spectrum of explained variance is displayed in Figure 2.9. Here, the percentage of explained variance of each model is calculated as the ratio of the eigenvalue to the total variance of the original field. This is definition is in principle similar to the definition adopted by Yun et al. (2021), but there is one important difference. Yun et al. (2021), according to their methodological description, calculate the portion of explained variance of each mode as the ratio between the eigenvalue and the total variance of the respective field (either original or reconstructed). This choice could, however, cause a statistical artifact. For instance, when using the PCR regression method, we could choose to reconstruct only the leading EOF pattern. This pattern alone will explain 100% of the reconstructed variance by definition, but this result would be obviously not informative. The choice of the total variance of the original field as reference thus leads to more results that are informative in general. The spectra for model simulation and three method-based ideal PPEs in this text are computed as the ratio between each of the first four reconstructed eigenvalues and the cumulative sum of all eigenvalues from the target variable.

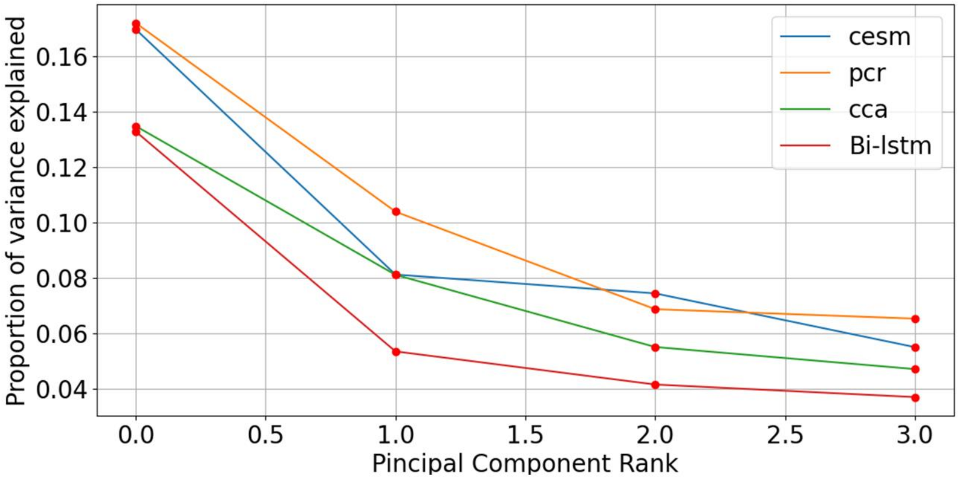

Figure 2.9: Eigenvalue spectra for CESM simulation and three method reconstructions: the spectra for CESM simulation and three method-based ideal PPEs are computed as the ratio between each of the first four reconstructed eigenvalues and the cumulative sum of all eigenvalues from the target CESM model

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.

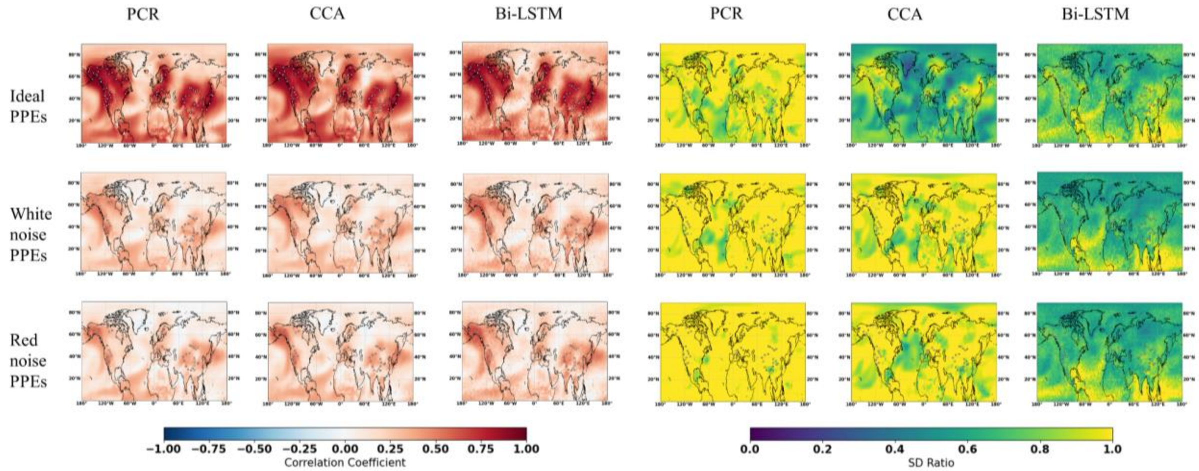

Figure 2.10: Summary of the pseudo-reconstructions derived from the CESM model-based pseudo-proxies using the original PAGES proxy network. The panels display the maps of the temporal correlation

coefficients at the grid-cell level (*cc*) and the ratio of standard deviations (SD ratio) between reconstructed and target temperature field

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. 2.10, 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. 2.10 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. 2.10. 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. 2.11. 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.

Table 2.2. *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	CESM	MPI	CESM	MPI	CESM
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 2.3. The same as Table 2.2, but for decadal AMV index

Method	*cc*		SD		RMSE	
	MPI	CESM	MPI	CESM	MPI	CESM
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	0.829	0.623	0.600	0.108	0.084

	(0.573/*0.344*)	(0.435/*0.450*)	(0.539/*0.576*)	(0.440/*0.536*)	(0.154/*0.182*)	(0.126/*0.125*)

The temporal evolution of the original AMV indices (Fig. 2.12) 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. 2.12 and in Table 2.3.

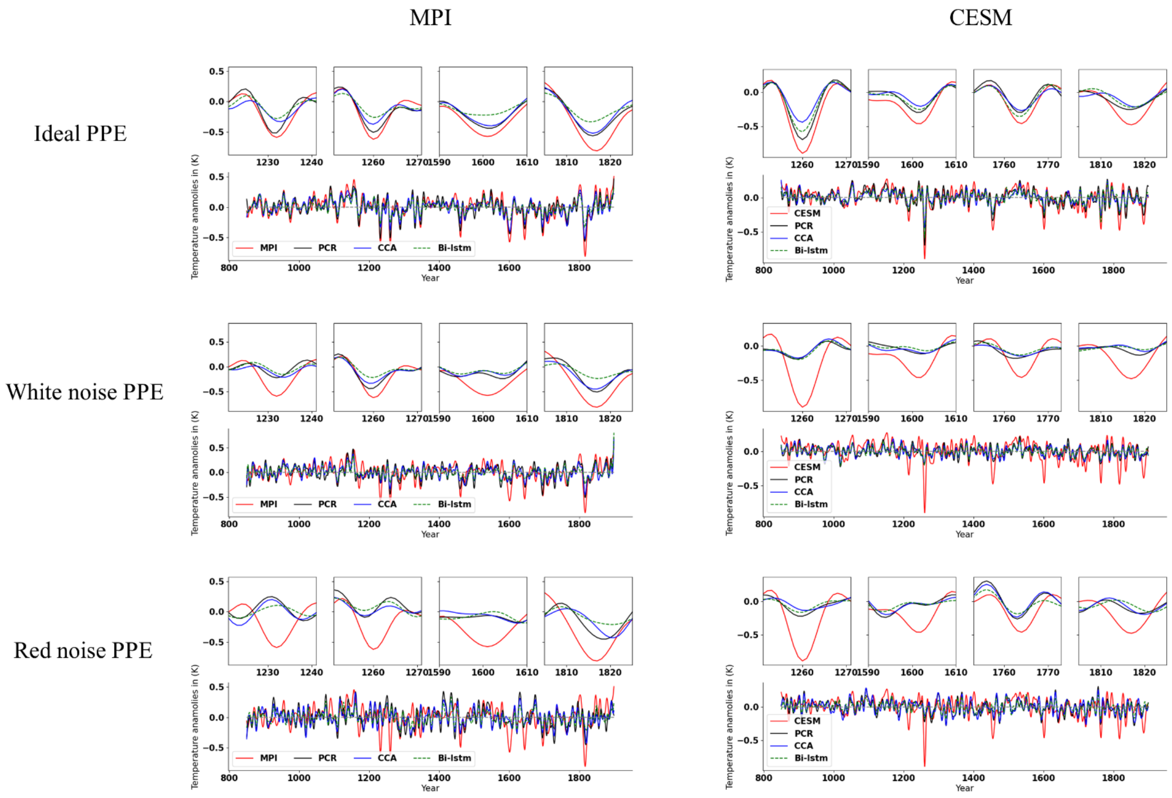

Figure 2.11: mean time series evolution of the validated reconstructions for NH summer temperature anomaly using full 48 pseudoproxies based on PCR, CCA, Bi-LSTM CFR methods. All time series have been smoothed using a butter worth low-pass filter to remove temporal fluctuations less than 10 years. MPI and CESM represent MPI/CESM 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.

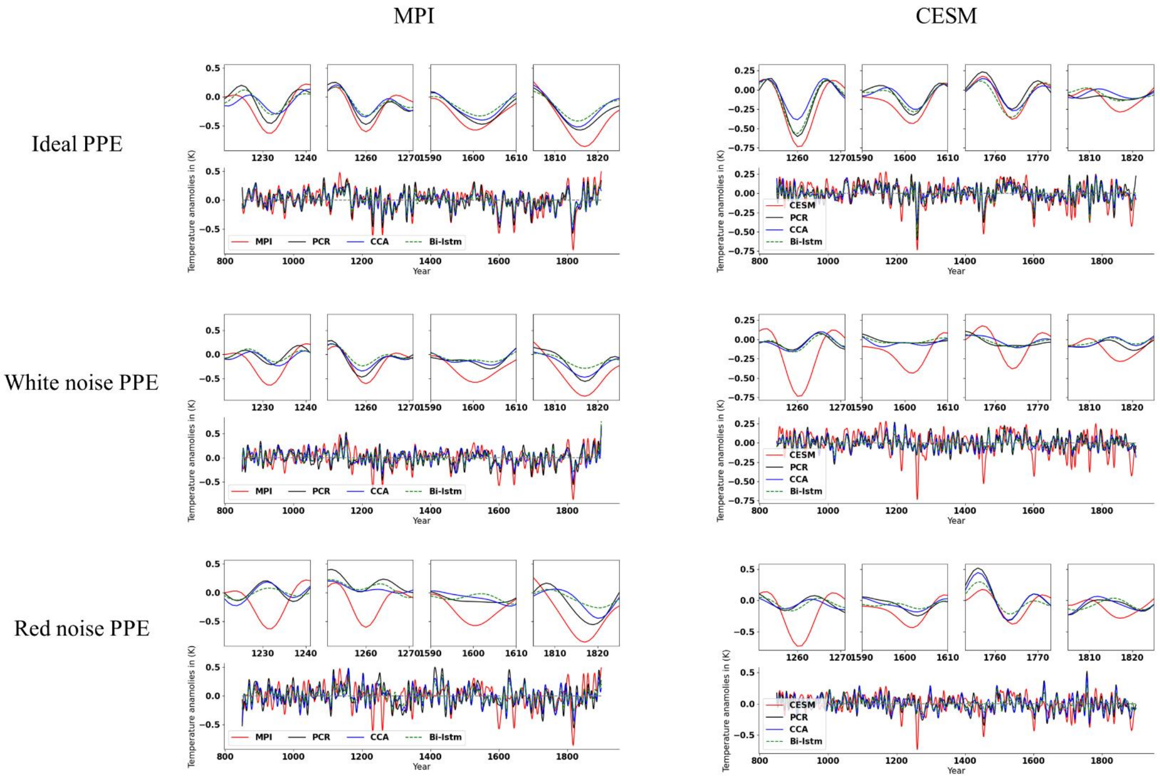

Figure 2.12: The same as Figure 2.11, but for Atlantic Multidecadal Variability (AMV) index.

The NH and AMV indices derived from more realistic noise contaminated CFRs are shown in Fig. 2.11 and Fig. 2.12 correspondingly. The larger noise contamination results in substantial skill deterioration (*cc*, SD and RMSD displayed within brackets in Table 2.2 and 2.3). 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. 2.13 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. 2.13, all three methods generally underestimate the power density, whereas this underestimation is more significant for the noise-contaminated derived PPE.

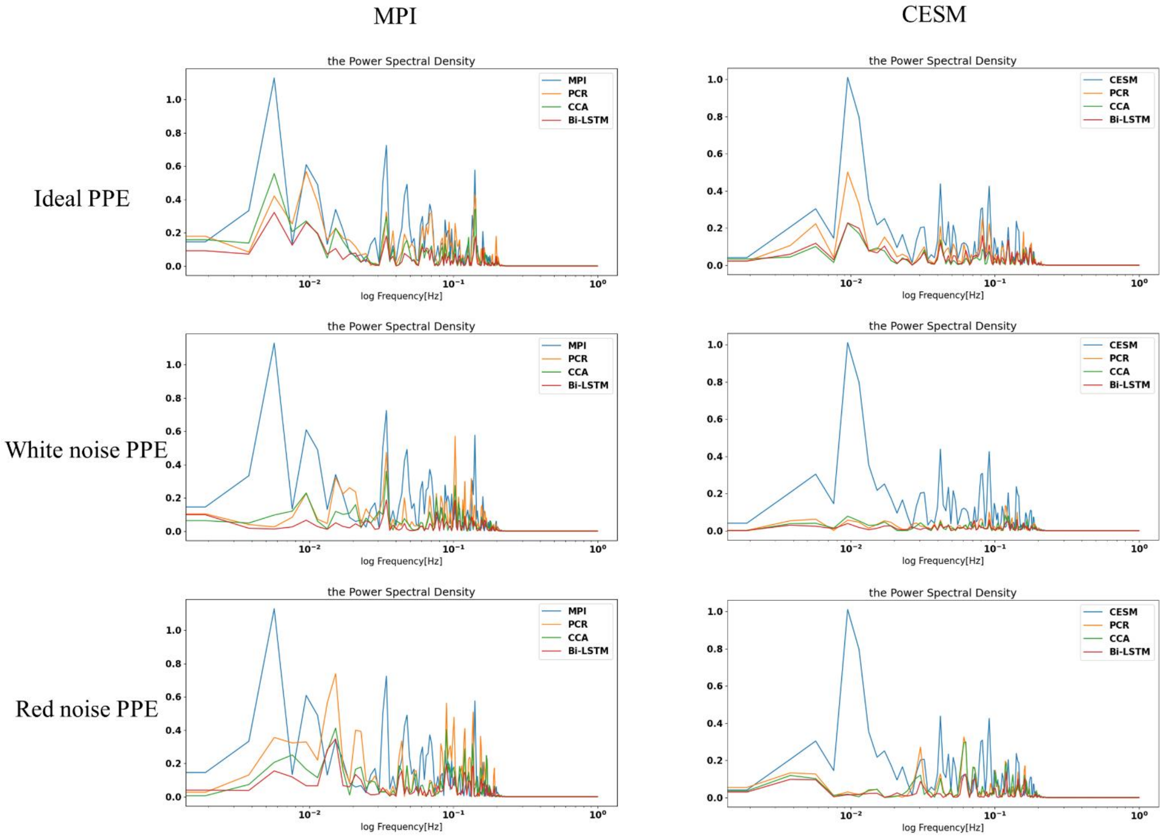

Figure 2.13: North Hemisphere indices power spectral density.

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.

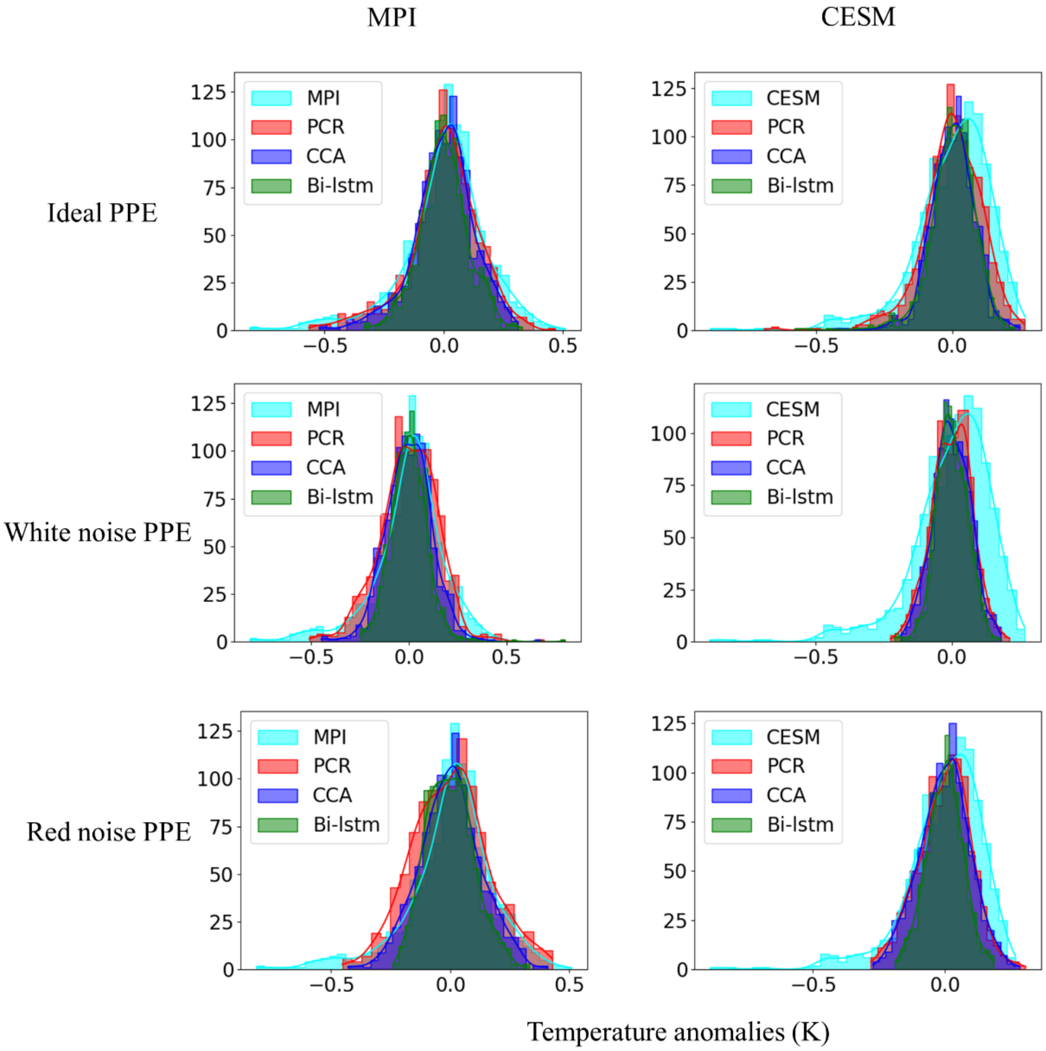

Figure 2.14: Histogram for decadal filtered NH mean index. The x axis denotes temperature anomaly values, and y axis is the number of data in each bin. Totally 30 bins are selected to plot each of the histogram.

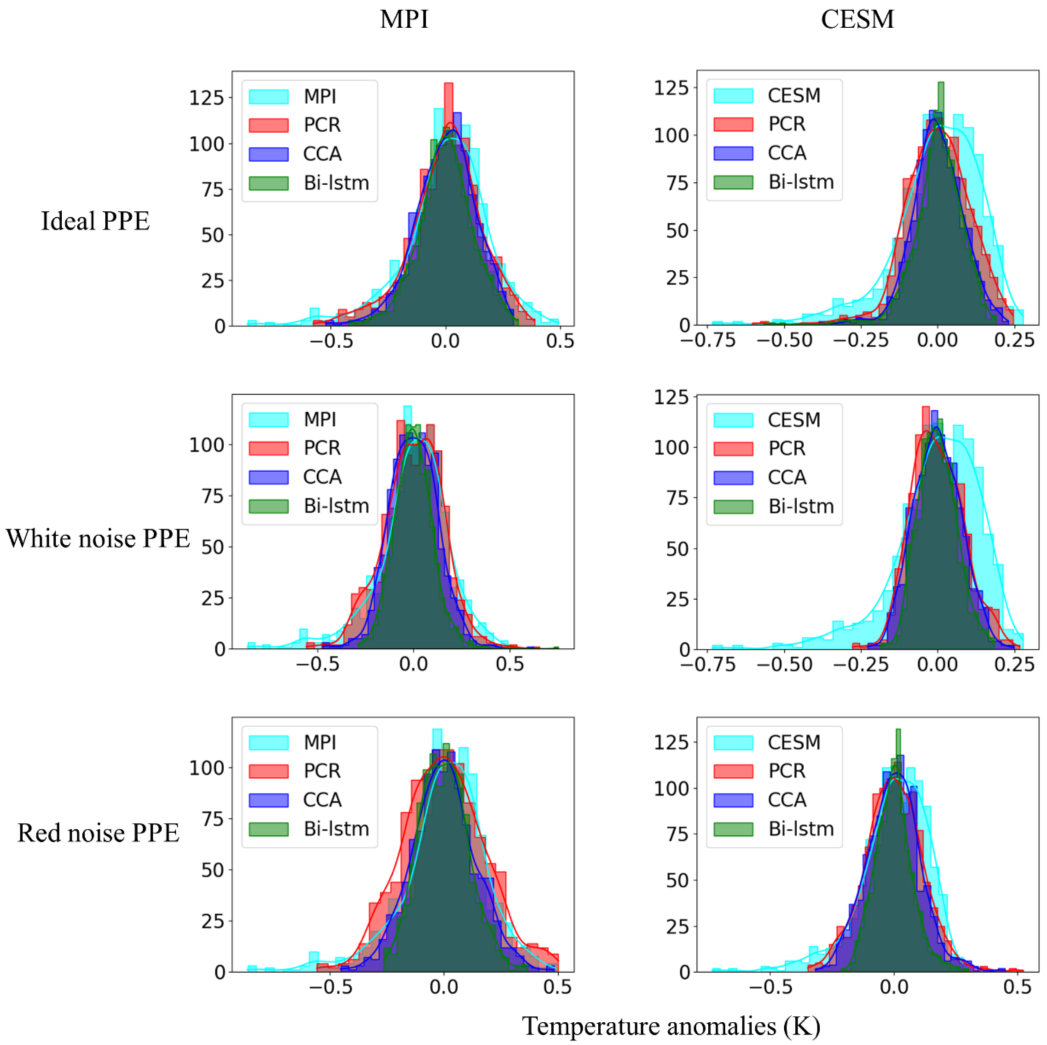

Figure 2.15: The same as Figure 2.14, but for decadal filtered AMV index.

Fig. 2.14 and 2.15 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.

Table 2.4. 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	CESM	MPI	CESM
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 2.5. The same as Table 2.4, but for AMV index.

Method	KS statistic		p-value	
	MPI	CESM	MPI	CESM
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 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 2.4-2.5). A smaller value of the KS statistic indicates a stronger overall similarity between the two probability distributions. The smallest KS statistic is achieved by the PCR method (see Table 2.4-2.5), confirming the impression that the PCR outperforms the other two methods for indices reconstructions in both the ideal and noise contaminated PPEs.

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. 2.14 and 2.15). 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.

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 2B.

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 l*eaky 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. 2.2-2.4).

For spatially resolved NH field, the PCR overestimates the variabilities both in ideal and noisy PPEs (see spatial SD ratio maps in Fig. 2.5-2.7 and mean statistics skills Table 2.1), and the CCA method shows relatively lower overestimatied varaince in noisy PPEs, the Bi-LSTM presents relatively reasonable reconstructions without clearly overestimation both in ideal and noisy PPEs (see Fig. 2.5-2.7 and Table 2.1). 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. 2.5-2.7 and mean skill statistics in Table 2.1). Whereas, both PCR and CCA exhibit overestimated variability reconstructions within noisy PPEs, the Bi-LSTM presents relatively robust reconstructions especially without variance overestimations in noisy PPEs (see Fig. 2.5-2.7 and mean skill statistics in Table 2.1).

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 2.11-2.12). 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. 2.13 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 machine learning (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 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 in the amplitude of reconstructed variability (see SD ratio patters in Fig. 2.9-2.10 and mean skills in Table 2.1). 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 2.1).

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. 2.5-2.7). Interestingly, the tropical regions do show some reconstruction skill especially in the derived reconstructions based on Bi-LSTM (spatial SD ratio patterns in Fig. 2.5-2.7), 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 Conclusion

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:

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.

Chapter 3: Evaluation of the Bilinear Long-Short-Term-Memory and Echo State Network Methods

Evaluation of the Bilinear Long-Short-Term-Memory and Echo State Network machine learning methods to reconstruct the Northern Hemisphere summer temperature

Manuscript (to be) submitted to Climate of the Past (Zhang et al., 2022b)

Summary

In order to compare the performance between linear regression and machine learning methods for climate field reconstruction, we employed four different methods to reconstruct Northern Hemisphere summer temperature field over last millennium. Two employed method including Principal Component Regression (PCR) and Canonical Correlation Analysis (CCA) are classical linear regression methodologies, while the rest two methods including Bidirectional Long-short-term-memory neural network (Bi-LSTM) and Echo State Network (ESN) belong to the family of machine learning methodology. In PCR and CCA, a general assumption is usually applied before driving reconstructions, which is several dominant climate patterns are assumed constant with time, whereas no assumptions need to be accounted in the two machine-learning methods. Furthermore, Machine-learning methods, especially the neural network methods, can provide superiorities in nonlinear function mapping tasks, it could incorporate some underlying nonlinearities in temperature variability. Based on ideal and noise-contaminated Pseudoproxy reconstruction experiments (PPEs) in the context, the ESN shows superiorities in preservation more temperature variance compared to other three CFR methods, and presents outperformance in capturing some extreme cooling events in the reconstructed temperature indices. Generally, the ESN method employed herein PPEs show a certain degree of superiorities both in hemispheric temperature field and indices reconstructions.

1 Introduction

Climate field reconstructions play an important role in understanding climate variability and evolution. For instance, paleo temperature field reconstructions could provide us with an expanded and clear data basis for better understanding of the past temperature evolution. Learning from the past could provide us with a broader perspective to better predict the future (Mann and Jones, 2003; Jones and Mann, 2004; Jones et al., 2009; Schmidt, 2010; Evans et al., 2014; Smerdon and Pollack, 2016; Christiansen and Ljungqvist, 2017). Nevertheless, observational or instrumental climatic records are only available back to the 19 century, so that there is a need to produce climate reconstructions from indirect proxy records (such as ice core, coral, sediments. Jones and Mann, 2004).

Past climate reconstructions are usually derived by employing statistical methods that translate the proxy information in units of physical environmental variables. However, these statistical methods can add errors, additional statistical uncertainty and bias to the reconstructions. In addition, the variations of proxy records are also caused by non-climate factors that blur the true climate signal in them. Both the non-climatic variability signals from proxy records and uncertainties from reconstruction methods can introduce statistical bias and uncertainties into the final climate reconstructions.

In order to estimate these potential deficiencies of climate reconstructions, pseudo-proxy experiments are proposed (PPEs, Smerdon, 2012, Gómez-Navarro et al, 2017). Using PPEs, climate reconstructions methods can be tested in a controlled situation provided by numerical simulations with Earth System Models. These state-of-the-art climate models can provide virtual climate trajectories with a physical consistence perspective for climate reconstructions. The impact of selected proxy network, the climate signal originated from proxy records and the reconstruction performance of statistical methods can then be evaluated in the virtual reality of climate simulations.

Amongst many statistical reconstruction methods, several common deficiencies have been identified, for instance a general tendency to ‘converge to the mean’; this can lead to an underestimation for reconstructing the total climate variability. Specifically, when the availability of proxy information - the climate signal included in proxy records is limited, this underestimation will be more significant. Besides, the limited spatial coverage of proxy network might result in biased reconstructions (Evans et al., 2014; Wang et al., 2014; Po-Chedley et al., 2020; Amrhein et al., 2020). Consequently, significant scope remains for evaluating and developing CFR techniques furtherly to mitigate those common deficiencies (Christiansen and Ljungqvist, 2017).

In present study, we employ and test a non-linear CFR method, which belongs to the family of machine learning, an Echo State Network (ESN). We compare the reconstruction performance of this ESN method with a Bidirectional Long Short-Term Neural Network (Bi-LSTM) that we have tested in a previous publication, and also with two traditional linear regression methods: Principal Component Regression and Canonical Correlation Analysis. These two linear regression methods employed in CFRs are usually established with an assumption that the relationship between target climate fields and local variables captured by indirect proxy records is linear and temporally stable. For instance, the spatial patterns of surface temperature field is usually considered as stationary. Nevertheless, climate system is chaotic and dynamic, it means that many underlying nonlinearities/uncertainties existed in climate evolution systems, and many links and relationships between climate fields can be highly non-linear (Dueben and Bauer, 2018; Schneider et al., 2018; Huntingford et al., 2019; Nadiga, 2020). The machine learning methods which have been demonstrated with highly non-linear mapping capability, for example, Artificial Neural Network, could help us to better capture the underlying relationships between targeted climate fields and proxy records (Schneider et al., 2018; Rasp and Lerch, 2018; Rolnick et al., 2019; Chattopadhyay et al., 2020; Huang et al., 2020; Nadiga, 2020; Lindgren et al., 2021). In addition, machine learning methods do not necessarily require previous preprocessing methods, such as Principal Component Analysis (PCA), to firstly extract several dominant climate patterns. In principle, the inherent variability of original climate field dataset is dynamically captured in Machine-learning methods with the optimized hyper-parameters during training procedure (Goodfellow et al., 2016).

From the available linear and nonlinear methods that we could have employed for CFRs experiments, recurrent neural networks (RNNs) could be a potentially candidate with the objective of better reconstructing the true climate variability, because they have been demonstrated to be able to learn the serial correlations (Bengio et al. 1994). Nevertheless, RNNs usually only learn short-term serial correlation (Bengio et al. 1994). The Bi-LSTM network also belongs to the category of RNN, and it has been demonstrated that the Bi-LSTM is able to learn and remember long-term temporal dependencies from sequential dataset (Hochreiter and Schmidhuber, 1997).

The Echo State Network (Maass et al., 2002; Jaeger, 2001) introduces a novel paradigm into traditional RNNs' training process, of which an RNN (usually refer to as reservoir) is randomly generated in the hidden-reservoir layer and only the output part (so called readout) of this ESN method is trained. In general, this new paradigm is known as reservoir computing (Pathak et al., 2018; Chattopadhyay et al., 2020; Arcomano et al., 2020), which greatly outperforms the classical RNNs in different regression and classification tasks, and promotes the practical applications of RNNs (Chattopadhyay et al., 2020; Huang et al., 2020; Nadiga, 2020). One more import reason for employing ESN in our study is that despite a number of successful applications of RNNs, it has been revealed that RNNs are difficult to train based on gradient descent (Bengio, et al. 1994; Pascanu et al., 2013). All model parameters of RNNs are gradually changed and optimized within training processes, which may result in poor convergence properties. Besides, the model parameter updates can be computationally expensive during training processes, which also leads to unnecessary long training times. The ESN method was proposed for mitigating these previously mentioned shortcomings of RNNs. The ESN usually consists of three layers, the input layer for incorporating data information, the hidden layer - the reservoir layer for mapping input data information into a nonlinear high dimensional feature space, and the output layer - readout of which the desired output information is derived as a linear combination of reservoir' signals. Unlike RNNs, all the parameters in input and reservoir layer are fixed in the initialization process, and only the readout parameters needs to be trained and tuned by a least square linear regression, which can reduce the convergence time significantly, and mitigate potential uncertainties in training processes. Our underlying assumption, to be tested in this study, is that this property of ESN might lead to the reconstruction variance and to computationally more efficient reconstructions. Our working hypothesis is, that a more simplified type of RNN might better replicate past variability and perhaps even more so for extreme values compared to a more complicated Bi-LSTM method. We would like to test whether this property of the ESN is helpful for paleoclimate research, especially when calibration/training dataset is limited , as it is in real reconstruction (Najafabadi et al., 2015). In addition, it has been demonstrated that ESN can present some superiorities in spatiotemporal chaotic system predictions (Pathak et al., 2018; Chattopadhyay et al., 2020; Arcomano et al., 2020). To our knowledge, the ESN method is applied for the first time in the context of paleo CFRs.

In CFRs experiments for evaluating different statistical reconstruction methods, we focus on the Northern Hemisphere summer temperature anomalies, and we evaluate the reconstruction skills for both spatially resolved summer temperature anomalies and spatially averaged temperature anomalies derived from the reconstructed spatial fields. The reconstruction of mean temperature indices could provide a general evaluation of the skill to reconstruct extreme temperature phases (e.g. related to volcanic eruptions), which could also provide as benchmarks to assess the potential performance of CFR methods for those anomaly reconstructions.

In this present study, a new Northern Hemisphere Tree Ring Network Development database (NTREND), a relatively small, but highly temperature-sensitive tree-ring proxy network (Wilson et al., 2016; Anchukaitis et al., 2017), is mimicked for testing the performance of different CFR methods. An unscreened, large collection of proxy network generally results in a relatively lower reconstruction performance, while a small expert-selected extensive proxy records distributed in high-latitude region can provide a skillful extratropical temperature field reconstruction but might not be expected to provide additional climatic information for other regions (Franke et al., 2020). Nevertheless, combing and merging all available proxy records and leaving the weighting to a statistical reconstruction method might generally not provide optimal results (Franke et al., 2020). Thus, in our study, we mirror the NTREND tree ring records which consists of the sites of 54 records (tree-ring-width, maximum wood density or, blended) selected and filtered by dendrochronological experts to be the best temperature-sensitive proxies (Wilson et al., 2016). That is to say, they are our highest quality, but low quantity tree ring collection dataset with the least spatial coverage. Considering the networks of real proxies used so far, St. George and Esper (2019) reviewed contemporary studies on previous NH temperature reconstructions based on tree rings (Mann et al., 1998, 2008, 2007, 2009a, 2009b; Emile-Geay et al., 2017). They made a general conclusion that the present generation of tree-ring proxy based reconstructions exhibit high correlations with seasonal hemispheric summer temperatures and display relatively better skills in tracking inter-annually climatic variabilities and decadal fluctuations than former proxy networks, as also found by Wilson et al., (2016) and Anchukaitis et al., (2017).

Climate model simulations can provide paleoclimate reconstructions as test bed for evaluating both method and model dependencies in real and pseudoproxy CFR experiments (Smerdon et al., 2011, 2012, 2015; Amrhein et al., 2020; Parsons et al., 2021). Thus, we use three comprehensive Earth System Models (ESMs) which can serve as 'numerical laboratories' to evaluate the performance of four CFR methods employed in our study. The Max-Planck-Institute climate model MPI-ESM-P, the Community Climate System Model CCSM4, and the Community Earth System Model CESM1-CAM5, totally three model simulations are employed as surrogate climatic database for setting up PPEs in this context.

2 Data and Method

2.1 Data

2.1.1 Proxy Network

We construct the pseudoproxies using the simulated grid-point summer mean temperature time series originated from three climate model outputs over the past millennium. In this study, totally 30 tree ring based proxy locations across extratropical Northern Hemisphere region are filtered and selected from the sites of NTREND network. The original NTREND network was trimmed down by rejecting and removing sites in which the tree ring records did not show strong and clear climatic signals. Specifically, 30 dendrochronology locations are filtered and selected according to Figure 4 of Anchukaitis et al., (2017) which illustrates the field correlation coefficient between dendroclimatological proxy records and summer mean temperature field. The field correlations between each proxy and the mean temperature field at most of the retained grid-cell locations is generally higher than 0.4. Figure 3.1 illustrates the spatial distribution of tree ring proxy we employed in our context.

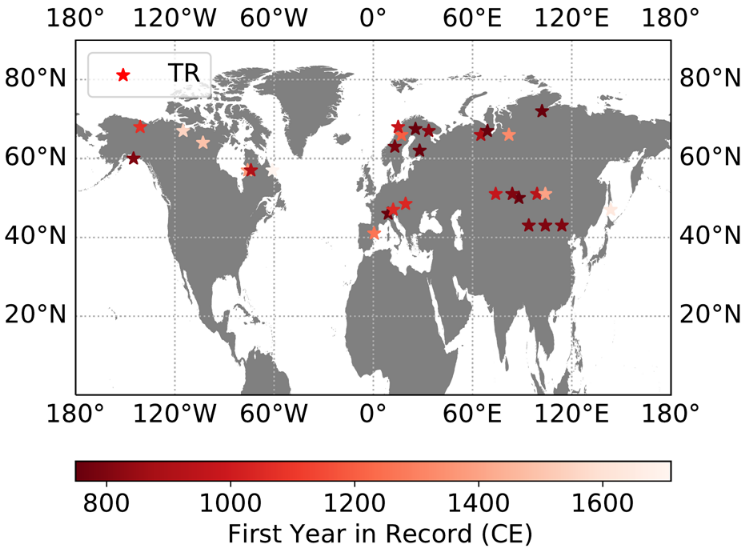

Figure 3.1: 30 real Tree ring Proxies remained after removing uncorrelated and weakly correlated (Figure 4 in Kevin J. Anchukaitis et al., 2017 indicated at some proxy sites where the local proxy response to MJJA is weak (e.g. locations in east Asia, Yakutia, and ring width-only chronologies from the North American treeline) proxies from 54(ntrend2015 from Kevin J. Anchukaitis et al., 2017

2.1.2 Climate models

In this study, we employ three climate model simulations as test bed for evaluating CFR methodologies. The first climate is the Max-Planck-Institute Earth System model MPI-ESM-P with a spatial horizontal resolution of about 1.9x1.9 degree. The simulation spans the period 100 BC to 2000 CE. It consists of the ocean model MPI-OM (Jungclaus et al., 2013), the spectral atmospheric model ECHAM6 (Stevens et al., 2013), the bio-geophysical model HAMOCC (Ilyina et al., 2013) and the land model JSBACH (Reick et al., 2013). Nevertheless, the present simulations based on MPI-ESM-P does not officially belong to the CMIP5 project. The driven forcings utilized in this simulation and additional technical setting ups are explained in the Appendix 2A. The second climate model employed in our PPEs is a Paleoclimate model: Community Earth System Model CESM1-CAM5 originated from National Centre for Atmospheric Research (NCAR) (Otto-Bliesner et al., 2016) with a spatial resolution of 2.5x1.9 degree (https://www.cesm.ucar.edu/projects/community-projects/LME/). In In these simulations, the CESM-CAM5-LME model covers the period from 850 CE to 2006 CE based on reconstructed climatic forcing for the land use conditions, transient evolution of aerosols, solar irradiance, orbital parameters, greenhouse gases and volcanic emissions, following the CMIP5 climate forcing reconstructions (Schmidt et al. 2011). CAM5 as the atmosphere model used in CESM1 (Hurrell et al., 2013) is a significant advancement of CAM4 model (Neale et al., 2013). We employ the last simulation member, 13th from the Last Millennium Ensemble. The third climate model employed in this study is the Community Climate System Model

CCSM4 model (Gent et al., 2011) of which the ocean model (POP2/Smith et al., 2010), the land (CLM4/ Lawrence et al., 2012), the sea ice model (CICE4/Hunke et al., 2008) and atmosphere (CAM4/Neale et al., 2013) are used as its components. One point needs to be emphasized is that the CESM1-CAM5 uses the same land, ocean and sea ice models as CCSM4 (Hurrell et al., 2013). CCSM4 model has a spatial resolution of 1.25x0.9 degree. In this study, the simulations labelled past1000 and r1i1p1 originated from CMIP5 pool are utilized; the past1000 simulations cover the period from 850 CE to 1849 CE and the historical simulation spans the period from 1850 CE to 2005 CE. The past1000 and historical simulations are concatenated together for our PPEs in this study. The driven forcing and boundary conditions follow the PMIP3 protocols (Schmidt et al. 2011).

2.2 Method

2.2.1 Construction of Pseudo-proxies network

The general methodology to prepare pseudoproxy climatic dataset from climate simulations is to subsample the model simulated time series at selected grid-cell which are co-located with the real tree ring proxy locations. In order to mimic real climate situation, we usually introduce random noise to contaminate the subsampled simulated time series, so that the correlations between the contaminated and original subsampled temperature time series can resemble the typical real proxy-temperature correlations. This correlation parameter can be modulated and adjusted in the pseudo record by adding an amount of random noise to simulated temperature. Ideal/perfect pseudoproxies only consist of model simulated temperature signals, Gaussian white noise is then introduced to contaminate the ideal Pseudoproxies temperature signals for setting up a more realistic PPEs. In this study, the percent noise by variance (PNV, Smerdon, 2012; Wang et al., 2014) is employed to define the noise level convention. The PNV indicates the ratio between added noise variance and total variance of resulting the pseudo-proxy time series.

$$PNV = NVAR/(1 + NVAR) \tag{3.1}$$

In general, real proxy records usually contain some temporal gaps and may span with different time period. For the sake of simplicity, we assume that the pseudoproxy data has no temporal gaps and cover the same whole period of simulations, and also an uniform distribution of noise is assumed throughout the whole pseudoproxy network.

We split the dataset employed herein for constructing PPEs into calibration period that covers 1900-1999, and validation period that covers 850-1999. All statistical metrics in the following sections are derived against the validation period from 850-1999.

2.2.2 Principal Component Regression

In order to reduce the dimensionality and reduce the risk of overfitting in processing high dimensional spatiotemporal dataset, we employ a statistical data preprocessing method, the Principal component analysis (Hotelling, 1957; Luterbacher et al., 2004; Pyrina et al., 2017), to decompose the original dataset into several new variables that can ideally represent a certain part of the total variability of the original climate field. In general, the dominant variability of original climate field can be described by several principal empirical orthogonal functions (EOFs). Combined with corresponded principal components (PCs), these EOFs time series could describe a major part of climate field variability. In this study, we retained several PCs that can capture 90% of the cumulative temporal variance of original climate field. The predictands in PCR are those PCs time series derived by PCA of the original climate field dataset. Based on the derived PCs and the empirical orthogonal functions-EOFs pairs from calibration dataset, we establish a desired linear regression relationship between proxy dataset (the predictors) and the PCs (the predictands), and then the desired PCs in validation period can be regressed by employing the estimated regression coefficients. The fully desired climate field is then reconstructed by using the linear combination of their corresponding EOFs and the regressed PCs.

We want to train a linear function to get appropriate equation parameters so as to map the relationship between predictors and predictants. Given a set of input climate field $\boldsymbol{x}_{\mathrm{t}}$, the *t* represents each time step, and at time step *t* input $\boldsymbol{x}_{\mathrm{t}}$ can be decomposed:

$$\boldsymbol{x}_{m,t} = \sum_{n=1}^{k} PC_{n,t}\, \boldsymbol{EOF}_{m,n} \tag{3.2}$$

where *t* indicates the time index, *m* represents the grid point index of climate field, and *k* is the total retained numbers of PCs.

The relationship between input proxies and desired climate field is built up by the following equation:

$$PC_{n,t} = \sum_{m=1}^{j} \omega_{n,m} Proxy_{m,t} + \varepsilon \tag{3.3}$$

where ω is the coefficient of the linear function, j represents the total numbers of proxies and ε is a residual term. In general, this residual term can be an unobserved random variable that will add noise to the linear function between targeted climate information (the proxy or pseudoproxy) and the dependent variable (PCs), in addition, it will involve all effects that are not related to the dependent variable on the desired regressors (Christiansen, 2011). Ordinary Least Square is employed to estimate the parameters ω.

In our PCR method, it is assumed that the temperature-sensitive proxies are linearly related with the derived PCs. The desired PCs in the validation will be reconstructed assuming that the linear coefficients calculated in Eq. (3) are constant in time:

$$\widehat{PC}_{n,t} = \sum_{m=1}^{j} \omega_{n,m} Proxy_{m,t} \tag{3.4}$$

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

2.2.3 Canonical Correlation Analysis

Similar to PCA method, CCA also belongs to an eigenvector method, it decomposes the total variance of the climate fields as a linear combination of amplitude time series and their corresponding spatial patterns. In contrast to PCA of which the final object is to derive several new variables that are the eigenvectors of the cross-covariance matrix of climate fields for maximizing the explained variance, the CCA maximizes the temporal correlation of the relevant amplitude time series to construct pairs of predictor-predictand. These predictor-predictand variables are usually derived by calculating the inverse of the covariance matrices of each climate field. Besides, these calculated matrices can be pseudo-degenerate which usually leads to numerically unstable calculation processes. Regularization can be introduced to mitigate this unstable calculation by first projecting the original climate field onto several leading EOF patterns (Widmann, 2005; Pyrina et al., 2017), which also can help to reduce the number of degrees of freedom and eliminate underlying noise variance. A relatively small number of pattern pairs with high correlation will be retained by this dimensional transformation. In this study, we keep the number of PCs than can capture at least 90% cumulative temporal variance of target climate field. The canonical coefficients (CCs) time series and canonical correlation patterns (CCPs) for both target climate field and proxy can be derived by employing retained PC time series as input variables of CCA. The desired climate field can be reconstructed by a linear combination of the CCs with CCPs for each time step *t*:

$$\boldsymbol{x}_{m,t} = \sum_{n=1}^{l} CC_{n,t}^{field}\, \boldsymbol{CCP}_{m,n}^{field} \tag{3.5}$$

$$\boldsymbol{Proxy}_{m,t} = \sum_{n=1}^{l} CC_{n,t}^{proxy}\, \boldsymbol{CCP}_{m,n}^{proxy} \tag{3.6}$$

Proxy denotes the desired proxy time series, and *l* indicates the number of CCA pairs. The canonical correlation represents the correlation between each pair CC (proxy and climate field), which can be derived by the square root of CCA-eigenvalues. Thus, once each $CC^{proxy}(t)$ is derived from the input proxy data through the validation interval, the relevant $CC^{field}(t)$ can be estimated as proportional to $CC^{proxy}(t)$, since there are no correlation between the different $CC_n^{proxy}(t)$. The final desired climate field can be reconstructed by linear combination of $CC^{field}(t)$ and $CCP^{field}(t)$ with an assumption that the dominant CCPs of climate variability are stationary in time.

2.2.4 Bidirectional Long short-term memory neural network

We employ here a Bidirectional Long short-term memory neural network (Bi-LSTM), as one of nonlinear machine leaning method, for testing its capacity in reconstructing climate fields. The LSTM network is able to capture the potential information of the serial co-variability involved in climate data, which, especially in tacking with temporal climate dataset, can thus provide a suitable alternative CFR method.

From the available CFR methods that we could employ, recurrent neural networks (RNNs) are potentially an appropriate candidate with the objective of better reconstructing the true climate variability, because they can learn the serial correlation. In general, RNNs learn only the short-term serial correlation (Bengio, et al. 1994). Bi-LSTM is a special type of RNN, which it has been demonstrated that it is capable of learning and capturing long-term dependencies from a sequential dataset (Hochreiter & Schmidhuber, 1997). 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. We would like to test this property of the Bi-LSTM for paleo climate research based on our experiments. 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.

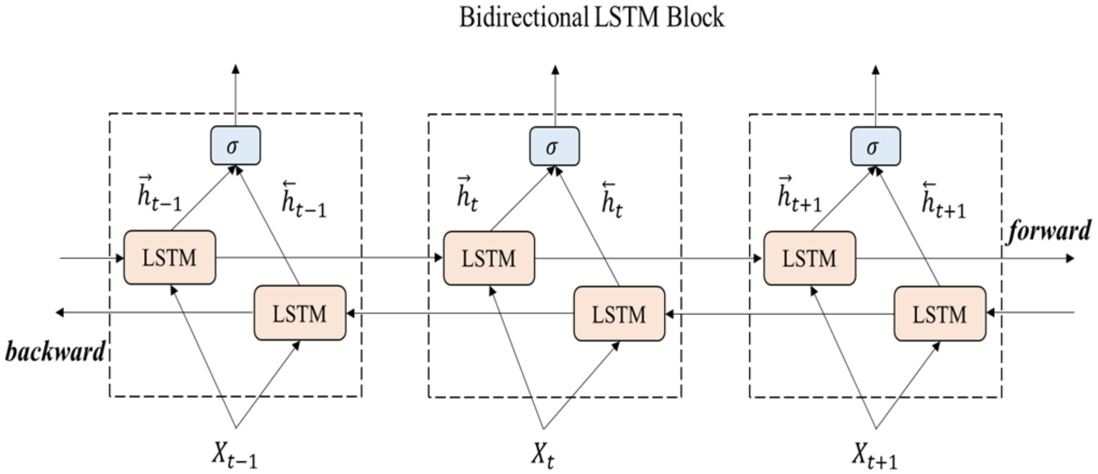

Figure 3.2: the bidirectional structure of the Bi-LSTM network

The training process of this network is achieved by feeding it with sequential data iteratively, backwards towards the past and forwards towards the future. Both backward and forward assimilations are derived by two individual LSTM layers, which are then connected and merged to the same output layer. Figure 3.2 shows the general structure of the Bi-LSTM method. Given a set of predictor-predictand pairs ($\boldsymbol{X_t}$, $\boldsymbol{Y}_\mathrm{t}$), our final object is to train a nonlinear function:

$$\widetilde{\boldsymbol{Y}_t} = F(\boldsymbol{X}) \tag{3.7}$$

The objective function employed in our CFR experiments for Bi-LSTM to be minimized during training process is the Huber loss that denotes the mismatch between the target climate field and the reconstructed climate field from simulations. The loss is minimized with gradient descent (Goodfellow et al., 2016). In addition, Huber loss can provide a key superiority that can be less sensitive to outliers:

$$L_\delta\big(\boldsymbol{Y}, f(\boldsymbol{X})\big) = \begin{cases} \frac{1}{2}\big(\boldsymbol{Y} - f(\boldsymbol{X})\big)^2 \\ \delta|\boldsymbol{Y} - f(\boldsymbol{X})| - \frac{1}{2}\delta^2 \end{cases} \tag{3.8}$$

where *f* represents the neural network, the brackets indicates the Euclidean norm. The Huber loss function varies from a quadratic to linear loss function when δ (a positive number) changes from small to big values (Meyer, 2020). This loss function can approach L1 when δ tends to be positive infinity and L2 loss when δ tends to be 0, here L1 is the sum of absolute deviations and L2 is the square root of the sum of squared deviations. We test the value of δ and finally set δ 1.35. The Bi-LSTM architecture employed in our CFR experiments is 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

2.2.5 Echo State Network

The reservoir computing (RC) methodology was developed by Maass et al. (2002) and Jaeger (2001), RC is essentially a category of Recurrent Neural Network. Echo State Network is one of the representative RC method, which usually consists of three basic layers: the input layer, the hidden layer-reservoir layer and the output layer-readout. For constructing the ESN method, a randomly reservoir will be initially created, which has the capability of preserving various non-linear transformations from input/predictor variables, and these non-linear transformations consist of hidden reservoir sates. Given the desired predictors-predictands pairs for methodology calibration and validation, ESN will be trained by employing calibration dataset based on linear regression so as to derive the relatively optimal weights of ESN architecture (input-to-reservoir, reservoir sate, reservoir-to-output), that is the desired target/predictand variables are derived by the linear combination of input and hidden reservoir sate weights (see Lukoševičius & Jaeger, 2009). In general, the training process in most type of artificial neural networks, including deep learning methods (RNN, LSTM), is accomplished by iteratively adapting and refining all connected thresholds and weights of the model architecture. In the RC-ESN method, the weights associated with input layer and hidden reservoir are held constant and randomly initialized, and the training process is realized by optimizing weights of reservoir-to-output employing linear regression (least square), which indicates that weights of hidden reservoir-to-output are the only trainable parameters in ESN. This outstanding characteristic enables

RC-ESN method to perform faster and even more accurately on linear or non-linear regression tasks (Pascanu et al., 2013; Goodfellow et al., 2016; Nadiga, B, 2020). The ESN structure is shown in Figure 3.3.

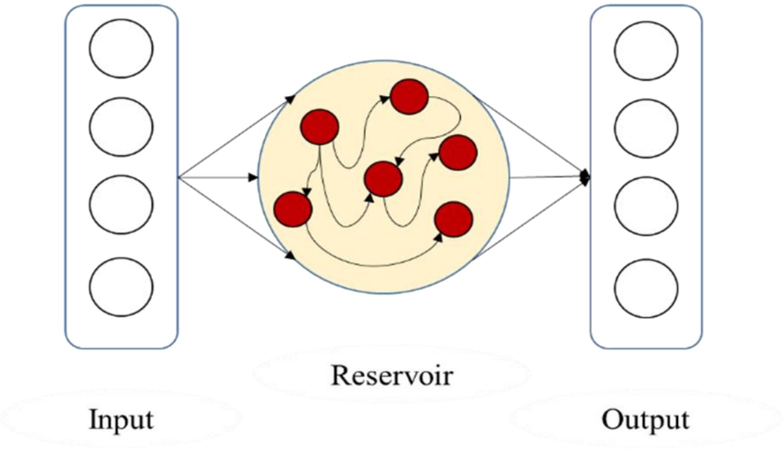

Figure 3.3: Structure of RC-Echo State Network

In this context, the classical RC-Echo State Network (Jaeger et al 2004, 2007; Chattopadhyay et al., 2020) is established. The ESN generally resembles an architecture of traditional neural networks, but the optimization process is different. Assuming we have input and target climate predictor-predictand pairs ($\boldsymbol{X}_t$, $\boldsymbol{Y}_t$, proxy and climate field correspondingly), $\boldsymbol{X_t}$ indicates the input proxy time series with length P that is equal to the proxy numbers employed in out CFR experiments, and it is also the input information for the input and hidden reservoir layer. In the construction of ESN, four weight matrices are initialized amongst input and reservoir layer; the initial hidden reservoir state $\boldsymbol{r}(t)$ which is a vector with dimension N (indicating N neurons are created in the hidden layer of ESN, here in this context, we set $N = 535$), weight matrix $\boldsymbol{M}$ of N neurons for each neuron connections with size of NxN, and the weight matrix $\boldsymbol{W}_{\text{in}}$for connecting input and hidden reservoir state layer with size of NxP; and the unit matrix $\boldsymbol{B}$ with size of NxN which can provide bias modulation in training process of ESN. Weight matrices $\boldsymbol{W}_{\text{in}}$ and $\boldsymbol{M}$ are initialized and originated from a uniform distribution within interval [-1, 1] (Jaeger et al 2004, 2007; Chattopadhyay et al. 2020). Thereby the updated hidden reservoir state can be derived based on these initialized weight matrices:

$$\hat{\boldsymbol{r}}(t) = relu[\boldsymbol{M}\boldsymbol{r}_t + \boldsymbol{W}_{in}\boldsymbol{X}_t + \boldsymbol{B}] \tag{3.9}$$

the derived $\hat{\boldsymbol{r}}(t)$ will be employed by combining the reservoir to output weight matrix $\boldsymbol{W}_{\text{out}}$, the desired climate field can then be reconstructed by:

$$\widehat{\boldsymbol{Y}}(t) = \boldsymbol{W}_{out}\hat{\boldsymbol{r}}(t) \tag{3.10}$$

where $\boldsymbol{W}_{\mathrm{out}}$ indicates the only trainable parameter amongst all initialized parameters, is defined and minimized as:

$$\boldsymbol{W}_{out} = \arg min\, \boldsymbol{W}_{out}\|\boldsymbol{W}_{out}\hat{\boldsymbol{r}}(t) - \boldsymbol{Y}_t\| + \alpha\|\boldsymbol{W}_{out}\| \tag{3.11}$$

this trainable matrix can fit the relationships between target climate field $\boldsymbol{Y}_{\mathrm{t}}$ and the reservoir state $\hat{\boldsymbol{r}}(t)$. $\|\quad\|$ represents least square norm for a vector and α denotes ridge regression coefficient (Lu et al., 2018).

This ESN architecture actually yields a simple training process compared to traditional neural networks, and thus can provide several advantages: only the weight matrix $\boldsymbol{W}_{\mathrm{out}}$ of readout can be trained in the training process, this indicates an orders of magnitude faster than backpropagation algorithm (Goodfellow et al., 2016); It does not suffer from the exploding and vanishing gradient issues especially compared with the training process in RNNs (Pascanu et al., 2013). The hypothesis of employing RC-ESN in our paleoclimate field reconstruction is that we would like to test whether this type of structural and theoretical simpler machine learning method compared to the relatively complex Bi-LSTM could provide some potential outperformance. Detailed ESN parameters are shown in Appendix 3A.

3 NH CFR experiments

The NH summer temperature anomalies reconstructions using four CFR methodologies amongst three climate models are derived and illustrated in Fig. 3.4-3.5. The employed 30 tree ring based pseudoproxies' geolocation is plotted as white circles in all the sub-figures.

Table 3.1. Skill reconstruction statistics for the North Hemisphere mean temperature in the verification period for ideal PPEs. The table shows the result for four CFR methods (PCR, CCA, ESN and Bi-LSTM) and three climate models (MPI, CAM and CCSM). The numbers in parenthesis indicate the skill statistics of noise contaminated PPEs.

Method	SD Ratio			*cc*		
	MPI	CAM	CCSM	MPI	CAM	CCSM
PCR	0.7175(0.6316)	0.7250(0.6697)	0.8027(0.6218)	**0.3754**(0.1753)	0.4610(0.1809)	**0.5264**(0.2438)
CCA	0.5077(0.5169)	0.5526(0.5357)	0.5679(0.5412)	0.3696(0.1876)	0.4374(0.1925)	0.5017(0.2490)
Bi-LSTM	0.5607(0.5317)	0.6402(0.5690)	0.6107(0.4079)	0.3658(0.1852)	**0.4699(0.2025)**	0.5093(0.2193)
ESN	**0.8586(0.7838)**	**0.8529(0.8182**)	**0.9042(0.8032)**	0.3371(**0.2015**)	0.3943(0.1839)	0.4705(**0.2658**)

The derived temperature anomaly *cc* maps and SD ratio distributions for ideal PPEs of NH are shown in Fig. 3.4, a general conclusion is that all four CFR methodologies yield generally consistent spatial distributions amongst each of the climate models employed herein, whereas, these spatial patterns differ across MPI, CAM5 and CCSM models. For NH CFRs based on all methods, skillful reconstructions are achieved over regions where denser pseudoproxies exist (over the extratropical regions including North American and Eurasia regions as shown in Fig. 3.4). Higher *cc* and SD ratio derived from four methods generally coincide with pseudoproxy samplings with high density, yet very low values arise over sparsely sampled regions. Relatively weak *cc* values occur in tropical regions that are not adjacent to the pseudoproxy positions that we sampled amongst almost all derived CFRs in Fig. 3.4, indicating that the teleconnections between tropics and mid-latitude regions could be detected and identified. All derived CFRs utilizing ideal PPEs suffer from variance losses as shown in SD ratio maps in Fig. 3.4 and in Table 3.1 correspondingly. The spatial distributions of the skill vary specifically between climate models and CFR methodologies, and are spatially heterogeneous. For instance, comparing the spatial patterns of derived SD ratio in Fig. 3.4 amongst the four models, more variance (the mean SD ratio is approximately 0.1 higher) in the high-density pseudoproxy sampling and the tropical regions are preserved in high-resolution CCSM model based PPEs especially in the PCR and ESN methods. CCA and Bi-LSTM method present relatively consistent derived SD ratio patterns as shown in Fig. 3.4, which also seems to suffer more variance losses (also see Table 3.1) over the NH comparing that to PCR and ESN methods. Moreover, the ESN method captures the most variance in the ideal PPEs compared with the rest methods as shown in Fig. 3.4, also see Table 3.1 for spatial mean statistics.

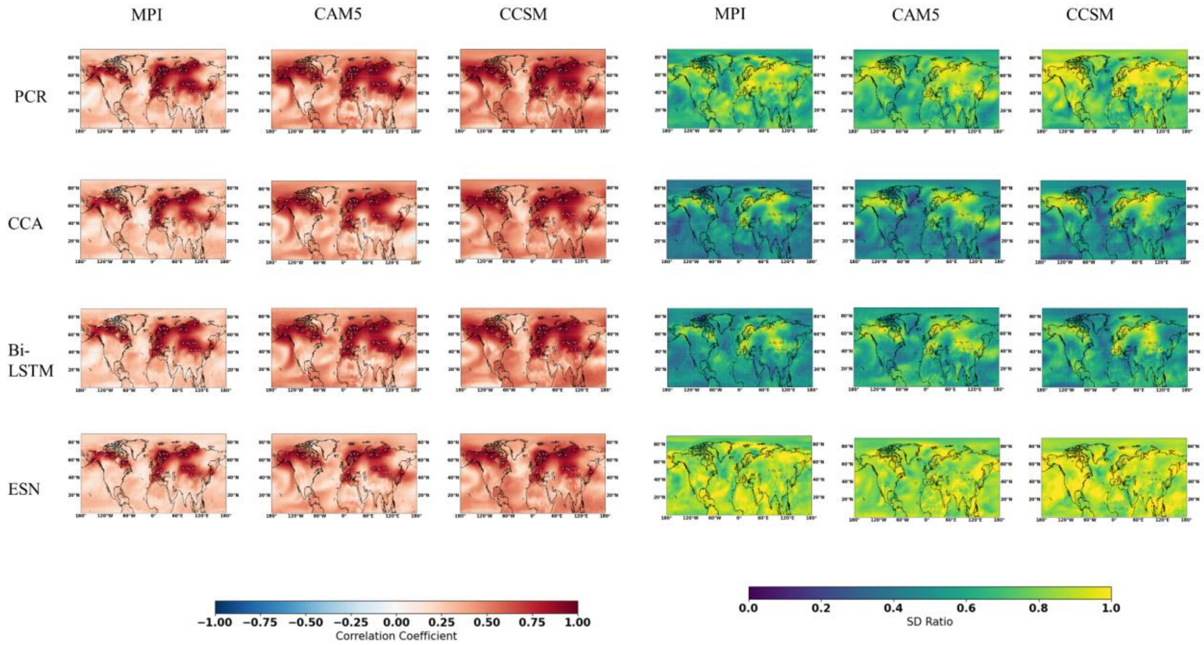

Figure 3.4: NH Reconstruction results of CFR methods (including PCR, CCA, Bi-LSTM, ESN) using MPI, CESM1-CAM5 and CCSM numerical simulation as target temperature field, all the CFR methods employ

the same proxy network with full 30 ideal pseudoproxies. 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

The noise contaminated PPEs are illustrated in Fig. 3.5. The performance deterioration can be expected, and significant CFR skill reduction occurs over regions where dense pseudoproxies are located. The CCA and the nonlinear method Bi-LSTM seem to suffer more from spatial variance losses again over the NH region (see Table 3.1). Both PCR and ESN seem to be able to preserve more variance in the noisy PPEs compared with CCA and Bi-LSTM (see mean statistics in Table 3.1 and the derived spatial statistical patterns in Fig. 3.5). ESN successfully captures most temperature anomaly variance amongst all four CFR methods, although presents more similar derived temporal covariance as others shown in Fig. 3.5, which might indicate that nonlinear-based ESN could be more robust against the presence of noisy in the records. The overall methodological performance is relatively consistent across all of the three model based PPEs. ESN and PCR generally outperform the Bi-LSTM and CCA methodology with higher mean SD ratio in noisy PPEs, but the ESN method achieve best performance in both ideal noise contaminated PPEs on the aspect of capturing more variance, which could be seen in Table 3.1.

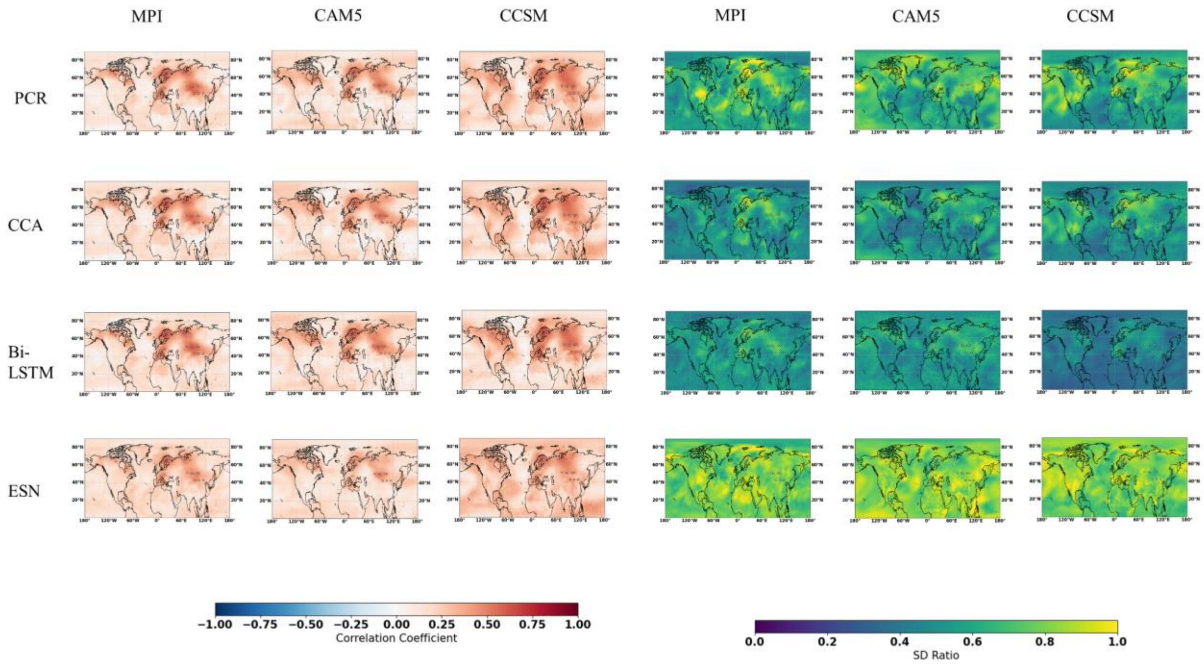

Figure 3.5: the same as Figure 3.4, but for noise-contaminated PPEs (NH temperature reconstructions using noise PPEs)

Despite the relatively different spatial performance across all evaluated CFR techniques illustrated in *cc* and spatial SD ratio patterns (Fig. 3.4-3.5) for preserving variance, the total errors across all CFR methods within same model based PPEs respectively are similar. The general methodological skill, as indicated by the derived *cc* and SD ratio values in Table 3.1, the CCA and Bi-LSTM generally present worse performance with relatively lower mean *cc* and SD ratio in both ideal and noise contaminated PPEs. Overall, ESN and PCR generally outperforms CCA and Bi-LSTM with highest mean cc and SD ratio values across all PPEs in the three model.

4 NH indices reconstructions

The decadal time series of NH mean summer temperature anomalies for different CFR methodologies across the three models employing ideal tree-ring network is illustrated in Fig. 3.6 on the left panel. All time series have been smoothed by using a Butter worth low-pass filter to remove temporal fluctuations less than 10 years. Based on these derived mean indices, all four CFR methods generally exhibit a certain degree of underestimation in reconstructing the total temperature amplitude and variance of targets, especially during the most recent 20th century. An interesting finding is that the nonlinear-based ESN method might be capable of achieving a skillful performance in capturing some variation signals over extreme cooling events compared with PCR, CCA and Bi-LSTM (different volcanic eruptions period selected as shown in the panel of each subfigure in Fig. 3.7). Nevertheless, considering the comprehensive performance on capturing extreme temperature signals, all four method perform comparably. In the derived mean indices based on ideal PPEs, all four CFR methods generally achieve skillful reconstructions compare to each target simulations as shown on the left panel in Fig. 3.6. In addition, the derived indices based on noisy PPEs as shown on the right panel in Fig. 3.6 reveal significant underestimation especially in the anomaly amplitude aspect amongst all model simulations and reconstruction methods. Specifically, in both MPI and CCSM models, all four CFR methods tend to underestimate the most 20th century temperature variabilities (see Fig. 3.6), which emphasizes that more internal climatic variabilities involved and existed in tropical regions (since we just employ extratropical proxy locations in this context to reconstruct entire Norther Hemisphere summer temperature anomalies). The derived noisy PPEs results in substantial skill reduction (*cc*, SD and RMSD displayed within brackets in Table 3.2). All four CFR methods generally fail to preserve the complete variance of targeted model simulations and, the magnitude of strong cooling events are significantly underestimated especially in noisy PPEs (see the left panel in Fig. 3.7).

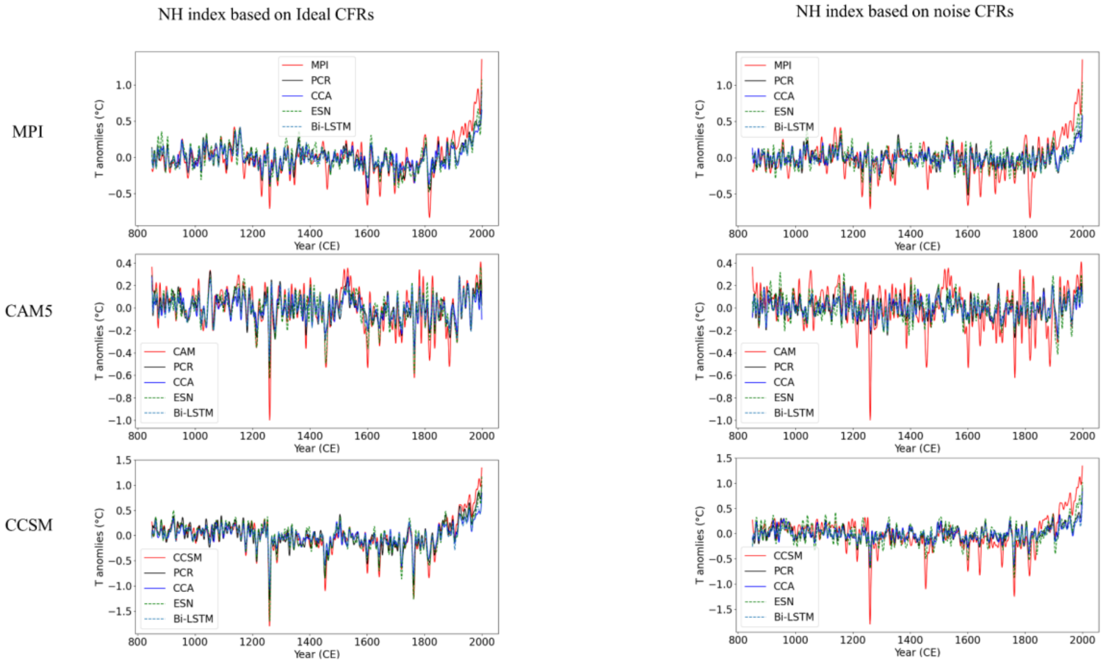

Figure 3.6: Decadal NH temperature indices derived from CFRs

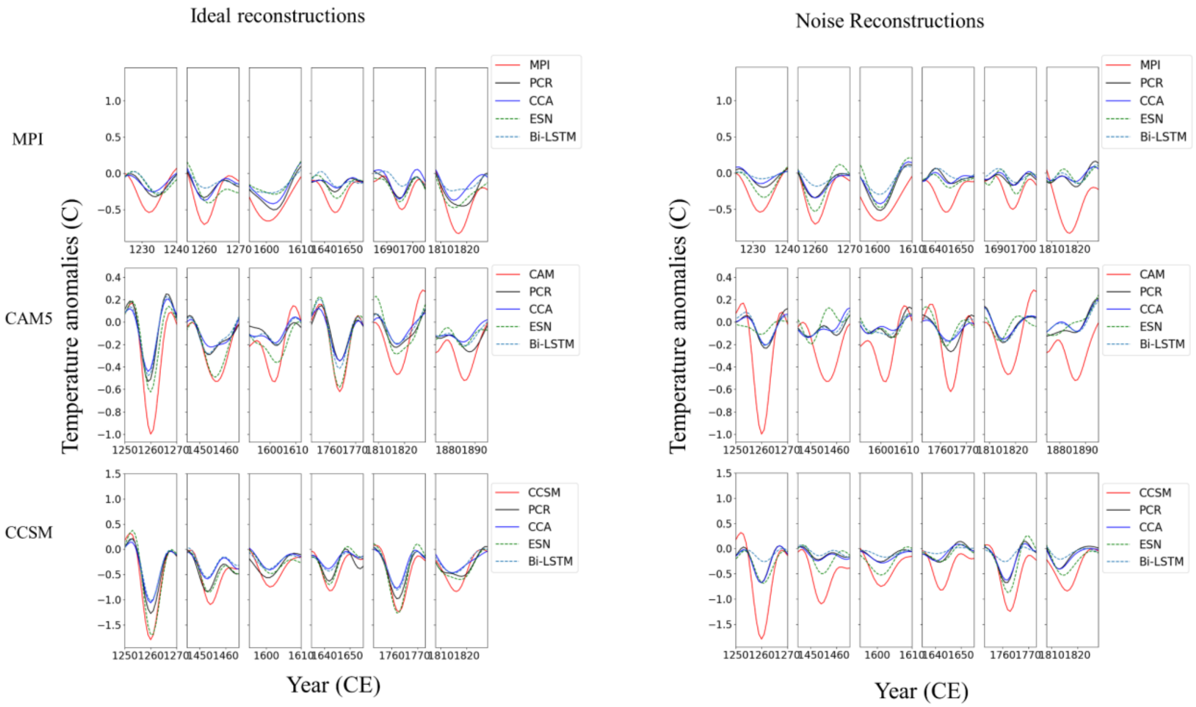

Figure 3.7 Selected extreme cooling events reconstructions

Table 3.2. *cc*, SD and RMSE (Celsius) during the verification interval for decadal NH mean temperature derived from ideal PPEs. The numbers in parenthesis indicate *cc*, SD and RMSE (Celsius) of noise contaminated PPEs.

Method	*cc*			SD			RMSE		
	MPI	CAM	CCSM	MPI	CAM	CCSM	MPI	CAM	CCSM
PCR	**0.8655**	0.8222	**0.9335**	0.6939	0.6593	0.8301	**0.1258**	0.1048	**0.1209**
	(0.6532)	(0.5206)	**(0.7341**)	(0.4873)	(0.4991)	(0.4737)	(0.1843)	(0.1513)	(0.2357)
CCA	0.8526	**0.8344**	0.9277	0.5876	0.5602	0.6154	0.1393	0.1090	0.1578
	(**0.6905**)	(0.5103)	(0.7337)	(0.4251)	(0.4173)	(0.4443)	(0.1832)	(0.1532)	(0.2394)
Bi-LSTM	0.8115	0.8274	0.9293	0.5958	0.5998	0.6069	0.1481	0.1073	0.1589
	(0.6198)	**(0.5373)**	(0.7075)	(0.3844)	(0.4399)	(0.2850)	(0.1948)	(**0.1504**)	(0.2669)
ESN	07945	0.8261	0.9242	**0.7511**	**0.7664**	**0.8755**	0.1448	**0.1005**	0.1248
	(0.6639)	(0.4039)	(0.7191)	**(0.6433)**	(**0.5868**)	(**0.6596**)	(**0.1783**)	(0.1654)	(**0.2261**)

In general, all CFR methods in the three climate models based PPEs generally capture both the amplitude and phase of the NH mean variability. Nevertheless, based on the derived spatial *cc* and SD ratio patterns, relatively noticeable discrepancies could arise amongst different CFR results and different model runs, which makes it relatively difficult to advocate one CFR technique over another (Smerdon et al,. 2011), resulting in a conclusion that area mean time series might be generally insufficient for assessing spatial reconstruction skills.

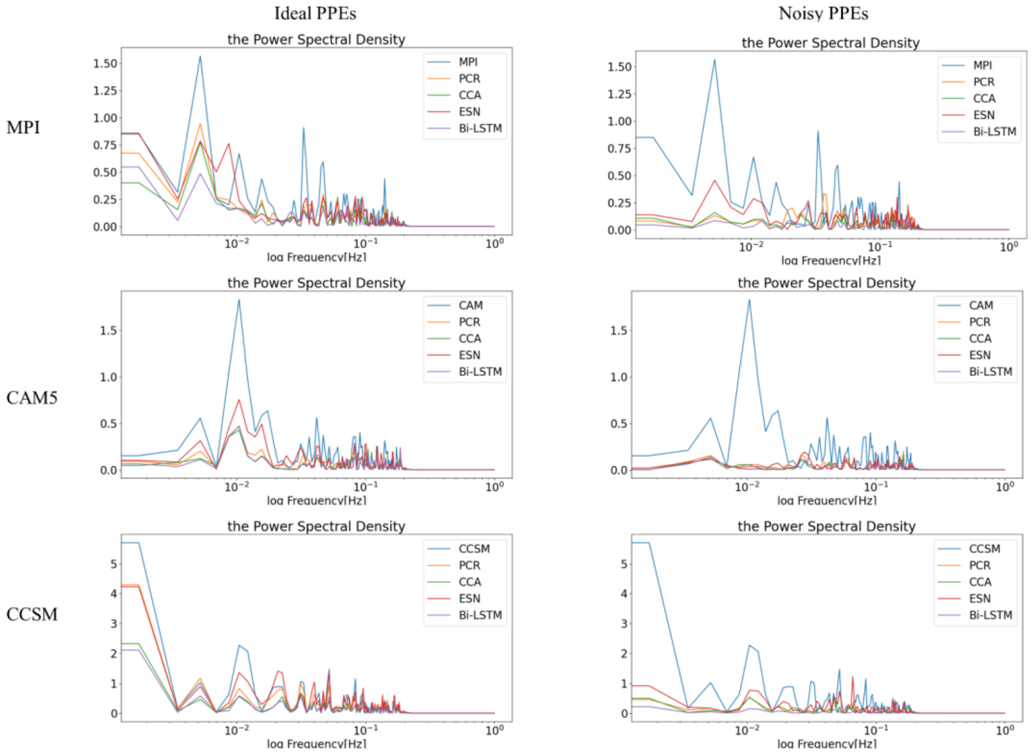

Figure 3.8: North Hemisphere indices power spectral density

Fig. 3.8 shows the comparison of mean NH summer temperature indices of power spectral density for both ideal and noisy PPEs between reconstructions and target model simulations. As illustrated in Fig. 3.8, all four CFR methods generally underestimation the targeted power spectra; specifically, this underestimation is more significant in the derived noisy PPEs; these power spectral plots could also provide us a deep insight of potential underestimation CFR performance. Moreover, it seems that the ESN method employed herein our PPEs generally exhibits superiorities in capturing the targeted power spectral amongst all models and across ideal and noisy PPEs (as shown in Fig. 3.8).

5 Additional metrics for derived NH mean indices

Even though the four CFR methods tend to underestimate the total variability in noisy PPEs, an interesting question is whether they can perform skillfully in reproducing the probability distributions of the temperature indices. Specifically, a relevant question is whether these CFR methodologies are able to capture some extreme phases of those indices. Fig. 3.9 illustrates the histogram for decadal NH mean indices. Each subfigure in Fig. 3.9 indicates the histogram of reconstructed temperature indices across the four CFR methods, compared with the histogram of the target temperature indices.

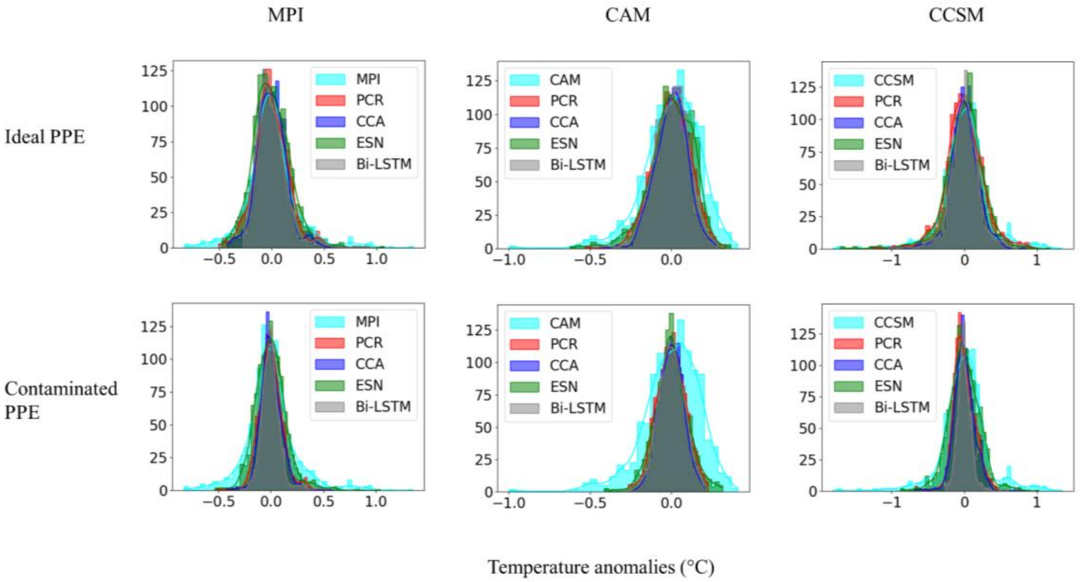

Figure 3.9: Histogram for decadal filtered NH mean index. The x axis denotes temperature anomaly values, and y axis is the number of data in each bin. Totally 30 bins are selected to plot each of the histogram.

For both perfect and noisy pseudoproxies in Fig. 3.9, the ESN reconstructions seem to reproduce the overall target distribution best. It can generally capture the lower tail better than the rest three CFR methodologies.

Moreover, these differences between the CFR methods become smaller in noisy PPEs, with the ESN still generally performing better than the other methods (subfigures for the noisy PPEs in Fig. 3.9). The nonlinear Bi-LSTM method however fails to capture the lower and upper tails of distribution amongst the four methods.

Table 3.3. 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 noise contaminated PPEs.

Method	KS statistic			p-value		
	MPI	CAM	CCSM	MPI	CAM	CCSM
PCR	0.056(0.127)	0.115(0.165)	**0.035**(0.141)	6e-2(1e-8)	5e-7 (4e-14)	5e-1(2e-10)
CCA	0.096(0.150)	0.164(0.213)	0.086(0.144)	5e-10(9e-12)	5e-14 (3e-23)	3e-4(7e-10)
Bi-LSTM	0.103(0.188)	0.142(0.198)	0.083(0.218)	1e-5(3e-18)	2e-10(4e-20)	8e-4(2e-24)
ESN	**0.038(0.082)**	**0.096(0.154)**	0.039(**0.045**)	3e-1(7e-4)	4e-5 (2e-12)	3e-1(2e-1)

In order to statistically quantify the similarity of indices distributions between targets and reconstructions, the two-sample Kolmogorov-Smirnov are employed as a metric (Hodges, 1958) (see Table 3.3). The smallest KS statistic is generally obtained by the ESN method (see Table 3.3), confirming that the ESN can generally outperform other CFR methods for temperature mean indices reconstructions in both the ideal and noisy PPEs.

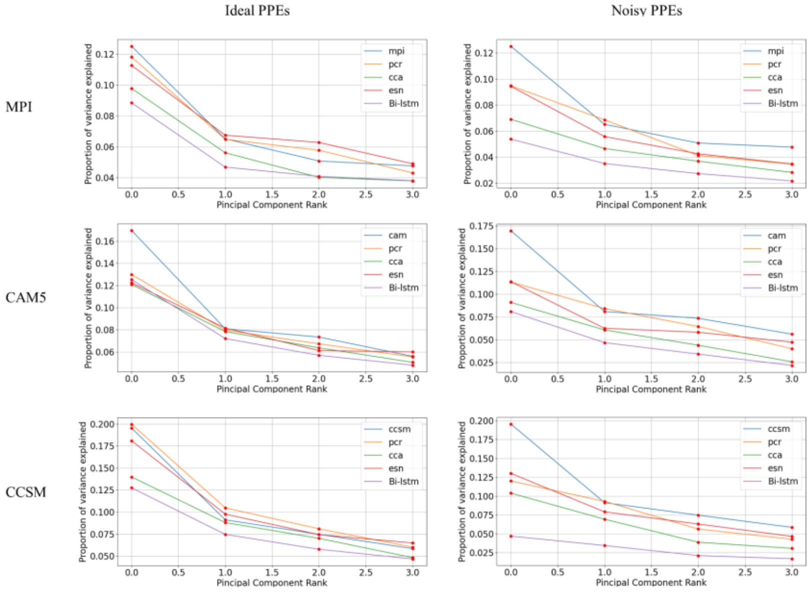

Figure 3.10: Eigenvalue spectra for model simulations and four method reconstructions: the spectra for model simulation and four method-based ideal PPEs are computed as the ratio between each of the first four reconstructed eigenvalues and the cumulative sum of all eigenvalues from target climate model

Several leading EOF-PC pairs are derived from reconstructions utilizing four CFR methods amongst three models as shown in Fig. 3.10. These four leading EOF-PC pairs could provide us a deeper insight about the general CFR performance, since dominant climate patterns can usually explained by several leading EOFs, and better reproductive leading EOF-PC patterns indicate better reconstruction performance. More explained variance could be reconstructed compared with targets indicates better reconstruction performance. As illustrated in Fig. 3.10, amongst all three model based PPEs, the Bi-LSTM and CCA method generally fail to better capture the variance in the first four leading patterns. The ESN and PCR method generally present comparable performance on reconstructing explained variance in these leading EOF patterns.

6 Discussion

6.1 Reconstruction performance based on Machine learning method

The ESN method firstly employed in our context for paleo-temperature field reconstructions exhibits an encouraging performance in capturing more temperature variance amongst all three models, and can generally perform better than the Bi-LSTM method of which more complicated structures and parameters exist compared to the ESN. In addition, the ESN method also achieves better reconstructions than the two traditional multivariate linear-based regression method PCR and CCA, although PCR shows more superiorities compared with CCA and Bi-LSTM.

Both ESN and Bi-LSTM method generally present resemble derived *cc* patterns as shown in Fig 3.4-3.5, yet ESN shows relatively worse mean *cc* see Table 3.1. Nevertheless, The ESN method successfully capture more variance in both ideal and noisy PPEs (see Fig 3.4-3.5 and Table 3.1). These indicate that the ESN method could preserve more temperature variance based on our reconstruction experiments. In the derived mean indices, the ESN method still exhibits superiority in capturing more variance (see SD values in Table 3.2) amongst all three models although with a relatively worse mean *cc*. Interestingly, the ESN seems to better capture the generally temperature rising tendency especially during the most 20th century in both noise and ideal PPEs as illustrated in Fig. 3.6, whereas all the rest three CFR methods fails. In additional, based on several selected extreme cooling events prior to 20th century plotted in Fig. 3.8, the ESN provides us with an encouraging performance in capturing some extreme cooling signals amongst all four methods. The reconstructed power spectral density compared with target models in Fig. 3.8 further provides us with a deep insight about the better reconstruction performance of ESN method.

Our implementation of the more complex Bi-LSTM does not show superiority in CFRs, at least in our specific experiments, compared to traditional CFR methods and the relatively simpler ESN methods, so our conclusion is that more complicated architecture might not be helpful to reconstruct the surface temperature field at least based on our specific experiments and the employed architecture of Bi-LSTM. Besides, one factor needs to be emphasized is that the degradation of performance may well be expected when a limited dataset is employed to train a neural network model (Najafabadi et al., 2015). Also splitting data inappropriately might cause unexpected effects on the general performance of one neural network model, since the data might contain trends or shifts in covariance in time - these could result from variations in the way the data was gathered or from varying choices over what information to collect (Riley, 2019) as well as from variations in climate conditions. Nevertheless, based on the general performance achieved by the ESN employed in our study, we would like to draw a general conclusion that employing relatively simpler architecture-based nonlinear machine learning method might be helpful for CFRs at least based on our specific experiments and the selected architecture of ESN. Both ESN and LSTM belong to RNN, yet ESN is much simpler than LSTM (Lukosevicius and Jaeger 2009), and has been demonstrated to be able to outperform the RNN methods in different applications (Chattopadhyay et al., 2019; Nadiga, B. 2020). We thus encourage testing ESN in different paleoclimate research directions for further study.

Nonlinear machine methods are usually capable of mapping chaotic and dynamic systems with highly nonlinearity. In this context, we utilize two nonlinear neural network methods ESN and Bi-LSTM to test their performance on NH paleo temperature reconstructions. Compared with linear methods PCR and CCA, the more complicated Bi-LSTM neural network model did not show clear advantages, whereas the employed simpler ESN method presents encouraging advantages. Nadiga (2020) pointed out that the capability of some machine learning methodologies are strongly dependent on the selection and setting of hyper-parameters. An extensive range of complexity can be involved in Machine learning methods, and thus it remains an open question as to which ML method is relatively or most suitable for paleoclimate analysis. However, until now it is still not very clear how to optimize the architecture of the ML methods in a systematically way, and especially how to illustrate the interpretability of ML methods. A considerably amount of ‘trial and error’ still remains in the design of the neural layers. Here, we have tested two ML methods originated from the family of, and finally employed two separated hidden layers for Bi-LSTM, and three layers for ESN, and evaluated their reconstruction performances on paleoclimate experiments, which could be a preliminary attempt.

6.2 Method and climate model dependencies

For the large-scale NH PPEs, skillful reconstructions are achieved over regions where denser pseudoproxies exits. Furthermore, weak climatic teleconnections are detected amongst tropical regions and some extratropical areas (Fig. 3.4-3.5). Potential uncertainties also need to be considered when the interpretation is performed on short temporal experiments using different methodologies (Qasmi et al., 2017). The relatively different spatial *cc* and SD ratio patterns across different models also indicate that the teleconnection is model dependent. For instance, the temporal stationarity and spatial pattern of teleconnections between local regions of the global climate field and the tropical Pacific region could vary substantially across CMIP5 ensemble runs (Coats et al., 2013a, 2015b; Lewis and LeGrande, 2015). In general, our NH CFR performance is relatively consistent across the three model based PPEs, yet some indispensable discrepancies occur. For instance, all of the CFRs display reasonable skills over extratropical regions with denser sampled proxy network, while this skillful reconstruction is relatively better for CCSM, compared to MPI and CAM5 (as also shown in Table 3.1 with relatively higher mean SD ratio and *cc*). The model dependency is also deeply revealed by mean correlation coefficient when we compare the reconstruction results within same proxy network in Table 3.1, which is that the mean *cc* in CCSM is generally better than that of CAM5 and MPI regardless of the employed CFR methods. Moreover, the derived spatial SD ratio maps as shown in Fig. 3.4-3.5 reveals an evidently methodological dependency, both ESN and PCR present relatively skillful variance preservation in both ideal and noisy PPEs better than CCA and Bi-LSTM, and the ESN seems to preserve the most temperature variance amongst all the four methods and across all three models.

Considering methodological performance on the aspect of indices reconstructions, all NH indices derived from CFR experiments demonstrate that four CFR techniques generally exhibit consistent and comparable performance both in phase and amplitude aspect amongst the three models. Generally, the overall 10-year filtered low-frequency signal (Fig. 3.6) is reasonably reconstructed by all CFR techniques. Besides, all decadal indices time series derived from CFRs suffer variance/magnitude losses compared with target model means, and the underestimation in amplitude varies relatively amongst CFR methods and amongst target models; such failures to completely capture the variance of some colder or warmer years play an indispensable implications for interpreting the variability of past climate (McCarroll et al., 2015). A potential interpretation could be that even if the best PPEs are constructed with best proxies, these PPEs might be imperfect recorders of natural variability of climate. This relates to the amount of noise being inherent in these best PPEs and, these internal noises will be inherited in any large or local reconstructions sequentially (Christiansen, 2011). Another interpretation is that reconstructions with low skill over sub-regions with a sparse or no proxy network will contaminate the overall performance directly. However, mean temperature indices cannot reflect climate variability and uncertainties for regional and or local spatial scales. These results could be the indicator that regional/hemispheric climate reconstructions could underestimate low-frequency time series signals both in amplitude and trends (Smerdon et al., 2011; Christiansen, 2011; Wang et al., 2014; Guillot et al., 2015).

In general, the ESN outperforms other three methods on capturing the overall temperature signal variance including extreme events; however, underlying uncertainties could appear across all nonlinear and linear methods regarding the large spatial variations in CFRs for the different spatial reconstruction metrics (*cc* and SD ratio).

7 Conclusion

Based on our PPEs setup using a set of state-of-the-art Earth System Models, NH summer temperature field could be reasonably reconstructed employing different CFR methodologies. Specifically, continental-based tree ring proxies provide a great contribution to the reconstructions of land temperature for the last millennium. For each specific PPE networks established in this context, all CFR methodologies produce generally skillful reconstructions in regions where denser pseudoproxy networks are available. Moreover, teleconnections are detected by these CFR methodologies, leading to weak spatial reconstruction skills outside of the proxy-sampled regions. All four CFR methods generally underestimate the target temperature variations to a rising tendency as random noise introduced into the perfect PPE setup. Furthermore, based on the derived spatial metrics (*cc* and SD ratio), inter-methods and inter-models discrepancies are noticeably exposed within different PPEs, and the spatial performance of CFR techniques employed in the context reveal some potential limits on the capability of currently constructed regression methods. For instance, extracting potential information from noisy and sparse perfect proxies, also indicates that no individual CFR methods can produce field reconstructions with universally better performance.

Another finding is that the amplitude of the 10-year filtered mean time series over the NH are generally underestimated, especially in some extreme cooling periods and recent warming decades. All CFR techniques systematically exhibit large underestimations and biases both in amplitude and evolution trends of the low-frequency signals especially in the PPEs with noise contamination (Fig. 3.7), while its general phase is basically reconstructed. Relatively large biases and variance occur in the entire regional mean combining with relatively skillful spatial reconstructions based on spatial metrics (cc and SD ratio) indicating that entire regional means are relatively insufficient for assessing spatial performance of CFR techniques.

In general, based on our specific experiments, the ESN method employed herein generally present more superiorities compared to the rest three CFR methods both in spatial and temporal temperature reconstructions. In addition, reasonable spatial and temporal reconstruction performance of CFRs achieved in our experiments by employing fundamental nonlinear machine-learning method based on relatively limited database confirms that machine learning/deep learning methodologies represent certain generalization capability (Zhang et al., 2017).

Chapter 4: Reconstruction of the Basin-Wide Sea Level Variability in the North Sea

Reconstruction of the Basin-Wide Sea Level Variability in the North Sea Using Coastal Data and Generative Adversarial Networks

Manuscript published in J. Geophys. Res. (Zhang et al., 2022)

Summary

We present an application of generative adversarial networks (GANs) to reconstruct the sea level of the North Sea using a limited amount of data from tidal gauges (TGs). The application of this technique, which learns how to generate datasets with the same statistics as the training set, is explained in detail to ensure that interested scientists can implement it in similar or different oceanographic cases. Training is performed for all of 2016, and the model is validated on data from three months in 2017. Tests with datasets generated by an operational model ("true data") demonstrated that using data from only 19 locations where TGs permanently operate is sufficient to generate an adequate reconstruction of the sea surface height (SSH). The machine learning (ML) approach appeared successful when learning from different sources, which enabled us to feed the network with real observations from TGs and produce high-quality reconstructions of the basin-wide SSH. Individual reconstruction experiments using different combinations of training and target data during the training and validation process demonstrated similarities with data assimilation when errors in the data and model were not handled appropriately. The proposed method demonstrated good skill when analyzing both the full signal, as well as the low frequency variability only. It was demonstrated that GANs are also skillful at learning and replicating processes with multiple time scales. The different skills in different areas of the North Sea are explained by the different signal-to-noise ratios associated with differences in regional dynamics.

1 Introduction

In the first part of the 20th century, Proudman and Doodson [1924] demonstrated how the fundamental dynamical equations of the tides may be used to obtain knowledge of the distribution of the surface elevation over the entire North Sea from observational data. They showed that the cotidal and corange lines can be easily determined, provided the elevation in some coastal stations and limited amount of open sea locations is known, and some local current observations exist. Additionally, they used some hypotheses about the frictional forces. In the present study, approximately 100 years later, we encounter the same issue from a different perspective.

Many important developments in the physical oceanography of the North Sea have followed the study of Proudman and Doodson [1924]. Numerical modeling of tides and storm surges initiated by Hansen [1956] and Heaps [1969] has become a fundamental tool in surge prediction [Soetje and Brockmann, 1983; Peeck et al., 1983; Flather and Proctor, 1983]. A dense network of tidal stations has been developed around the North Sea coasts, which operates over long periods and provides high-quality records [Wahl et al., 2013]. A comparison between numerical simulations and satellite observations [Andersen, 1999] revealed good agreement between the two data sources. The low-frequency variability in altimeter and tide gauge data over the North Sea shows a reasonably good correlation [Cipollini et al., 2017]. A recent important development in predicting the sea level in the North Sea is achieved within the framework of the North-West European Shelf forecasting system [Tonani et al., 2019].

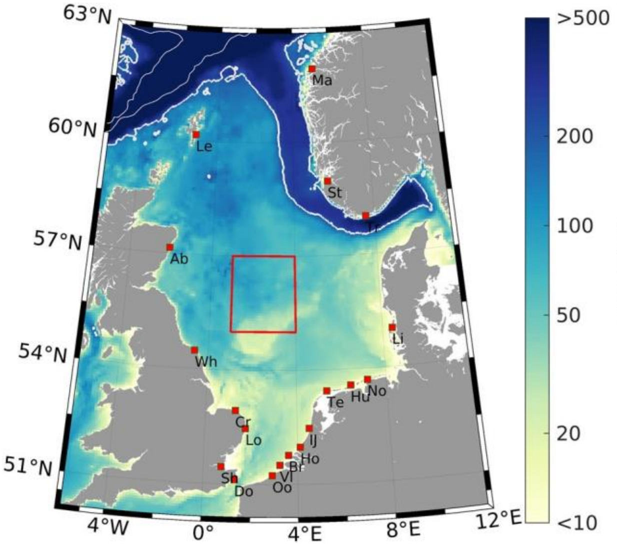

Figure 4.1 Topography of the North Sea, the positions of the TG stations (red squares), and the subsampled region (grid of 32 × 32 points) shown by the large red rectangle. Ab, Aberdeen; Br, Brouwershavensegat 8; Cr, Cromer; Do, Dover; Hu, Huibertgat; Ij, IJgeulstroompaal 1; Le, Lerwick; Li, List; Lo, Lowestoft; Ma, Maloy; No, Hoek van Holland; No, Norderne; Oo, Oostende; Sh, Sheerness; St, Stavanger; Te, Terschelling Noordzee; Tr, Tregde; VL, Vlakte van de Raan; Wh, Whitby

The North Sea (Figure 4.1) is a shallow sea located at the European continental shelf with an average depth of ~ 90 m [Otto et al., 1990; Huthnance, 1991; Becker et al., 1992]. The sea-level dynamics in this basin can be considered as a response to different forcings, such as barotropic and baroclinic tides [Haigh et al., 2019], wind and atmospheric pressure, air-sea heat and water exchanges, as well as forcing from the open boundaries and rivers. The processes that dominate the dynamics are, in most cases, coupled; that is, one cannot easily consider the response to individual drivers in isolation. Thus, there is a need to use analysis methods tailored to detect and reproduce nonlinear dynamics. Deep learning, which has much in common with neural networks, is well suited to resolve such processes. Our first objective in the present study is to explore the performance of deep learning techniques when reconstructing the basin-wide sea level in the North Sea using data from coastal stations. In our specific application, we will use generative adversarial networks [GANs; Goodfellow et al., 2014]. This technique learns how to generate datasets with the same statistics as the training set. Unlike some previous studies [e.g., Chipolini et al., 2017], we will focus both on the shorter-term and longer-term variability ranging from intratidal to monthly time scales. Under “exploring the performance of deep learning techniques” we understand also identifying the application limits. This will be illustrated by setting up several experiments, with different reconstruction potential.

Our second objective is to compare the goodness of reconstructions based on adversarial networks against other known reconstruction methods. One such method, which uses a Kalman filter approach, was applied by Grayek et al. [2011] to extrapolate one-dimensional FerryBox data acquired along the ferry routes to larger two-dimensional areas.

We are not aware of any applications of deep learning to sea-level analysis and reconstruction, particularly in the region of the North Sea. This justifies our third objective, which is to present our results in a way that they ensure reproducibility by interested scientists and motivate potential oceanographic applications using similar or different data sets. Therefore, we will analyze many individual steps that led to the final application of the method to the entire North Sea area. The major focus is on what GANs can reconstruct successfully and what they cannot. The analysis of the results demonstrates the power of reconstructions based on GANs. The paper is structured as follows. In section 2, we present the methods used. Section 3 presents the experiments; section 4 presents the results, followed by discussion in section 5 and conclusions.

2 Methods

2.1 Data

2.1.1 Data from the operational model for the North West European Shelf

The reconstruction of the basin-wide sea level using data from coastal stations necessitates high-quality data from observations over the entire North Sea. Data with such coverage are available only from satellites. However, they cannot perfectly resolve the spatial and temporal variability, particularly at scales shorter than the time of the repeat cycle. Furthermore, close to the coast, these data are not quite accurate [Chipolini et al., 2017]. Data from numerical models, although not absolutely correct, provide spatial and temporal coverage over the entire basin. Therefore, when developing and testing our method, we will use data from numerical simulations to represent the "true" sea level. In this research, the dataset was obtained from the operational numerical Forecasting Ocean Assimilation Model (FOAM) with 7 km horizontal resolution, known as Atlantic Margin Model-7 (AMM7) [O'Dea et al., 2012] for 2016 and 2017. For brevity, we will refer to these data as to the AMM7 data below.

For the objective of present research, the sea-level data over one year is sufficient to cover some of the most important periodic variations. Therefore, we chose one-year sea surface height (SSH) data, which are from approximately 8640 hourly SSH maps, as the training dataset. In addition, it is important to note that the training SSH map (in 2016) is independent from the validation dataset. Our validation dataset (from 2017) covers a total of 3 months (2158 hourly SSH maps). This choice limits the analyses to processes with periods ranging from over-tidal to monthly. In our study area and for the time ranges defined above there are two basic processes, which explain most of the variability. These are the short-periodic tides (daily and shorter periods) and atmospherically-induced motions (e.g. due to synoptic variability in the atmosphere). As shown by Jacob and Stanev (2017) both type of motions are non-linearly coupled and their separation is not a trivial problem. In order to quantify the potential of deep learning techniques when analyzing and reconstructing SSH, we filtered signals with periods less than 48 hours by using the Butterworth filter, thus we will process two data sets: one data set containing all frequencies (briefly called AF) and the low-passed filtered data set (briefly called LF).

Example variability patterns of the AMM7 data are shown in Figure 4.2. The first two panels show the phase lines of the semidiurnal principal lunar (M2) tide (Figure 4.2a) and its amplitude (Figure 4.2b). They describe the known pattern of the dominant tidal oscillations consisting of three amphidromic areas; the Kelvin wave propagates counterclockwise [Proudman and Doodson, 1924]. The standard deviation ($s = \sqrt{1/(N-1)\sum_{i-1}^{N}(x_i - \bar{x})^2}$) between the current sea-level height xi from the LF-data set computed from AMM7 and its mean x¯ for the period from 01.01.2016 to 31.12.2016 is shown in Figure 4.2c. This panel quantifies the magnitude of low-frequency variability, which is largest in the coastal area, particularly in the German Bight. Notably, the spatial distribution of amplitudes caused by tides and wind is different. In the German Bight, the magnitude of low-frequency signal is approximately two times lower than the one of the signal associated with the M2 tide. Along the coasts of the British Isles, this ratio is larger than 5.

2.1.2 Data from the GCOAST model for the North West European Shelf

For the experiments discussed later in this study, we will need the output of another (independent) model. To this aim, we chose the numerical simulations performed in the Helmholtz-Zentrum Geesthacht (HZG) based on the Nucleus for European Modelling of the Ocean [NEMO v3.6; Madec, 2016] with 3.5 km horizontal resolution, which is two times finer than in the AMM7. The respective model setup is part of the Geesthacht COAstal model SysTem (GCOAST), which is a coupled modeling framework that includes atmospheric, oceanic, wind wave, biogeochemical and hydrological parts [Ho-Hagemann et al., 2018]. For the purposes of the present study, we use only the ocean circulation part. The model area covers the Baltic Sea, the Danish Straits, the North Sea and part of the Northeast Atlantic. The data used in the present study cover only the region shown in Figure 4.1. The vertical discretization uses 50 hybrid s-z* levels with partial cells. The model forcing for the momentum and heat fluxes is computed using bulk aerodynamic formulas and hourly data from atmospheric reanalyses of the European Centre for Medium Range Weather Forecasts (ERA5 ECMWF with a horizontal resolution of 0.25°). The tidal potential is also included in the model forcings [Egbert and Erofeeva, 2002]. The daily climatology for the river run-off is based on river discharge datasets from the German Federal Maritime and Hydrographic Agency (Bundesamt für Seeschifffahrt und Hydrographie, BSH), the Swedish Meteorological and Hydrological Institute (SMHI) and United Kingdom Meteorological Office (Met Office). The boundary conditions at the open boundaries use input from the AMM7 [O'Dea et al., 2012] distributed by the Copernicus Marine Environment and Monitoring Service. The output is stored hourly for 2016 and 2017. Data assimilation is not used.

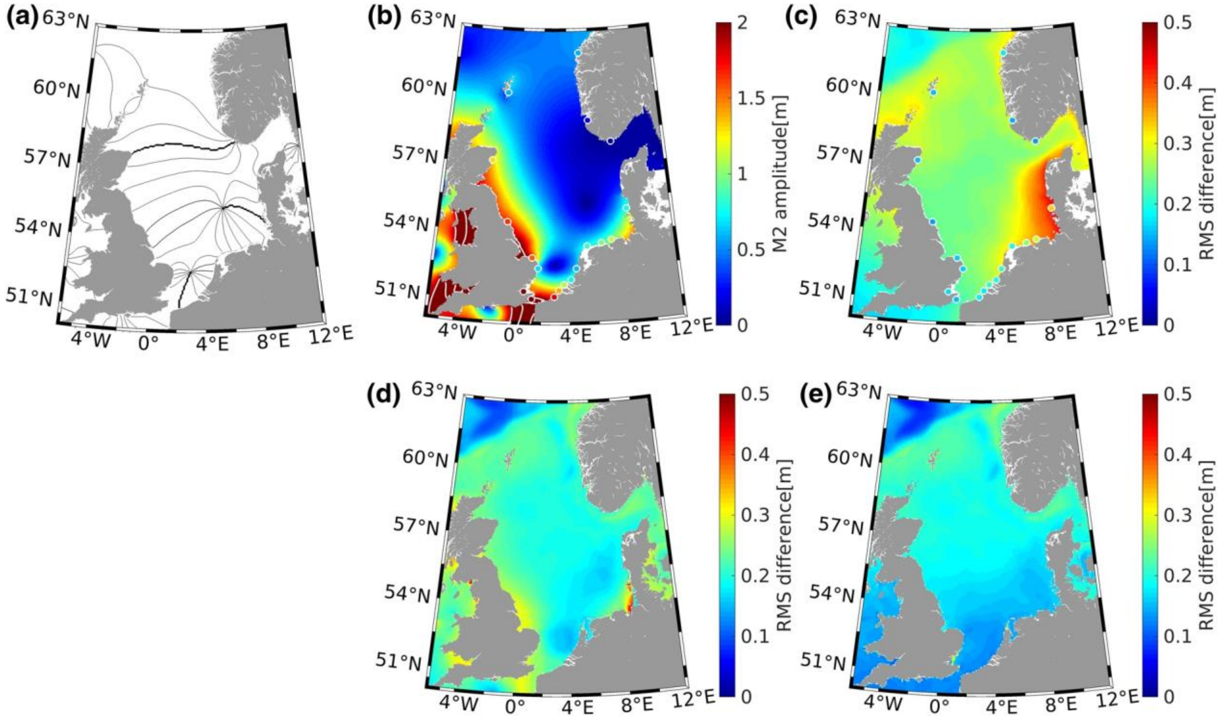

Figure 4.2. Phase lines of the M2 tide for the period computed from the AMM7 data using UTide (a). (b) is the amplitude corresponding to (a). The white isolines in (b) are lines of equal amplitude of 3, 4, and 5 m, respectively. (c) is the RMS of the LF variability of SSH. The respective values of the magnitudes of the M2 tide and the RMS variability from the TG sea level are superimposed with circles in (b) and (c). (d) and (e) are the RMS differences between the AMM7 and GCOAST models (AF, [d] and LF, [e]) shown in the AMM7 grids. AF, all frequencies (full data set); AMM7, Atlantic Margin Model-7; GCOAST, Geesthacht COAstal model SysTem; LF, low frequencies (low-pass filtered data set); RMS, root mean square; SSH, sea surface height; TG, tidal gauge.

Figures 4.2d and 4.2e show the RMSs difference between the simulations produced by the AMM7 and GCOAST models over one year for the AF and LF-data sets, respectively. Regarding the tidal signal, the differences between the two models are far below the level of variability (compare Figure 4.2d with Figure 4.2b). The largest deviations between the two models are located in the English Channel, in front of the mouth of the Elbe and around the Wash. Over most of the model area, the difference between the LF sea level in the two models is approximately two times lower than the standard deviation of the signal in each of them. This quantitative similarity between the GCOAST and AMM7 data is explained by the similar model setups, forcing and boundary conditions. The major difference between the two models, which is that the horizontal resolution in GCOAST is two times finer than in the AMM7, explains most of the differences between the two data sets.

2.1.3 Tidal gauge data

Observational data along the North Sea coast have been obtained from the Copernicus Marine Environment Monitoring Service (CMEMS, http://marine.copernicus.eu/). Altogether, 19 gauge stations with hourly resolution are used. Their positions are shown in Figure 4.1. The magnitudes of the M2 tide and the respective RMS variability of the observed sea level are superimposed with circular symbols in Figure 4.2b and Figure 4.2c to illustrate the differences between the model and observational data. Obviously, these differences, which are quantified in Table 4.1, are one order of magnitude smaller than the magnitude of the respective signals (Figure 4.2). In this table, we show the RMS deviation $RMS(P,Q) = \sqrt{\sum_i^n (P_i - O_i)^2/n}$, where P_i and O_i are the SSHs from the two datasets (the observed and the modeled SSH, respectively, or the model-1 and model-2 SSHs, respectively) at the positions of tidal gauges (TGs). In the above equation, n is the number of observations (the index is i).

Table 4.1. Quantification of differences and agreements between datasets in coastal stations.

Abbreviations: AMM7, Atlantic Margin Model-7; GCOAST, Geesthacht COAstal model SysTem; RMS, root mean square; TG, tidal gauge.

Tidal station	**RMS (TG, AMM7) (m)**	**RMS (TG, GCOAST) (m)**	**RMS (AMM7, GCOAST) (m)**
Lerwick	0.16	0.24	0.23
Aberdeen	0.17	0.20	0.23
Whitby	0.23	0.23	0.29
Cromer	0.33	0.24	0.26
Lowestoft	0.22	0.20	0.19
Sheerness	0.49	0.45	0.48
Dover	0.48	0.33	0.33
Oostende	0.22	0.22	0.31
Vlakte van de Raan	0.25	0.22	0.22
Brouwershavensegat 8	0.18	0.17	0.18
Hoek van Holland	0.15	0.15	0.14
IJgeulstroompaal 1	0.31	0.29	0.16
Terschelling Noordzee	0.25	0.22	0.17
Huibertgat	0.28	0.20	0.19
List	0.25	0.27	0.26
Tregde	0.17	0.13	0.23
Stavanger	0.15	0.13	0.20
Maloy	0.18	0.12	0.24
List	0.25	0.27	0.26

The time versus the along-coast distance diagrams (Figures 4.3a and 4.3c) give a clear illustration of the propagation characteristics of the tidal waves. Starting from Lerwick and traveling up to Whitby, the coastal wave propagates with the coast on its right (Figure 4.2a), and the slope of the contours gives a measure of the wave propagation speed, ranging between several to several tenths of ms^{-1} depending on the local conditions (the average depth of North Sea of ~ 90 m would result in a propagation speed of ~30 ms-1). At around the Wash, the propagation pattern changes dramatically because, to the south, the small amphydrome in the Southern Bight (Fig. 4.2a) wedges into the big amphydrome in the southern North Sea. This is the reason the contours undergo a rapid transition until the Terschelling Noordzee station. If we omit the data between Cromer and Terschelling Noordzee (and just linearly interpolate the data between them), the contours would present much smoother patterns. The tidal amplitudes decrease strongly at around the Tregde station (see Figure 4.1 for its position) when passing from the southern to the northern amphidromic area. Figure 4.3c is the same as Figure 4.3a; however, the data come from the AMM7. Visually, the model and observations agree quite well, and the quantitative comparison between them can be better estimated from Table 4.1.

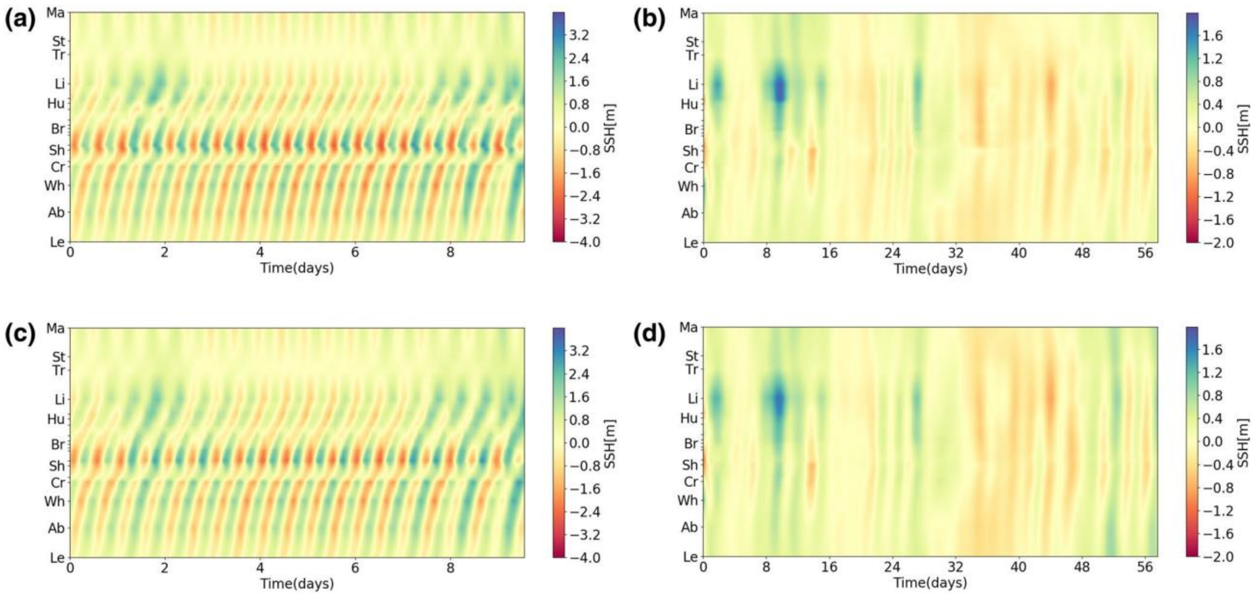

Figure 4.3. Time versus the along-coast distance (starting from the Lerwick station) diagram of the sea level from TGs (a) and AMM7 (c). The panels on the right, (b) and (d), show the same as (a) and (c) but for the LF signal and for longer periods. AMM7, Atlantic Margin Model-7; LF, low frequencies (low-pass filtered data set); TGs, tidal gauges

The LF signal (the panels in of Figure 4.3b and 4.3d) shows a temporal variability dominated by the synoptic time scales (in the atmosphere). Because of the much longer time axis compared with the panels on the left-hand side of the figure, the slope of the contours looks rather small. Along the eastern and western coasts, the propagation direction is from north to south; the change in the slope of contours occurs in the Southern Bight. Again, the consistency between the data from the TGs and the AMM7 seems quite good; all major low and high sea-level events in the observation dataset have their counterparts in the numerical simulations. The simulated amplitudes are slightly lower than the observed amplitudes, which is explained by the quality of the atmospheric forcing. The above comparison between observations and simulations shows that both datasets are similar but far from identical. The difference between the two model datasets (AMM7 and GCOAST) is comparable to the difference between each of them and the observations (see Table 4.1). As we will show in the next sections, these comparisons are important to understand the results from the experiments using machine learning (ML).

2.2 GAN

2.2.1 Brief introduction

LeCun et al. [2015] defined deep learning as a method allowing "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction". In many applications, deep learning uses feedforward neural networks, which learn to map an input dataset (in many examples, an image is used as input) to an output (e.g., more abstract information such as the probability of belonging to a certain category). Artificial neural networks (ANNs) are inspired by biological neural networks, which learn by considering examples. Their structure consists of connected units (nodes) called artificial neurons, which receive and transmit a signal to other neurons. In deep learning, multiple levels of information transformation from the previous layer to a higher layer (more abstract information) are used. Filters are applied to the input images to create feature maps that summarize the presence of those features in the input. The filter (e.g., a 3x3 matrix) is moved across the image. This movement, which is usually symmetrical in the x and y directions, is referred to as the stride. The default stride is (1, 1). A stride of (2, 2) would mean moving the filter two pixels in the horizontal and vertical directions. Thus, the neurons combine the input in such a way that the output is presented as a nonlinear combination of its inputs. A series of weights determine how the inputs are fed to the outputs. In many applications, the weight vectors are adjusted following the stochastic gradient descent (SGD) algorithm.

Goodfellow et al. [2014] introduce a framework for estimating generative models via an adversarial process by training two models. The first model is a generative model. This model captures the data distribution. The second model is a discriminative model, and its role is to estimate the probability that a sample comes from the training data rather than the generative model. As a result, the generative model recovers the training data distribution.

Normally, the structure of a convolutional neural network consists of convolutional layers followed by pooling layers. The role of the latter is to reduce the amount of redundant information. The most commonly used method is the max-pooling method, which keeps only the most active neurons (out of every 2 × 2 square of neurons in the convolutional layers, the "max"). Experience shows that this pooling step does not reduce the performance of the network. In the U-Net architecture [Ronneberger et al., 2015], the pooling operations are replaced by upsampling operators. An expansive path is developed, which is more or less symmetric to the contracting part and yields a u-shaped network architecture (Figure 4.4). In the expansive path, in every other layer, the resolution of the output is increased. Thus, in the upsampling part of the network, information is propagated to higher-resolution layers. Two distinct models, a generator and a discriminator, constitute the GAN. The generator is trained adversarially by optimizing a minimax objective together with a discriminator. In the following, the specific application of the U-Net architecture to analyzing sea-level maps is described.

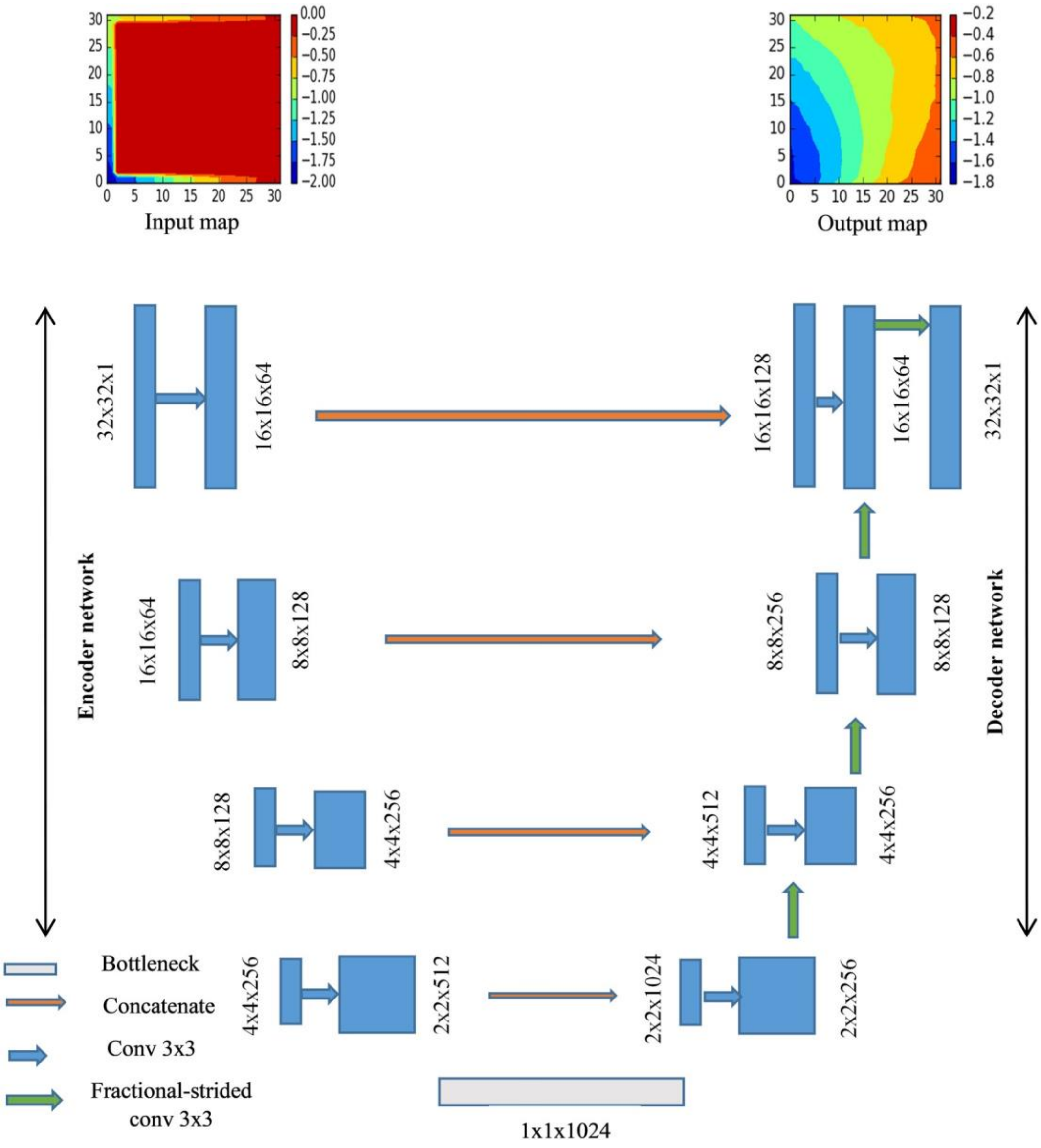

Figure 4.4. Schematic presentation of the generator part of the deep neural network model for sea surface height map reconstruction

2.2.2 GAN for tidal reconstructions

2.2.2.1 Generator

The U-Net structure of the generator part of our deep neural network model for tidal reconstruction (Figure 4.4) is illustrated in the following using the 32x32 SSH hourly maps (rectangle in Figure 4.1) for 2016 from the AMM7. In the example considered here to train the model, we use only the SSH records along the sides of rectangle as an input dataset. This dataset is named in Figure 4.4 as "Input map". The target dataset is the full AMM7 dataset (see, e.g., "Output map" in Figure 4.4). The task of the generator is to provide a model of high-quality reconstructed SSH maps (as close as possible to the AMM7 maps) by using only the information at the boundary.

The U-Net convolutional neural network (Figure 4.4) consists of two parts: an encoder (on the left) and a decoder (on the right). The encoder transforms an image (map) into a compact latent feature representation. The decoder uses that representation to produce the missing image content. Thus, the encoding-decoding process learns the image features and generates full maps.

The encoder-decoder pipeline works as follows. The encoder takes an input image with missing data and produces a latent feature representation of that image. The decoder takes this feature representation and produces the missing image content. The encoder process consists of the repeated application of 3x3 convolutions with a stride of 2 for downsampling, each followed by a batch normalization layer [Ioffe and Szegedy, 2015] and Leaky logarithmic rectified linear unit [Leaky-L_ReLU, Maas, 2015] activation.

In the example shown in Figure 4.4, the first convolution layer is a 32x32x1 (width × length × depth) map. In each subsequent convolution calculation, we obtain more latent feature maps with a larger depth index and narrower width and length. The feature map represents higher-dimensional data distribution characteristics from the image. The bottleneck layer (Figure 4.4) represents the image fully compressed into a feature map with a depth of 1024.

Decoding is the opposite of encoding; we call this process deconvolution. It consists of repeated applications of 3x3 convolutions with a stride of 2 using a transposed operator (also called a transposed convolution or fractionally strided convolution); that is, it performs a deconvolution. The upsampling layers in the original U-Net structure [Zador, 2019] are replaced with fractionally strided convolutional layers in our U-Net-like structure.

In image completion problems, corrupted images and output images share a certain amount of low-level features, such as prominent information from the noncorrupted regions, luminance and resolution. However, deep network-based methods with bottleneck layers may lose details of images when propagating feature maps in the training stage. Moreover, these methods may suffer from the vanishing gradient problem as the network deepens. To shuttle the image information through the networks and reduce the training burden, we apply the skip connections strategy [Mao et al., 2016].

2.2.2.2 Discriminator

The discriminator (Figure 4.5) is used to determine the possibility that the prediction map comes from the training set (i.e., whether it is a real training image) or the prediction set (i.e., whether it is fake image from the generator). During training, increasing better fake images are generated, and the role of the discriminator is to correctly classify the real and fake images. When the generated prediction map is consistent with the ground truth of the image content (we will call this the target for short) and the GAN discriminator cannot determine whether the prediction map is from the training set or the prediction set, then the network model parameters are considered to have reached the optimal state.

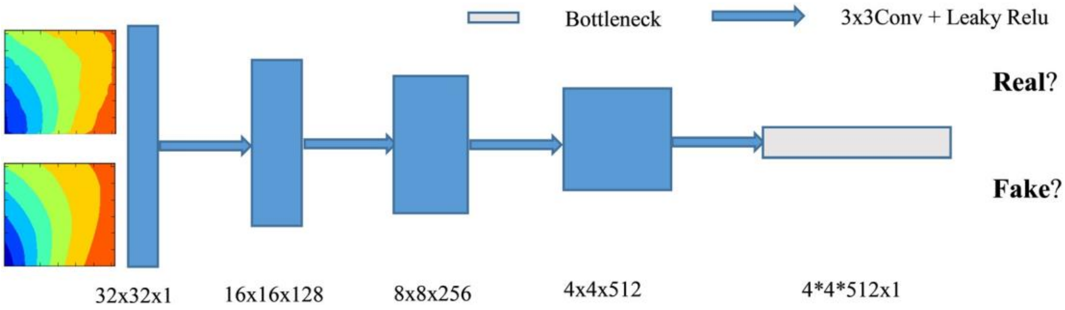

Figure 4.5. The discriminator part of the reconstruction model

The discriminator can be understood as the inverse of the generator with five 3x3 convolutions with stride of 2 layers, where the last convolutional layer is fed into a single sigmoid activation function. The L_ReLU activation function is used for all the layers in the discriminator except for the output. Following GAN technology, the generator is trained adversarially against a discriminator, which is simultaneously trained with the generator.

2.2.2.3 Technical details

We use 8640 SSH maps with a size of 32x32 to train the GAN model. This dataset is too large to be passed to the computer at once; therefore, we divide the data into smaller sizes. Two hyperparameters are defined: the number of training epochs (how many times we train the model) and the batch size (the number of samples used to train the model in one epoch). During the training stage, mini-batch learning is introduced [Cotter et al., 2011]. This method divides the data into several small batches and updates the parameters in batches. Thus, a set of data in a batch determines the direction of the gradient, which is more stable and converges faster. In the present tidal application, 12 is selected as the batch size, which corresponds to the period of the M2 tide in the studied region [see also Riley, 2019]. The generator model uses only the observations in tidal gauges locations to generate the entire 2D SSH maps. This generated SSH map is feed into Discriminator together with the SSH map from the numerical model (true data set) for checking whether the generated SSH resembles the true one (see Annex 1).

The model epoch parameter, which is a number optimizing the gradient decent (to avoid overfitting or underfitting), is set to 60. The initial learning rate of the Adam optimizer of the GAN model is set to 0.00003. The learning rate determines the step size of gradient descent (i.e., how fast the model converges). Too large a rate may cause the parameters to move back and forth on both sides of the optimal value. Too small a rate will greatly reduce the optimization speed. To solve the problem, we introduce an exponential decay method from the TensorFlow framework [Loshchilov and Hutter, 2019]. The learning rate was gradually reduced to make the model more stable in the later stages of training. The number of convolution levels is set to 9 (later in the text, we explain why this number must be changed in other experiments).

In this model, the discriminator loss is the same as the basic deep convolutional GAN (DCGAN) model [Radford et al., 2015], while for the generator loss, we introduce (on the basis of original generative loss) the least square errors (known as the L2 loss function) as the consistency content loss into the generative loss function. The basic loss function of the GAN model meets the Nash equilibrium condition as much as possible [Osborne and Rubinstein, 1994]. The principle behind this equilibrium is based on game theory and aims at continuously optimizing the generator and discriminator so that the generated data approach the real data [Dong and Yiang, 2018]. In this way, the GAN makes the samples generated by the generator approach the real sample, in terms of both authenticity and diversity. To adapt the technology to our specific study and obtain more accurate results, we also established a new loss function combination for our pixelwise GAN model by adding the pixelwise reconstruction loss generation part based on the basic loss function [Zhao et al., 2017]. The equations describing the loss functions of the generator and discriminator are given in Annex 1.

After 60 training epochs, we obtain a suitable generator structure that remembers the SSH high-dimensional features of the selected ocean region. To validate the generator model, the discriminator part is dismissed, since the generator part is the main structure for reconstructing the completed SSH maps. Now, the validation dataset (2158 hourly, incomplete SSH maps) is fed into the generator part of the neural network model to generate feasible SSH maps.

2.2.3 GAN for LF SSH reconstruction

As shown in section 2.1.1, the amplitude of the remaining signal is lower than that of the dominant partial tides. Furthermore, the variability is less regular because meteorological drivers such as wind, storms or atmospheric pressure have a certain level of randomness. Therefore, we introduce some changes in the GAN model described in section 2.2.2 for the sake of obtaining more accurate reconstruction results [see Zador, 2019].

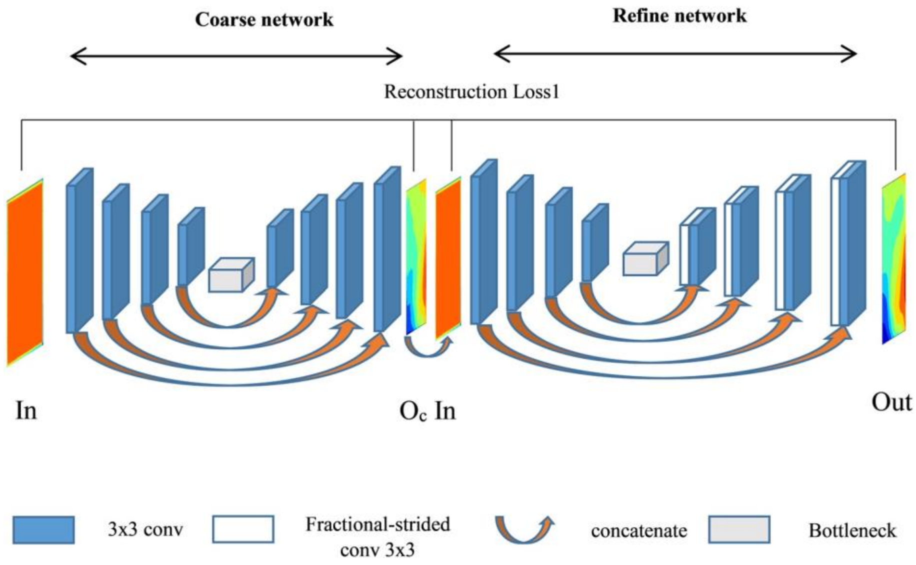

Figure 4.6. Generator structure for residual sea surface height map reconstruction

Our new model for LF SSH reconstruction consists of two steps: coarse and fine reconstruction (Figure 4.6). This architecture helps to stabilize training and enlarge the receptive fields, as mentioned by Yu et al. [2018]. Figure 4.6 represents a 32x32x1 map with real data only at the boundary, and O_c represents the coarse-reconstructed SSH map. The refinement step takes the O_c and In maps together as input pairs to output the final result Out. Thus, O_c conditioned on In is selected as the input of the refined network that reconstructs the complete SSH map called Out. This type of input stacks information of the known areas to urge the network to capture valid features faster [Liu et al., 2019], which is critical for rebuilding the content of missing regions. Our refined structure also consists of an encoder and decoder, where a skip connection is adopted, similar to a coarse network. In the encoder, each of the layers is composed of a 3×3 convolution, while in the decoder, a fractional stride convolutional layer with stride of 2 is adopted together with a 3×3 convolution. Finally, the discriminator has the same structure as the discriminator in our tidal reconstruction model.

The training step and reconstruction process are the same as in the tidal reconstruction model. However, the training sample batch size is set to 336 (two weeks of hourly data). The number of training epochs is set to 5000, since more uncertainty and greater magnitudes of the variations lead to training difficulties, which will require more training epochs to obtain a stable and desirable model. The initial learning rate of the Adam optimizer in this GAN model is set to 0.00007 to adapt to these model parameter changes.

2.3 Kalman Filter Approach

Schulz-Stellenfleth and Stanev (2010) proposed an optimal linear estimator to reconstruct ocean state parameters from observations knowing the prior distribution of the state and measurement errors. The method is similar to the approach of Frolov et al. (2008) and uses standard concepts of estimation theory.

It will be very briefly presented below; for more detail, the interested reader is referred to Schulz-Stellenfleth and Stanev (2010) and Grayek et al. (2011). The method uses the background covariance matrix derived from the AMM7 data as a priori information. We will denote the global state vector of dimension m by $\boldsymbol{x}$. The data from 19 tidal gauges represent the measurement vector $\boldsymbol{y}$ of dimension n. The global state vector $\boldsymbol{x}$ contains SSH from AMM7 data at the individual position of the model area. The task is to find a reconstruction matrix A such that:

$$J(A) = \sum_{j=1}^{q} \left\| x(t_j) - Ay(t_j) \right\|^2 \tag{4.1}$$

is minimum, where q is the number of SSH maps (hourly maps in 1 year). This would ensure that the reconstruction error is as small as possible. Assume that the observations can be derived from the global states according to

$$\boldsymbol{y} = \boldsymbol{Hx} \tag{4.2}$$

where H is the linear measurement operator. Schulz-Stellenfleth and Stanev (2010) showed that $J(A)$ is minimum if A is the Kalman gain matrix

$$A = PH^T(HPH^T + R^{-1}) \tag{4.3}$$

where P is the background covariance matrix for the state $\boldsymbol{x}$ and R is the observation error.

It was demonstrated in the same study that if the dynamics of the state variables can be described by only a few empirical othogonal functions (EOFs), the dimension of the reconstruction problem can be significantly reduced. For the AMM7 SSH data set, only three EOFs describe more than 95% of the variance. Therefore, in the analyses addressed in the following, we used three EOFs only. R is taken as a diagonal matrix, with a constant error value of 1 cm.

3 Experiments

3.1 Experiments in reduced area

The first group of experiments presented below aims first at analyzing how appropriate the GAN is to reconstruct the sea level in a relatively small area (only 32x32 grids) in the interior of the North Sea. In the experiments' nomenclature (Table 4.2), we use the abbreviations AF and LF. In the AF experiments, we use the one-hour data, as they are produced by the AMM7 model. In the LF experiments, we use also one-hour data, but the variability with periods higher than two days is removed by low-pass filtering as explained above. Therefore, in the LF experiments, the sea level can be considered mainly driven by the atmosphere and by low frequency tides (e.g., spring-neap variability). The training phase uses 8642 hourly maps, which corresponds to one year. In the first type of experiments, for which there is a column called "Training" in

Table 4.2, we use data from the same source in the training and validation step. In this way, the data at the two steps are consistent with each other.

Table 4.2. List of experiments in a small domain
Abbreviations: AF (all frequencies, full dataset), LF (low frequencies, low-pass filtered dataset)

Name	Type of experiment (AF)	Type of experiment (LF)	Training (input data-target data)	Validation (input data-validation data)	Comment
AF1	X	-	AMM7-AMM7	AMM7-AMM7	Randomly distributed no-data locations.
AF2	X	-	-	-	At the validation step, there are missing data only in the interior.
AF3	X	-	-	-	At the validation step, data are available only at the boundary.
AFK	X	-	-	-	-
LF	-	X	-	-	-
LFK	-	X	-	-	-

In AF1-3, we determine the quality of reconstruction if some input data are missing.(Here we use 0 to replace the missing data value, 0 does not have specific physical meaning here.) In AFTRA1, we randomly generate locations in the 32x32 matrix, where we assume that there are no available data (there are no “observations” in half of the locations in the original grid). Thus, we feed the GAN model with data only from the locations where there are “observations”. We use the data from all the locations as a target dataset. At the validation step, we use the “observations” in only half of the locations of the original grid to reconstruct the 32x32 field (all locations) over a period of 3 months. A comparison between the AMM7 data and the reconstructed data will be analyzed in section 4.

In AF2, we select a square area in the middle of the 32x32 matrix (i=5,…, 25, and j=5,…, 25), which is considered a no-data area. The basic difference from AF1 is that this no-data area is compact. In AF3, we extend the no-data area up to the boundary. This exercise can thus be interpreted as a reconstruction of the full dataset using data only at the boundaries (all the boundary locations).

Experiment LF is essentially the same as AF3; that is, only data at the boundary are used to train the model. The difference is that LF analyses the capability of using a GAN to reconstruct the data set from which the high- frequency tides have been removed. In all the experiments described above, the computational resources were relatively low. On one GPU node, which is a Nvidia Tesla V100 with 32 GB memory, it takes ~30 minutes to complete the training in the AF experiments and ~45 minutes to complete the training in the LF experiments. The latter takes a longer time than the former because the more stochastic signals associated with the atmospheric forcing compared to the periodic tidal signals make the convergence slower.

3.2 Experiments in the entire North Sea

The second group of experiments are for the entire North Sea basin (index "B" in Table 4.3). Experiment BAF1 is essentially the same as AF3. The difference is that in the training phase, we use only data from 19 locations where TGs operate. These data in BAF1 are taken from the AMM7. The training and validation periods are the same as those in the experiments with reduced area. BLF1 is essentially the same as BAF1; the difference is that we analyze the quality of the reconstruction of the LF-North Sea data set, that is the data used in this experiment are the low-pass-filtered data used in BAF1. The idea to carry out this experiment was two-fold. It was assumed that removing the high-frequency oscillations would result in a better model when reconstructing the low-frequency variability of basin-wide SSH using only coastal data. The second consideration was that the high-frequency oscillations are not included in some altimeter products; thus, it is worth trying to test whether ML can well resolve only the low-frequency variability.

BAF2 is the same as BAF1; however, real observations from TGs are used in the validation step along with the same model developed in BAF1. Obviously, this experiment uses data of different origins (in the "Validation" column of Table 4.3, the data sources are different). Thus, these data are not fully consistent with each other. In the following, we will refer to this type of experiment as experiments with "inconsistent data". BLF2 is the same as BAF2, but BLF2 addresses the quality of the reconstruction of low-frequency variability. One important difference between the AF and LF experiments is that we use different ML models (see section 2) because of the different spatiotemporal characteristics of the tidally and atmospherically driven sea level. In BAF-G and BLF2-G, we do not use real observations as in BAF2 and BLF2 but rather data from the GCOAST model in the observation locations.

The next two experiments, BAF3 and BLF3, use partially inconsistent data for training and validation. By "partially inconsistent", we mean the following. At the training step, at the positions of the TGs, we use data from the TGs. The target dataset is the basin-wide SSH, which is produced using the AMM7 (TG-AMM7 in the "Training" column of Table 4.3). It is expected that the ML model learns the consistency between the forcing data and the target. Therefore, the product is partially consistent with the data from the TGs and the AMM7. At the validation step, we use tidal gauge data and the ML model to reconstruct the

SSHs and compare them to the AMM7 data. BLF3 is the same as BAF3, but BLF3 addresses the quality of the low-frequency reconstructions.

In the final two experiments (BAF3-G and BLF3-G), we use the same approach as in BAF3 and BLF3. In this case, in the observation locations (Figure 4.1), the data are sampled from the GCOAST model, which has a horizontal resolution two times better than that of the AMM7. These data are considered pseudo-observations with "different quality" than the quality of the coarser AMM7.

Table 4.3. List of experiments in the entire North Sea
Abbreviations: B (basin-wide), AF (all frequencies, full dataset), LF (low frequencies, low-pass filtered dataset)

Name	Type of experiment (AF)	Type of experiment (LF)	Area	Training (input data-target data)	Validation (input data-validation data)	Comment
BAF1	X	-	Entire North Sea	-	-	Similar to the comment for AF (see Table 4.2).
BLF1	-	X	-	-	-	Similar to the comment for LF (see Table 4.2).
BAF2	X	-	-	-	TG-AMM7	Similar to the comment for BAF1 for the observations used at the validation step.
BLF2	-	X	-	-	-	Similar to the comment for BLF1 observations used at the validation step.
BAF2-G	X	-	-	-	GCOAST-AMM7	Similar to the comment for BAF2 GCOAST

						data used at the observation locations at the validation step.
BLF2-G	-	X	-	-	-	Similar to the comment for BLF2 GCOAST data used at the observation locations at the validation step.
BAF3	X	-	-	TG-AMM7	TG-AMM7	The data at the boundary are the same in the training and validation steps.
BLF3	-	X	-	-	-	-
BAF3-G	X	-	-	GCOAST-AMM7	GCOAST-AMM7	-
BLF3-G	-	X	-	-	-	-

4 Results

4.1 Sea-level reconstruction in idealized (reduced) areas

The results of all the reduced area experiments are presented in Figure 4.7. As representative characteristics measuring the agreement between the reconstruction data and the observations, we use the index of agreement [Willmott, 1981]:

$$D(P,Q) = 1 - \sum_{i=1}^{n} (P_i - O_i)^2 \Big/ \sum_{i=1}^{n} (|P_i - \bar{O}| + |O_i - \bar{O}|)^2 \qquad (4.4)$$

In the above equation, the overbar indicates the temporal mean, and the other notations are explained in section 2.1.3. The above equation provides a statistical approach to compare model predictions (*P)* with observations (*O*). The numerator measures the average error magnitude and the denominator gives a basis of comparison. The index of agreement measures the model performance as the degree to which *P* matches *O*, where 1 indicates perfect agreement and 0, complete disagreement. Other possible indices for model-data comparison are defined by Nash and Sutcliffe [1970], Legates and McCabe [1999]; see also Willmott et al. [2012]. The results of the three AF experiments (AF1-AF3) are shown in Figure 4.7a-4.7c. They illustrate how the reconstruction results deteriorate if some input data are missing. However, “deterioration” is not an adequate word in this case because all three reconstructions are characterized by index of agreement greater than 0.99. The pattern of *D* reflects some characteristics of the data distribution and dynamics. In AF1, the no-data locations are randomly distributed. However, the results in Figure 4.7a show that the lowest values of the index of agreement appear predominantly in the coastal areas. This finding is explained by the fact that, in the basin interior, the no-data locations are uniformly surrounded by locations where observations are available. However, at the periphery of the studied area, the no-data locations are surrounded by fewer observations (because no observations exist outside of the area).

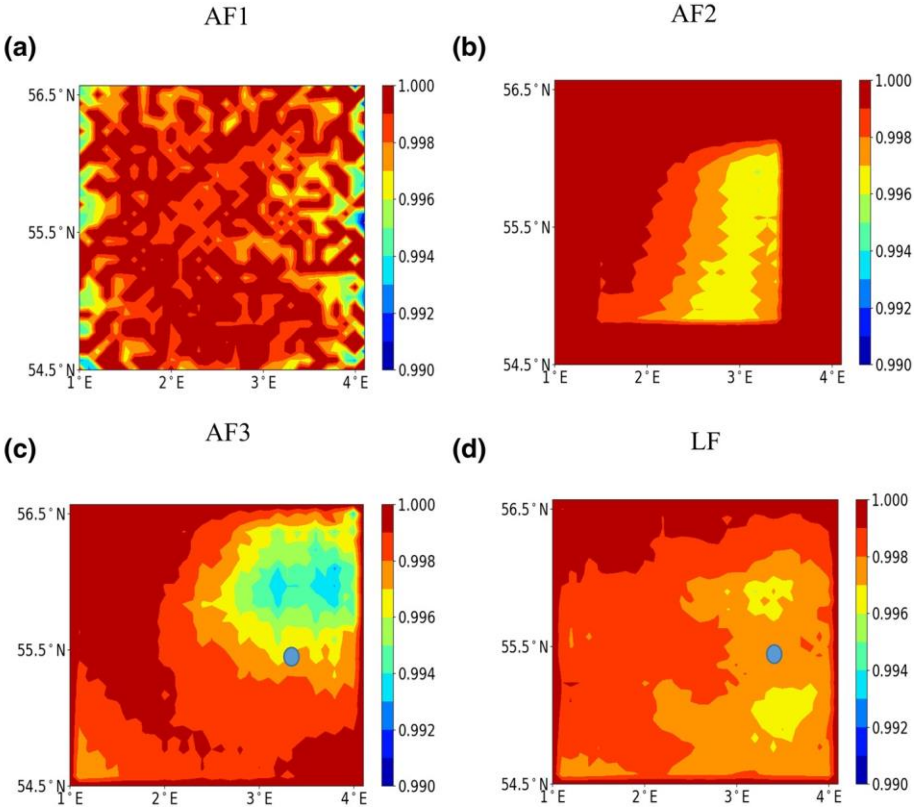

Figure 4.7. Index of agreement between the “true” and reconstructed SSHs in the experiments carried out in the reduced areas. The index was computed at the validation step. The numbers on the axes are the longitude and latitude (see Figure 4.1 for the position of this area). The blue dots are locations where time series are analyzed for the validation period. AF, all frequencies (full data set); LF, low frequencies (low-pass filtered data set); SSH, sea surface height

In AF2, where we prescribe a wide coastal area with data, the index of agreement is higher than 0.995. In the no-data area (in the middle of Figure 4.7b), the index of agreement shows a propagation pattern, which is in agreement with the propagation direction of the Kelvin wave (see the rectangle in Figure 4.1 and the phase lines in Figure 4.2a). The situation in AF3 (Figure 4.7c) is qualitatively similar to that in AF2 (a better agreement with the validation data set in the western part). However, in this experiment, only the data along the that in AF2. The lowest D∼0.994 in AF3 appears in the area closest to the amphydromic point (Figure 4.2a), boundary are used; therefore, the index of agreement is slightly lower, and its pattern is less regular than where the amplitude of the signal is lower; therefore, the signal-to-noise level is also lower.

Experiment LF (Figure 4.7d), which quantifies the capability of GAN to reconstruct the SSH using the LF data set, shows a comparable skill as AF1–AF3. In all four cases, the index of agreement is above 0.99, and its ranges are comparable. The fundamental difference between the AF and LF experiments is the pattern of the index of agreement, which is no longer tidally dominant in the LF case. for one location shown in Figure 4.7 where the reconstruction quality of GAN is relatively low (D∼0.997). The comparison between the performance of GAN and Kalman filter approach is presented in Figure 4.8 the AF reconstructions and AMM7 data are ∼3 cm. The RMS differences between the LF experiments and Obviously, the two methods perform very similarly. For this specific location, the RMS differences between low-pass filtered SSH are ∼1 cm. These values, as seen in Figure 4.8, are negligibly smaller than the amplitude of the respective signals. It is clearly seen from these illustrations, in particular in the plots on the bottom,that the largest differences between the reconstructions using GAN and Kalman filter methods, from one side, and data, from the other, occur almost at the same times.

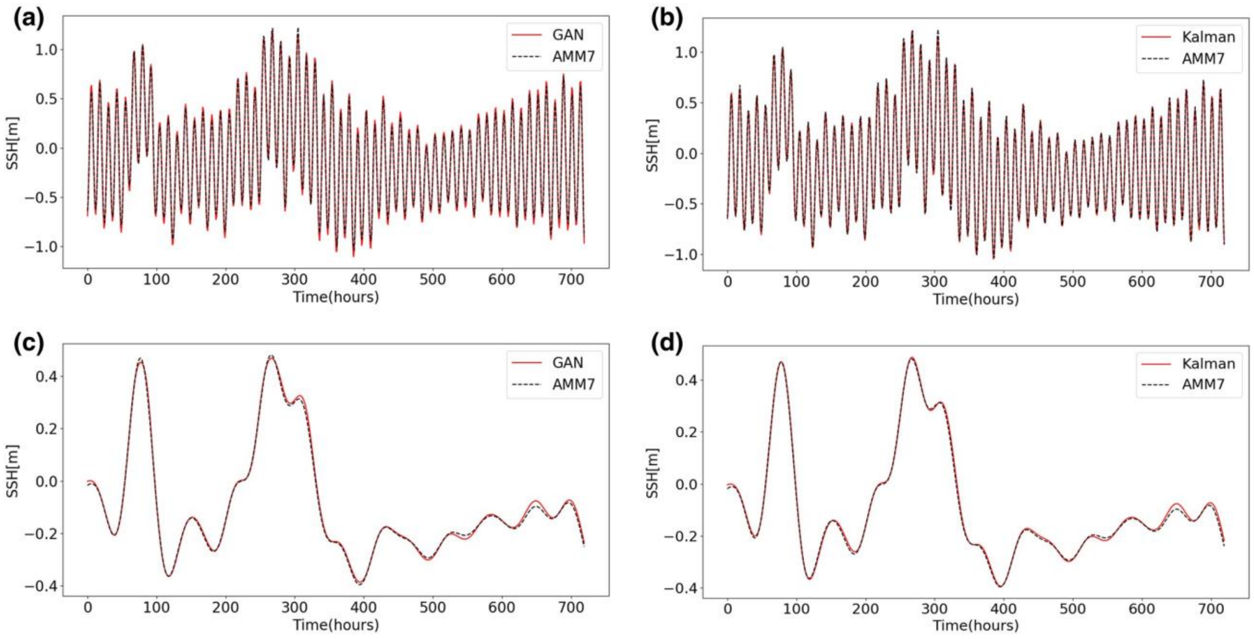

Figure 4.8. Sea level in the locations shown in Figure 4.7 for the 2-month validation period. (a and b) are AF experiments; (c and d) are LF experiments. Panels on the left are from GAN reconstructions; those on the right are from the reconstruction using the Kalman filter approach. AF, all frequencies (full data set); AMM7, Atlantic Margin Model-7; GAN, generative adversarial network; LF, low frequencies (low-pass filtered data set)

The above experiments are relatively easy, at least in terms of the volume of data used. In real-world applithe same performance if larger datasets were used. By increasing the data volume by ∼25 times, one reaches cations, data sets are usually much larger, and it was not clear a priori whether the used method would have paratory experiment in which we interpolated the 32 × 32 matrices with a resolution five times better than the volume of the data set generated from AMM7 over the entire North Sea. Therefore, we performed a prein the AF experiments and repeated AF3 experiments using the new data set. Because of the increase in initial learning rate is the same as in the case of the 32 × 32 data set. The computational time for training × the data size, we increase the number of convolutional layers to 13; the number of epochs is set to 600. The increased up to ∼6 h, which is approximately 12 times longer compared to the 32 x 32 cases. The index of agreement in this additional experiment (not shown here) is higher than 0.993, and its pattern is close to that in AF3.

4.2 Sea-level reconstruction over the entire North Sea

Here, we discuss the skill of experiments introduced in Section 3.2. As a measure of the skill of each of them, we will show maps of the index of agreement (Figure 4.9). The RMS difference between the reconstructed and “true” data is shown in the supporting material (Figure 4.S1). The results from experiment

BAF1 (Figure 4.9a) demonstrate that using the data from only 19 locations where the TGs operate is sufficient for adequate sea-level reconstruction over most of the analyzed domain. Only in the northwestern part of the points in the eastern North Sea, the index of agreement drops to ∼0.8. The smaller D in the area between study area and in Kattegat, where TGs are not available, and in the area between the two amphidromic points is explained by the low-amplitude tides in this zone (small denominator in Equation 1).

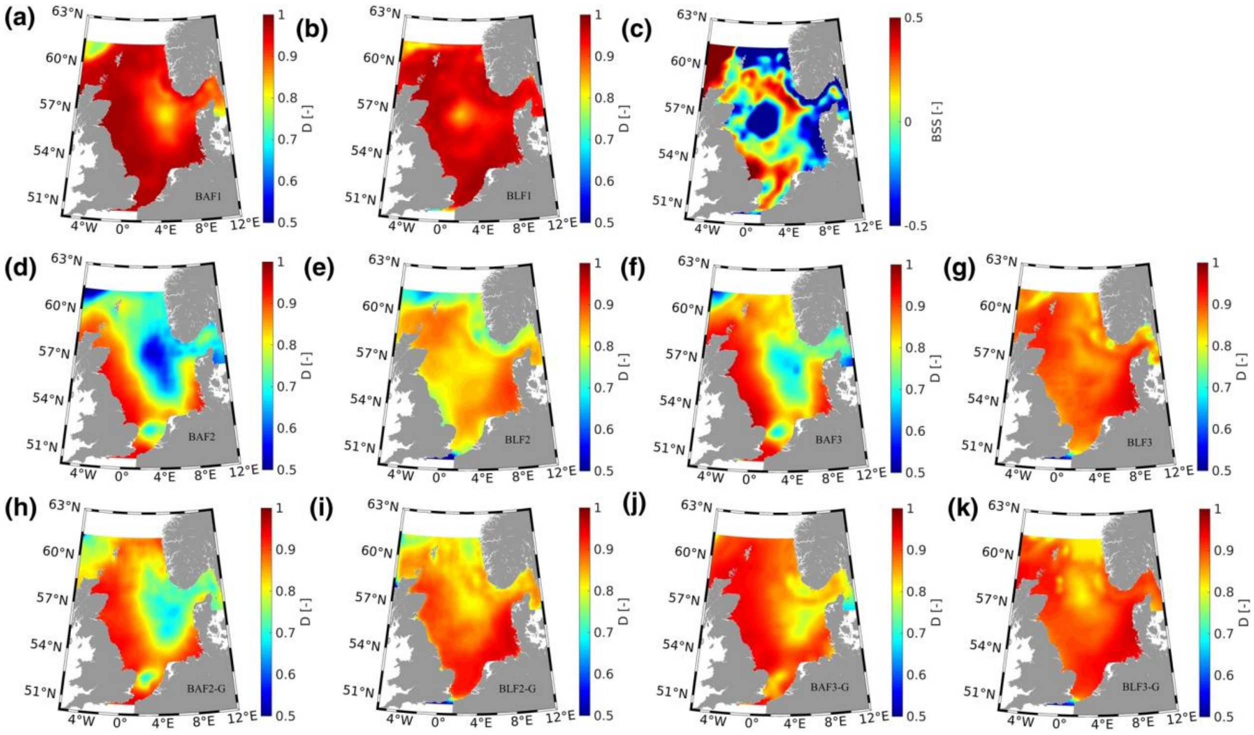

Figure 4.9. Index of agreement (see Equation 1) between the "true" and reconstructed SSHs in the experiments carried out over the entire North Sea (see Table 4.3). The training data set is from January 01, 2016 to January 01, 2017, and the validation data set is from January 01, 2017 to March 01, 2017. The names of the individual experiments are shown in each panel. The ML model BAF1 is used in all the experiments shown in the first column (a, d, and h). The panels in the second column (b, e, and i) use the BLF1 model. Panels f and j and panels g and k use the BAF3 and BLF3 models, respectively. (c) shows the Brier skill score (see Equation 2) of BLF1 against the low-pass output of BAF1. ML, machine learning

The reconstruction of the LF variability (experiment BLF1, see Figure 4.9b) is approximately as good as the reconstruction of the full signal. The lower index of agreement in the interior of the North Sea is explained by the relatively low amplitude of the signal there (compare with Figure 4.2c). Because of this phenomenon, the signal-to-noise ratio reduces the reconstruction skill. Overall, BAF1 and BLF1 demonstrate that if consistent datasets are used in the locations where TGs are located, the GAN model adequately reconstructs the basin-wide SSH.

We remind the reader here that the SSH reconstructed in BAF1 contains low- and high-frequency signals, while the output of the BLF1 experiment reproduces only the low-frequency variability. The initial expectation was that the GAN could better learn less complicated temporal and spatial variability, which is in the case when a high-frequency signal was removed from the data. Figure 4.9c shows the Brier skill score

$$BSS = 1 - BS_{SFA1i}/BS_{TFA1} \tag{4.5}$$

where $BS(P,Q) = 1/n \sum_i^n (P_i - O_i)^2$ is the mean-squared error in each experiment and BAF1 is taken as the reference experiment. The reconstruction is perfect when the BSS is equal to 1. BSS= 0 means that there is no improvement in BLF1 compared to the results in BAF1. If BSS<0, the quality of BLF1 reconstruction is poorer than that in BAF1. Obviously, there are areas where BAF1 shows better agreement with the observations than BLF1. This result suggests that processing a much more complex data set (BAF1) is superior in many areas than processing low-pass filtered data, which demonstrates that the GAN can learn about processes with multiple time scales. As demonstrated by Jacob and Stanev (2017), in the North Sea, processes with multiple time scales in the ranges studied here are nonlinearly coupled. The fact that the reconstruction of the full signal is superior in many areas compared with the reconstruction of the LF signal provides indirect proof that the nonlinear interactions between processes with different time scales are well captured by ML. These patterns would not occur if nonlinear interactions between processes with different time scales did not exist. Figure 4.S2 gives an illustration how M4 tides, which are due to nonlinear advection, are replicated. This result emphasizes the performance of ML method in the coastal regions.

In the BAF2 and BLF2 experiments, the reconstruction models are the same as in BAF1 and BLF1, respectively. However, unlike the BAF1 and BLF1 experiments, where the training and validation steps use consistent data, the BAF2 and BLF2 experiments belong to the class of experiments using inconsistent datasets at the validation step; that is, instead of using AMM7 data at the coast (consistent with the training data), we feed the model with real observations (which are inconsistent with the model). The level of inconsistency is quantified in Section 3 (see Figure 4.3 and Table 4.1). This substitution of data at the validation step resulted in a reduction in the reconstruction quality. What the GAN model can only adequately capture (index of agreement above 0.85) is the sea-level variability in the coastal areas of the western and southern North Sea (Figure 4.9d). The areas of low sea-level variability show a very low reconstruction skill (compare with Figure 4.2b, particularly the region of the small amphydrome in the Southern Bight). The reconstruction of the LF-signal is slightly better, particularly in the coastal zone of the German Bight, where the variability range of the LF-signal is the strongest (compare Figure 4.9e with Figure 4.2c). Obviously, the GAN model is not very flexible in using arbitrary types of data at the reconstruction step. The reduction in the reconstruction skill reminds us of the problems in data assimilation when errors in the data and model are not treated appropriately.

The BAF2-G and BLF2-G experiments, similar to the BAF2 and BLF2 experiments, belong to the group of experiments using inconsistent data sets at the validation step. In this case, the ML models are the same as in BAF1 and BLF1; however, the GCOAST data in the locations of TGs are used at the validation step. As shown in Section 2, the GCOAST data are slightly more consistent with the AMM7 data than with the TG data (Figure 4.3, Table 4.1). Therefore, the reconstruction skill improved in comparison to that in BAF2 and BLF2 (compare Figures 4.9h and 4.9i with Figures 4.9d and 4.9e, respectively). However, D is much lower than in the BAF1 and BLF1 experiments.

The “partial inconsistency” of the data for training and validation in BAF3 and BLF3 implies that, at the training step, the processing of the GCOAST data (at the coast) and the AMM7 data (as the target) tends to decrease the inconsistency between the two datasets in the GAN model. Thus, the results of new models, which are different from BAF1 and BLF1, respectively, are documented by the comparison between Figures 4.9f and 4.9g and Figures 4.9d and 4.9e, respectively. Obviously, the GAN model learns to adapt the solution to the data from different origins. The model skill is also dependent on the magnitude of the sea-level variability (compare with Figure 4.2b), which also explains the relatively good skill of BLF3 in the coastal zone of the German Bight. Notable is the more uniform and better reconstruction skill in the LF experiment than in the AF experiment over most of the area.

The final two experiments, also belonging to the group of experiments with “partially inconsistent” data (BAF3-G and BLF3-G), illustrate the improvement of the reconstruction skill if synthetic observations (the GCOAST data in the TG locations) are used in the training (compare Figures 4.9j and 4.9k with Figures 4.9h and 4.9i, respectively). The better agreement between the GCOAST and AMM7 data than the agreement of each of them with the real observations (see Table 4.1) explains why the index of agreement in Figures 4.9j and 4.9k are better than those in Figures 4.9f and 4.9g, respectively.

5 Discussion

One year of training data appeared sufficient for a model using a generative adversarial network to learn the structure of the SSH data and to adequately reconstruct the basin-wide SSH during the validation period using data from only 19 locations along the coast. The quality of the reconstructions was almost equally good for the full signal (AF) and low frequency one (LF). The high values of the index of agreement (area mean values of 0.937 and 0.945 for the BAF1 and BLF1 experiments, respectively) were possible provided data from the same source was used (in this case, the AMM7). Another reason for the reconstruction success is the relative smoothness of the SSH maps.

The use of the same model fed from either the observations (TGs) or independent numerical simulations (GCOAST data) reduced the quality of the reconstructions: 0.761 and 0.822 in BAF2 and BLF2, respectively, and 0.823 and 0.860 in BAF2-G and BLF2-G, respectively. The comparable numbers are explained by the comparable differences between the three data sets fed to the model. In the BAF experiments, the agreement between the reconstructions and validation data depends strongly on the patterns of the tidal amplitude: the lower the amplitude (in amphidromic points), the lower the agreement is. The BLF experiments show a much more uniform distribution of the index of agreement. In both the BAF and BLF experiments, the reconstruction model performs better if it is fed with the GCOAST data than when it is fed with actual observations.

Some drawbacks in the reconstruction could be avoided if data from the coastal stations are used in the training. In BAF3, BLF3, BAF3-G, and BLF3-G, the basin mean indexes of agreement are 0.823, 0.879, 0.882, and 0.889, respectively. Obviously, there are good perspectives by developing an optimal learning process to improve the reconstruction quality, which should include the study of the individual imprint of stations for the reconstruction of the basin-wide sea level. TGs are sometimes placed in locations that are not representative of the large-scale dynamics; therefore, the observed signal is not fully consistent with the basin-wide dynamics. Such stations would have low imprints but could also contaminate the learning process.

A further adjustment of the loss function or other parameters would improve the reconstruction quality, which is another technical task to solve in future research. This issue has not been addressed in the present study because our aim was to demonstrate the power of the GAN model in reconstructing SSH maps by using different types of inputs and targets.

Another issue that has not been discussed in the present study is the length of the data series that we use. As is well known, neural networks can reconstruct situations similar to those they encounter from the past. Therefore, another way to improve the reconstruction quality would be to extend the duration of the learning process and perhaps to set a clearer aim to the reconstruction exercise with respect to the time scales addressed.

One important question to discuss here is what we learn about physics. One zero-order answer would be “nothing” because what we see in the validation step is a synthesis of situations from the past. However, the basic message from Figures 4.9a and 4.9b is that a decent reconstruction capability is realized using a relatively short time series, which is an illustration of a substantial recurrence of patterns. This would imply that the spatial-temporal patterns repeat (quasi)periodically, and a relatively short-time record contains the most representative characteristics of SSH dynamics. While this was clear for the tides, it was not so obvious about changes in sea level caused by the atmosphere. Furthermore, the GAN has a good skill to learn and reconstruct dynamics with multiple time scales. The results presented in Figure 4.7a illustrate that the reconstruction capabilities of the model decrease when approaching the boundaries of the area addressed. This finding justifies that having data at the boundaries is an important prerequisite for optimal reconstructions. In other words, much information on the dynamics of the entire basin is encapsulated in the boundary data, which enables the good reconstruction skill of the specific GAN application.

The patterns in Figures 4.7 and 4.9 demonstrate that the errors in the reconstructions are closely linked to specific physical patterns; that is, one can also make useful analyses of the model errors to study the physical properties of the sea level. This analysis would be important when developing concepts to specify error covariance matrices in data assimilation models. Another aspect concerns the role of coasts, which constrain the circulation features in different ways. One example is the low index of agreement in the area of the Norwegian trench, which is a known challenge for numerical models. Evaluation of different types of nonlinearities and identification and quantification of the responsible processes with the help of ML is another issue of future development. Our preliminary analyses of other types of data (e.g., sea surface temperature and sea surface salinity in the German Bight) show that in some cases Kalman filter approach performs better, in some other cases, for example, reconstruction of sea surface salinity, it is the ML approach, which performs better. The studied here SSH maps is just one type of data with their respective temporal and spatial scales. They cannot be considered as a comprehensive set of different types of data with different spatial and temporal characteristics, as well as different level of stochasticity. Therefore, the experiments presented here do not allow to fully analyze the advantages and disadvantages of two methods. A deeper analysis of the performance of the ML and Kalman filter approaches when using different and more challenging data sets will be presented in a forthcoming study.

Longer periods are beyond the scope of the present research. Reconstructing basin-wide SSH over long times would require a different design of deep learning. One can expect that reducing the resolution of basin-wide data used in the training, both in time and space, would allow more efficient computations and extension of the addressed time scales to decadal and beyond. One fundamental issue to address is whether coarser resolution in space and time ML would ensure adequate decadal reconstructions. Such an exercise will be analyzed in a separate study using different data sampling and processing technologies.

Another natural extension of the present research would be the application of a GAN to data-only cases. One candidate is the amalgamation between satellite altimetry and TGs, which would open up the perspectives to improve and optimally use the observational networks in the North Sea.

6 Conclusion

The method proposed here to reconstruct the basin-wide SSH using TG data from a few coastal stations builds on the capability of GANs to detect and reproduce nonlinear dynamics, as well as learning the dominant relationships of different spatial and temporal signals. We presented the method in detail, motivating interested scientists to apply it to similar natural settings or other oceanographic datasets. In the case when the coastal and open ocean data are consistent (e.g., they are from the same source), as was the case in experiments BAF1 and BLF1, only 19 stations in the locations of the permanently operating TGs are enough for the GAN to ensure an adequate reconstruction. The relatively short time series, which is only 1 year, provides an illustration of a substantial recurrence of events. It was demonstrated that, in this case, the skills of the models used to reconstruct tidally and synoptically driven temporal and spatial variability were almost equally good and comparable to the skill when using the Kalman filter approach. However, differently from the case of optimal linear estimator (e.g., the Kalman filter approach), of particular value is the capability of the GAN to learn and replicate processes with multiple time scales and the associated nonlinear interactions between them.

Using data from different sources (real observations or data from another numerical model) resulted in a decrease in the skill, and the patterns of disagreement with the test data were constrained by the model dynamics, generally reflecting the signal-to-noise ratio. Thus, the index of agreement between the reconstructions and validation data depends strongly on the patterns of the tidal amplitude. Including real coastal observations in the learning process increased the skill of the model. Obviously, GANs optimally learn from data from different sources. The lower skill of the experiments, in which real coastal observations are not used in the training process, reveals a similarity with the problems in data assimilation when errors in the data and model are not treated appropriately. Using other independent observations when training the GAN has the potential to further increase the power of the proposed method in real applications. This method can be attempted in other oceanographic settings.

Chapter 5: Conclusion and Outlook

5.1 Conclusion

Reconstruction of climate fields, for instance, past surface temperatures and sea surface heights, can provide us with a better perspective to learn from the past, and to explore climate and earth system dynamics. Likewise, investigating the past can also help us to better analyze and project potential future climate changes. Especially, the increasing number of extreme events that could exacerbate global temperature rising and global sea level rise, human life and property are under serious threats. Yet, many state-of-the-art reconstruction studies have identified some deficiencies about most CFR methods, for example a general tendency to 'regress to the mean' (Christiansen and Ljungqvist, 2017), which leads to an evident underestimation of climate variability, especially for low-frequency signals. In addition, sparse networks and proxy records with less quality could result in biased reconstructions (Wang et al., 2014; Evans et al., 2014; Amrhein et al., 2020; Po-Chedley et al., 2020). Potential nonlinearities in the linkage between proxy and target climate variables, and the selection of different reconstruction methodologies could also introduce additional reconstruction bias and error. Thus, significant scope remains for further establishing and introducing new CFR methodologies, and in designing methods that are less prone to those previous mentioned common deficiencies.

The goal of this thesis is to employ the newly emerging machine leaning/deep learning methods for reconstructing climate field variabilities, and to compare reconstructions with several traditional CFR methods. Specifically, it was tested whether machine-learning methods, for instance, the artificial neural networks that has been proved with highly nonlinearity mapping capacity, could capture and reproduce more realistic climatic variability and hence reduce more reconstruction biases. In addition, it was tested whether specific neural network methods related to LSTM, ESN and GANs employed in this thesis, are superior compared to traditional CFR methods such as PCR, CCA and Kalman filter.

The thesis consists of **3** main studies: **1)** Evaluation of statistical climate reconstruction methods based on pseudoproxy experiments using linear and machine learning methods. **2)** A comparison and evaluation of Northern Hemisphere summer temperature reconstructions using machine learning and linear regression methods. **3)** Reconstruction of the Basin-Wide Sea-Level Variability in the North Sea Using the Kalman filter method and Generative Adversarial Networks. The according prospects and conclusions based on the initially formulated questions and hypotheses are presented in the following paragraphs.

1). Can a recurrent neural network improve CFRs compared to traditional Multivariate linear methods?

In the first study of the evaluation of statistical climate reconstruction methods based on pseudoproxy experiments using linear and machine learning methods, we implemented the Bi-LSTM method using different structures to reconstruct surface temperature fields for past millennium, and compared the reconstruction results of LSTM with PCR and CCA method, respectively. In contrast to PCR and CCA, the merit of Bi-LSTM is that it does not require a linear and temporally stationary connection between the proxy network and the target climate variable. In addition, Bi-LSTM can incorporate temporal information of the serial correlation related to the time series. Our working hypothesis in the first study is that the Bi-LSTM method could achieve a better performance in reconstructing the amplitude of past temperature variability. This study shows that all three CFR methodologies generally tend to more strongly underestimate the variability of spatially averaged temperature indices as the more noise is introduced into the pseudoproxies. The Bi-LSTM method tested in our experiments using a limited calibration dataset shows relatively low reconstruction skills compared to PCR and CCA. Therefore, our working hypothesis that this machine-learning method with a more complex structure would provide better performance for temperature field reconstructions was not confirmed: The nonlinear artificial neural network method Bi-LSTM employed in the first study is not superior for CFR reconstructions compared with two classical methods, at least based on our PPEs. In general, Bi-LSTM show lower skill metrics in spatial and temporal CFRs compared to PCR and CCA, including the representation of extremes. A perspective derived from this study is to employ a larger set of nonlinear CFR methods to evaluate different model structures, and further test their performance on CFRs.

A second conclusion based on the first study is the general inability to capture extreme cooling signals prior to 20th century. Here the Bi-LSTM fails at reasonably extrapolating temperature amplitudes beyond the training set - a phenomenon that is intrinsic to most ML-based methods. Therefore, compared with linear methods, the Bi-LSTM neural network model did not show clear advantages. The performance of the Bi-LSTM might be further improved by optimizing the architecture and hyperparameters of the network(e.g. by modifications of and inclusion of the type of objective function, type of neural activation function, network optimization function, number of hidden layers, the model-learning rate etc.). In this context, the optimization in the selection/settings of these hyper-parameters could be further explored, to test the according sensitivity and extend on the reconstruction skill. Nadiga (2020) indicated that the performance of some machine learning-methods is strongly dependent on these hyper-parameters settings. Machine learning methods usually include an extensive range of complexity, and therefore it remains an open issue as to which ML techniques are suitable for climate reconstructions.

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 robust and optimal architecture. This included two separated hidden layers, which could be seen as 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 might speculate that more complicated architectures might not be helpful for improving CFRs. One factor that needs to be emphasized is that a degradation of out-of-sample performance may well be expected when a limited amount of a dataset is employed to train a neural network model (Najafabadi et al., 2015). Finally, we would like to point out other methods, such as an Echo State Network (Lukosevicius and Jaeger, 2009; Nadiga, 2020) for climate reconstruction research. Since both ESN and LSTM belong to the family of RNNs, yet ESN is much simpler than LSTM (Lukoševičius and Jaeger 2009), and has outperformed the RNNs in many other applications (Chattopadhyay et al., 2020; Nadiga, B. 2020).

2). Does reservoir computing improve CFR skills compared to traditional Multivariate linear methods?

In this study, a comparison and evaluation of Northern Hemisphere summer temperature field reconstructions was implemented using machine learning and linear regression methods, a reservoir computing method - ESN is established together with Bi-LSTM, PCA and CCA methods. Based on pseudoproxy reconstruction experiments. The ESN method belonging to the family of reservoir computing was employed for temperature field reconstructions, and according results of ESN exhibit an encouraging performance in capturing more temperature variance. The ESN method also achieves better reconstruction skill compared to the two traditional multivariate linear-based regression methods PCR and CCA.

Our working hypothesis for the second study, that the more complex Bi-LSTM does not show superiority in CFRs compared to traditional CFR methods and the relatively simpler ESN methods, confirmed the conclusion of the first study that more complicated architecture might not be helpful for improvement of surface temperature field reconstruction, at least based on our specific experiments. One point that needs to be emphasized is that the employed Bi-LSTM method has the same structure and hyper-parameters as the one we have tested in the first study. Nevertheless, based on the general reconstruction performance achieved by the ESN employed in this study, we would like to draw a general conclusion that employing relatively simpler architecture-based nonlinear machine learning method might be helpful in designing CFRs.

The ESN and LSTM method both belong to the family of RNN. The ESN method which only consists of three layers is much simpler than LSTM (Lukosevicius and Jaeger 2009) both in model-training and model-establishing procedures, and has been demonstrated to be able to outperform the RNNs in different applications (Chattopadhyay et al., 2020; Nadiga, B. 2020). We thus encourage testing ESN in different paleoclimate research directions for future CFR studies.

3). Does the Convolutional neural network produce achieve reasonable CFRs?

In the study of the Reconstruction of the Basin-Wide Sea-Level Variability in the North Sea Using Kalman filter method and Generative Adversarial Networks, we established a deep learning method GANs together with a data assimilation method Kalman filter to reconstruct sea surface height fields of North Sea.

The method proposed here to reconstruct the basin-wide SSH using Tidal Gauges (TGs) data from a few coastal Tidal gauges stations builds on the capability of GANs to identify and reproduce nonlinear ocean dynamics, as well as learning the dominant relationships of different spatial and temporal signals. In the case when the coastal and open ocean data are consistent (e.g., they are from the same data source), only 19 stations in the locations of the permanently operating TGs are sufficient for the GANs to ensure an adequate SSH field reconstruction. The relatively short time series, which is only one year, provides an illustration of a substantial recurrence of events. It was demonstrated that, in this case, the skills of the models used to reconstruct tidally and synoptically driven temporal and spatial variability were almost equally good. Of particular value is the capability of the GAN to learn and replicate processes with multiple time scales and the associated nonlinear interactions between them.

Using data from different sources (real observations or data from different climate models) resulted in a decrease in the reconstruction skill, and the patterns of disagreement with the test data were constrained by the model dynamics, generally reflecting the magnitude of signal-to-noise ratio. Thus, the index of agreement between the reconstruction and validation data depends strongly on the patterns of the tidal amplitude. Including real coastal observations in the learning process increased the skill of the model. Obviously, GANs can learn from data of different sources. The lower skill of the experiments, in which real TG observations are not used in the training process, reveals a similarity with the problems in data assimilation when errors in the data and model are not treated appropriately. Using other independent observations when training the GANs has the potential to further increase the power of the proposed method in real applications. This method can be attempted in other oceanographic settings.

In general, the GANs method we employed were capable of reconstructing sea surface height fields with reasonable and comparable accuracy compared with the Kalman filter approach. In addition, based on the experiments in the third study, the GANs can learn from data of different sources adequately, and general ocean dynamics could be reproduced successfully.

To sum up, this dissertation tested different state-of-the-art machine learning methods, including LSTM, ESN and GANs, on climate field reconstructions, and compared CFRs results reconstructed from machine leaning methods with reconstructions from several traditional CFR methods, including PCR, CCA and a data assimilation method. In order to better and more accurately reconstruct past climate and ocean dynamics for global and regional scales, it is necessary to test different techniques, especially related to the newly emerging non-linear machine learning approaches. These machine-learning methods have been proven to be able to better capture and reproduce the underlying nonlinearities, and reduce potential uncertainties in physical processes, especially on the application of climate and ocean science.

Moreover, based on our three separated studies using different machine learning methods and traditional CFR methods for climate reconstructions, our general conclusion is that better or comparable climate reconstructions could be achieved compared to traditional CFR methodologies by employing state-of-the-art artificial neural networks with appropriate structures and optimized and tested hyperparameters.

5.2 Outlook

This thesis employs different machine learning methods on the application of climate reconstructions in three separated studies. Results from these three separated studies demonstrate that for neural network methods, when tested and selected with appropriate structures and models, comparable or better reconstructions can be realized compared to traditional CFR or data assimilation methods.

Nevertheless, it has been pointed out that the capability of machine learning methodologies are strongly dependent on the selection and setting of hyperparameters (Najafabadi et al., 2015; Nadiga, 2020). An extensive range of complexity can be involved in Machine learning, and thus it remains an open question as to which method is more suitable for specific applications. Until now, it is still not clear how to systematically optimize the architecture of ML methods for a particular purpose in climatic and oceanographic context. A considerably amount of 'trial and error' still remains in the design of the neural structures.

Here, we have tested three neural network methods originated from the family of machine learning; we employed Bi-LSTM, ESN, and the convolutional neural network based GANs, and evaluated their reconstruction performances on reconstruction experiments, which could be a preliminary attempt. In these studies, we heuristically tested and selected the hyperparatmeters of each neural network in several necessary numerical experiments to explore a more suitable structure. These heuristic approach could introduce additional uncertainties, for instance, the reconstruction results could be affected by subjective decisions and settings. If we want to search for an optimal machine learning method with optimal structures and hyperparameters, it will take additional work. In fact, currently employed machine learning architectures have been mostly established and selected by human experts. Therefore, the automated neural architecture search methodologies (Elsken, T et al., 2018), which can be categorized as a subfield of Auto machine leaning- **AutoML** (Hutter, F et al., 2019) may provide an additional perspective. It is valuable to test the AutoML method for further CFR studies, which can provide additional avenues to improve the efficiency of machine learning application on climate and ocean research.

For solving a specific task using Machine learning, we practitioners have a set of raw data that would be employed for training and testing a machine leaning method. In order to feed the raw dataset in to a neural network model and run it successfully, an expert with necessary domain knowledge may have to apply appropriate data pre-processing, for instance, feature extraction and feature selection. When the preprocessed dataset is prepared, we usually need to apply a selection algorithm, choose a model structure, and seek a hyperparamter optimization to establish an architecture that can potentially fulfill our tasks in an optimal manner. AutoML aims to simplify these procedures for non-experts, and make the application of machine learning for a specific task more efficiently. For instance, we could set a predefined network parameter within certain range (for instance fixed hidden neural numbers or layer numbers within a certain range); the automated neural architecture search method could select an optimal hidden neural number or layer number based on objective function (lowest mean square error) within the predefine range. Thus, we do not have to test each specific number one by one for determining which the hidden neural number or layer number would be the optimal in our specific task.

In the future, we therefore would like to test AutoML methods to optimize the structure and hyperparameters of machine-learning methods employed in this thesis for future CFR experiments. This will finally allow to check whether AutoML could help us find the best neural network architectures and hyperparameter for the application on climate field reconstructions.

In addition, the machine learning method can provide us with a better way to incorporate real climatic observation records into the final climate field reconstructions. For instance, it was demonstrated that machine-learning models can provide a better understanding of the connections between leaf physiognomy and climate, and improve the final reconstructions (Wei et al., 2021). Another new direction is that it can be employed to project the climate field by considering and fusing multisource data. For instance, Zhu et al. (2019) proposed a new temperature reconstruction method by employing mutilsource data, including NDVI (Normalized Difference Vegetation Index) and multisource remotely sensing data such as MODIS (Moderate Resolution Imaging Spectroradiometer) land surface temperature data, which can improve the inversion accuracy of surface temperature with high spatial resolution in a wide range significantly.

Based on our experiments tested in this thesis, machine learning also has a well-applied direction on the application of oceanography and biogeochemistry reconstructions (Bennington et al., 2022; Stanev et al., 2022; Liu et al., 2022). In addition, monitoring and extracting the sea ice cover based on remote sensing dataset using machine learning methods (Li et al., 2019; Wang et al., 2021) will provide us with a new direction for analyzing or predicting its general variability, which is import for monitoring global climate changes.

A review publication about the application of machine learning methods to costal sediment transport and morphodynamics summarized a set of practices for coastal researchers (Goldstein et al., 2019), for instance, predictions of coastal overwash on a developed island, small-scale projections of sediment transport to larger-scale sand bar morphodynamics. This overview presents a promising direction that machine-learning methods are effective alternative approaches involving, time-consuming fluid dynamics, data requirements, and numerical models (Kim and Lee, 2022).

Reference

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.

Andersen, O. B.: Shallow water tides in the northwest European shelf region from TOPEX/POSEIDON altimetry, J. Geophys. Res., 104, 7729–7741, doi:10.1029/1998JC900112, 1999.

Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., and Ott, E.: A Machine Learning-Based Global Atmospheric Forecast Model, Geophys. Res. Lett., 47, e2020GL087776, https://doi.org/10.1029/2020GL087776, 2020.

Baker, A., Fuller, L., Genty, D., Fairchild, I. J., Jex, C., and Smith, C. L., Annually laminated stalagmites: a review, Int. J. Speleol., 37, 193–206, 2008.

Barth, A., Alvera-Azcárate, A., Licer, M., and Beckers, J.-M.: DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations, Geosci. Model Dev., 13, 1609–1622, https://doi.org/10.5194/gmd-13-1609-2020, 2020.

Barth, A., Alvera-Azcárate, A., Troupin, C., and Beckers, J.-M.: DINCAE 2.0: multivariate convolutional neural network with error estimates to reconstruct sea surface temperature satellite and altimetry observations, Geosci. Model Dev., 15, 2183–2196, https://doi.org/10.5194/gmd-15-2183-2022, 2022.

Becker, G., Dick, S., and Dippner, J.: Hydrography of the German Bight, Mar. Ecol.-Prog. Ser., 91, 9–18, 1992.

Bennington, V., Galjanic, T., and McKinley, G. A.: Explicit physical knowledge in machine learning for ocean carbon flux reconstruction: The pCO2-Residual method. J. Adv. Model. Earth Syst., 14, e2021MS002960, https://doi.org/10.1029/2021MS002960, 2022.

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.

Boisvenue, C., and Running, S. W.: Impacts of climate change on natural forest productivity—evidence since the middle of the 20th century. Glob. Chang. Biol., 12, 862–882, 2006.

Burger, G., Fast, I., and Cubasch, U.: Climate reconstruction by regression-32 variations on a theme, Tellus A, 58(1), 227–235, 2006.

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.

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.

Cipollini, P., Calafat, F. M., Jevrejeva, S., Melet, A., and Prandi, P.: Monitoring sea level in the coastal zone with satellite altimetry and tide gauges, Surv. Geophys., 38, 35–59, https://doi.org/10.1007/s10712-016-9392-0, 2017.

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.

Coats, S., Smerdon, J. E., Cook, B. I., and Seager, R.: Are simulated megadroughts in the North American southwest forced?, J. Climate, 28, 124–142, https://doi.org/10.1175/jcli-d-14-00071.1, 2015b.

Cook, E. R. and Kairiukstis, L. A.: Methods of dendrochronology: applications in the environmental sciences, Kluwer Academic Publishers, Dordrecht, 394 pp., 1990.

Cotter, A., Shamir, O., Srebro, N., Sridharan, K.: Better mini-batch algorithms via accelerated gradient methods. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), Advances in Neural Information Processing Systems24 (NIPS 2011) (pp. 1647–1655), .2011.

http://papers.nips.cc/paper/4432-better-mini-batch-algorithms-via-accelerated-gradient-methods.pdf

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.

D'Arrigo, R., Wilson, R., Liepert, B., and Cherubini, P.: On the 'Divergence Problem' in Northern Forests: A review of the treering evidence and possible causes, Global. Planet. Change, 60,289–305, 2008.

Dong, J.; Yin, R.; Sun, X.; Li, Q.; Yang, Y.; Qin, X. Inpainting of Remote Sensing SST Images With Deep Convolutional Generative Adversarial Network. IEEE Geosci. Remote Sens. Lett., 16, 173–177, 2019.

Dong, H, W., Yang, Y, H.: Towards a deeper understanding of adversarial losses. arXiv[preprint]. https://arxiv.org/abs/1901.08753, 2019.

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, Franke, J., Valler, V., Brönnimann, S., Neukom, R., and Jaume-Santero, F.: The importance of input data quality and quantity in climate field reconstructions – results from the assimilation of various tree-ring collections, Clim. Past, 16, 1061–1074, https://doi.org/10.5194/cp-16-1061-2020, 2020.

Egbert, G. D. and Erofeeva, L.: Efficient inverse modeling of barotropic ocean tides, J. Atmos. Ocean. Tec., 19, N2, 183–204, 2002

Eiler, J. M.: Paleoclimate reconstruction using carbonate clumped isotope thermometry, Quatern. Sci. Rev., 30, 3575–3588, 2011.

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.

Esper, J., Cook, E. R., and Schweingruber, F. H.: Lowfrequency signals in long tree-ring chronologies for reconstructing past temperature variability, Science, 295, 2250–2253, doi:10.1126/science.1066208, 2002.

Fairchild, I. J. and Treble, P. C.: Trace elements in speleothems as recorders of environmental change, Quaternary Sci. Rev., 28, 449–468, 2009.

Flather, R. A., and Proctor, R.: Prediction of North Sea storm surges using numerical models: Recent developments in the U.K. In J. Sundermann, & W. Lenz (Eds.), North Sea Dynamics (pp. 299–317). Berlin, Heidelberg: Springer, https://link.springer.com/chapter/10.1007/978-3-642-68838-6_21, 1983.

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.

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.

Fritts, H. C.: Tree rings and climate, Academic Press, London, 1976.

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.

Gagen, M., McCarrol, D., and Hicks, S.: The Millennium project: European climate of the last millennium, PAGES News, 14, p. 4, 2006.

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, J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence & K.Q. Weinberger(Eds.), Advances in Neural Information Processing Systems27 (NIPS 2014) (pp. 2672–2680). Montreal, Quebec, Canada: Curran Associates, Inc, https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf, 2014.

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

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.

Gordon, E. M., Barnes, E. A., and Hurrell, J. W.: Oceanic harbingers of Pacific decadal oscillation predictability in CESM2 detected by neural networks. Geophys. Res. Lett., 48(21). https://doi.org/10.1029/2021gl095392, 2021.

Goldstein, E. B., Coco, G., and Plant, N. G.: A review of machine learning applications to coastal sediment transport and morphodynamics, Earth-Sci. Rev., 194, 97–108, https://doi.org/10.1016/j.earscirev.2019.04.022, 2019.

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.

Grayek, S., Staneva, J., Schulz-Stellenfleth, J., Petersen, W., and Stanev, E. V.: Use of FerryBox surface temperature and salinity measurements to improve model based state estimates for the German Bight, J. Mar. Syst., 88, 45–59, doi:10.1016/j.jmarsys.2011.02.020, 2011.

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.

Haigh, I. D., Pickering, M. D., Green, J. A. M., Arbic, B. K., Arns, A., Dangendorf, S., Hill, D. F., Horsburgh, K., Howard, T., Idier, D., Jay, D. A., Jänicke, L., Lee, S. B., Müller, M., Schindelegger, M., Talke, S. A., Wilmes, S.-B., and Woodworth, P. L.: The Tides They Are A – Changin: A Comprehensive Review of Past and Future Nonastronomical Changes in Tides, Their Driving Mechanisms, and Future Implications, Rev. Geophys., 57, e2018RG000636, https://doi.org/10.1029/2018rg000636, 2020.

Hansen, W.: Theorie zur Errechnung des Wasserstands und der Stromungen in Randemeeren, Tellus, 8, 287–300, 1956.

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.

Hausfather, Z., Drake, H. F., Abbott, T., and Schmidt, G. A.: Evaluating the Performance of Past Climate Model Projections, Geophys. Res. Lett., 47, e2019GL085378, https://doi.org/10.1029/2019GL085378, 2020.

Heaps, N. S.: A Two-Dimensional Numerical Sea Model Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences, 265(1160), 93–137, 1969.

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.

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

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

Ho-Hagemann, H. T. M., Hagemann, S., Grayek, S., Petrik, R., Rockel, B., Staneva, J., Feser, F., and Schrum, C.: Internal Model Variability of the Regional Coupled System Model GCOAST-AHOI, Atmosphere, 11, 227, https://doi.org/10.3390/atmos11030227, 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.

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.

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.

Huthnance, J. M.: Physical oceanography of the North Sea, Ocean and Shoreline Management, 16, 199–231, https://doi.org/10.1016/0951-8312(91)90005-M, 1991.

Hutter, F.; Kotthoff, L.; Vanschoren, J. Automated Machine Learning: Methods, Systems, Challenges; Springer Nature: Berlin/Heidelberg, Germany, 2019.

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.

Ioffe, S., and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, PMLR, (pp.448–456). http://proceedings.mlr.press/v37/ioffe15.pdf, 2015.

Jacob, B. and Stanev, E. V.: Interactions between wind and tidally induced currents in coastal and shelf basins, Ocean Dynam., 67, 1263–1281, https://doi.org/10.1007/s10236-017-1093-9, 2017.

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.

Jaeger, H. The "Echo State" Approach to Analysing and Training Recurrent Neural Networks-with an Erratum Note. GMD Technical Report 148 (German National Research Center for Information Technology, Bonn, 2001).

Jaeger, H. and Haas, H.: Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, Science, 304, 78–80, 2004.

Jaeger, H., Maass, W., and Principe, J.: Special issue on echo state networks and liquid state machines.Neural Networks. 20, 287–289, 2007.

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.

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.

Kim, T., Lee, W.D.: Review on Applications of Machine Learning in Coastal and Ocean Engineering, J. Ocean. Eng. Technol, 36, 194–210, https://doi.org/10.26748/KSOE.2022.007, 2022.

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

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.

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.

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.

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.

LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature, 521, 436–444. https://doi.org/10.1038/nature14539, 2015.

Legates, D. R. and McCabe, G. J.: Evaluating the use of “goodnessof-fit” measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, 1999.

Lewis, S. C. and LeGrande, A. N.: Stability of ENSO and its tropical Pacific teleconnections over the Last Millennium, Clim. Past, 11, 1347–1360, https://doi.org/10.5194/cp-11-1347-2015, 2015.

Liu, H, Y., Jiang, B., Xiao, Y., and Yang, C.: Coherent semantic attention for image inpainting. The IEEE International Conference on Computer Vision (ICCV 2019), (pp. 4170–4179). http://openaccess.thecvf.com/content_ICCV_2019/html/Liu_Coherent_Semantic_Attention_for_Image_Inpainting_ICCV_2019_paper.html, 2019.

Liu, H., Lin, L., Wang, Y., Du, L., Wang, S., Zhou, P., Yu, Y., Gong, X., and Lu, X.: Reconstruction of Monthly Surface Nutrient Concentrations in the Yellow and Bohai Seas from 2003–2019 Using Machine Learning. Remote Sens. 14, 5021, https://doi.org/10.3390/rs14195021, 2022.

Loshchilov, I., Hutter, F.: Decoupled weight decay reg-ularization. arXiv[preprint]. https://arxiv.org/abs/1711.05101, 2019.

Lu, Z., Pathak. J., Hunt, B., Girvan, M., Brockett, R., and Ott, E.: Reservoir observers: Model-free inference of unmeasured variables in chaotic systems, Chaos, 27, 041102, https://doi.org/10.1063/1.4979665, 2017.

Lu, Z., Hunt, B. R., and Ott, E.: Attractor reconstruction by machine learning, Chaos, 28, 061104, https://doi.org/10.1063/1.5039508, 2018.

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.

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.

Li, X.-M., Sun, Y., and Zhang, Q.: Extraction of Sea Ice Cover by Sentinel-1 SAR Based on Support Vector Machine With Unsupervised Generation of Training Data, IEEE T. Geosci. Remote, 59, 3040–3053, https://doi.org/10.1109/TGRS.2020.3007789, 2020.

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.

Maas, A. L., Hannun, A. Y., and Ng, A. Y.: Rectifier nonlinearities improve neural network acoustic models. http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf, 2013.

Maass, W., Natschläger, T., and Markram, H.: Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Computation, 14(11), 2531–2560, 2002.

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.

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.

Mao, X. J., Shen, C. H., and Yang, Y. B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, & R. Garnett(Eds.), Advances in Neural Information Processing Systems 29 (NIPS 2016), (pp.2810–2818), 2016.

McCarroll, D., Young, G., and Loader, N.: Measuring the skill of variance-scaled climate reconstructions and a test for the capture of extremes, Holocene, 25, 618–626, https://doi.org/10.1177/0959683614565956, 2015.

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.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models, Part I - A discussion of principles, J. Hydrol., 10, 282–290, 1970.

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.

O'Dea, E. J., Arnold, A. K., Edwards, K. P., Furner, R., Hyder, P., Martin, M. J., Siddorn, J. R., Storkey, D., While, J., Holt, J. T., and Liu, H.: An operational ocean forecast system incorporating NEMO and SST data assimilation for the tidally driven European North-West shelf, J. Oper. Oceanogr., 5, 3–17, 2012.

Osborne, M. J., and Rubinstein, A.: A Course in Game Theory. (pp. 14). Cambridge, MA: MIT. ISBN 9780262150415, 1994.

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.

Otto, L., Zimmerman, J. T. F., Furnes, G. K., Mork, M., Saetre, R., and Becker, G.: Review of the physical oceanography of the North Sea, Netherlands, J. Sea Res., 26, 161–238, 1990.

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)

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)

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 2k Consortium)

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.

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.

Pascanu, R., Mikolov, T., and Bengio, Y.: On the difficulty of training recurrent neural networks, in: International Conference on Machine Learning, 16–21 June, Atlanta, USA, 1310–1318, 2013

Pathak, J., Hunt, B., Girvan, M., Lu, Z., and Ott, E.: Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach, Phys. Rev. Lett., 27, https://doi.org/10.1063/1.5010300, 2018

Peeck, H.H., Proctor, R., and Brockmann, C.: Operational storm surge models for the North Sea. Continental Shelf Research, 2(4), 317–329, https://doi.org/10.1016/0278-4343(82)90024-3, 1983.

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.

Proudman, J., and Doodson, A. T.: The principal constituent of the tides in the North Sea. Philosophical Transactions of the Royal Society of London, A(244), 185–219, 1924.

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.

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

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, 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.

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

Rischard, M., McKinnon, K. A., and Pillai, N.: Bias correction in daily maximum and minimum temperature measurements through Gaussian process modeling, arXiv:1805.10214v2, 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.

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.

Schneider, T.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, J. Climate, 14, 853–871, 2001.

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.

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.

Sheldon, N. D. and Tabor, N. J.: Quantitative paleoenvironmental and paleoclimatic reconstruction using paleosols, Earth-Sci. Rev., 95, 1–52, doi:10.1016/j.earscirev.2009.03.004, 2009.

Sheppard, P. R.: Dendroclimatology: Extracting climate from trees. Wiley Interdiscip. Rev. Clim. Chang. 1(3), 343–352, https://doi.org/10.1002/wcc.42, 2010.

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, 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.

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L. V., Hegerl, G., Seneviratne, S. I., Vautard, R., Zhang, X., and Zwiers, F. W.: Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities, Weather and Climate Extremes, 18, 65–74, https://doi.org/10.1016/j.wace.2017.10.003, 2017.

Elsken, T., Metzen, J. H., and Hutter, F.: Neural architecture search: a survey. J. Mach. Learn. Res. 20, 1–21, 2019.

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.

Smerdon, J. E. and Pollack, H. N.: Reconstructing Earth’s surface temperature over the past 2000 years: the science behind the headlines, WIREs Clim. Change, 7, 746–771, https://doi.org/10.1002/wcc.418, 2016.

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.

Speer, J. H.: Fundamentals of tree ring research, The University of Arizona Press, Tucson, 2010.

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.

Stanev, E. V., Wahle, K., and Staneva, J.: The synergy of data from profiling floats, machine learning and numerical modeling: Case of the Black Sea euphotic zone.J. Geophys. Res.-Oceans, 127,e2021JC018012, https://doi.org/10.1029/2021JC018012, 2022.

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

Radford, A.; Metz, L.; Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv[preprint]. https://arxiv.org/abs/1511.06434, 2015.

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.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W.M., & Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, (Vol 9351, pp.234–241). Cham: Springer. https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Soetje, K. C. and Brockmann, C.: An operational numerical model of the North Sea and the German Bight, in: North Sea Dynamics, edited by: Sundermann, J. and Lenz, W., Springer-Verlag, Berlin, ¨ Heidelberg, New York, 693 pp., 1983.

Stoffel, M. and Bollschweiler, M.: Tree-ring analysis in natural hazards research – an overview, Nat. Hazards Earth Syst. Sci., 8, 187–202, https://doi.org/10.5194/nhess-8-187-2008, 2008.

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.

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.

Su, X., Yan, X., and Tsai, C.-L.: Linear Regression: Linear Regression. WIREs Comp. Stat., 4, 275–29, 2012.

Tan, M., Baker, A., Genty, D., Smith, C., Esper, J., and Cai, B.: Applications of stalagmite laminae to paleoclimate reconstructions: comparison with dendrochronology/climatology, Quaternary Sci. Rev., 25, 2103–2117, 2006.

Tejedor, E., Steiger, N., Smerdon, J., Serrano-Notivoli, R., and Vuille, M.: Global temperature responses to large tropical volcanic eruptions in paleo data assimilation products and climate model simulations over the Last Millennium, Paleoceanography and Paleoclimatology, 36, e2020PA004128, https://doi.org/10.1029/2020PA004128, 2021.

Tejedor, E., Steiger, N. J., Smerdon, J. E., Serrano-Notivoli, R., and Vuille, M.: Global hydroclimatic response to tropical volcanic eruptions over the last millennium, P. Natl. Acad. Sci. USA, 118, e2019145118, https://doi.org/10.1073/pnas.2019145118, 2021.

Tingley, M. P. and Huybers, P.: A bayesian algorithm for reconstructing climate anomalies in space and time. Part I: development and applications to paleoclimate reconstruction problems, J. Climate, 7, 2759–2781, doi:10.1175/2009JCLI3015.1, 2010.

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., MannshardtShamseldin, E., and Rajaratnam, B.: Piecing together the past: statistical insights into paleoclimatic reconstructions, Quaternary Sci. Rev., 35, 1–25, 2012.

Toms, B. A., Barnes, E. A., and Hurrell, J. W.: Assessing decadal predictability in an Earth-system model using explainable neural networks. Geophysical Research Letters, 48, e2021GL093842. https://doi.org/10.1029/2021GL093842, 2021.

Tonani, M., Sykes, P., King, R. R., McConnell, N., Péquignet, A.-C., O'Dea, E., Graham, J. A., Polton, J., and Siddorn, J.: The impact of a new high-resolution ocean model on the Met Office North-West European Shelf forecasting system, Ocean Sci., 15, 1133–1158, https://doi.org/10.5194/os-15-1133-2019, 2019.

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.

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.

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.

Wahl, T., Haigh, I. D., Woodworth, P. L., Albrecht, F., Dillingh, D., Jensen, J., Nicholls, R. J., Weisse, R., and Wöppelmann, G.: Observed mean sea level changes around the North Sea coastline from 1800 to present, Earth-Sci. Rev., 124, 51–67, doi:10.1016/j.earscirev.2013.05.003, 2013.

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.

Wang, Y.-R. and Li, X.-M.: Arctic sea ice cover data from spaceborne synthetic aperture radar by deep learning, Earth Syst. Sci. Data, 13, 2723–2742, https://doi.org/10.5194/essd-13-2723-2021, 2021.

Wei, G., Peng, C., Zhu, Q., Zhou, X., Yang, B.: Application of machine learning methods for paleoclimatic reconstructions from leaf traits. Int. J. Climatol, 41, E3249–E3262, 2021.

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.

Willmott, C. J.: On the validation of models, Phys. Geogr., 1, 184– 194, 1981.

Willmott, C. J., Robeson, S. M., and Matsuura, K.: A refined index of model performance, Int. J. Climatol., 32, 2088–2094, doi:10.1002/joc.2419, 2012.

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.

Wong, C. I. and Breecker, D. O.: Advancements in the use of speleothems as climate archives, Quaternary Sci. Rev., 127, 1–18, 2015.

Yu, J. H., Lin, Z., Yang, J. M., Shen, X. H., Lu, X., and Huang, T. S. Generative Image Inpainting with Contextual Attention. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 5505–5514), 2018. http://openaccess.thecvf.com/content_cvpr_2018/papers/Yu_Generative_Image_Inpainting_CVPR_2018_paper.pdf

Yu, Y., Si, X., Hu, C. and Zhang, J.: A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput, 31, 1235–1270, https://doi.org/10.1162/neco_a_01199, 2019.

Yun, S., Smerdon, J. E., Li, B., and Zhang, X.: A pseudoproxy assessment of why climate field reconstruction methods perform the way they do in time and space, Clim. Past, 17, 2583–2605, https://doi.org/10.5194/cp-17-2583-2021, 2021.

Zador, A. M.: A critique of pure learning and what artificial neural networks can learn from animal brains. Nature communications, 10(1):1–7, https://doi.org/10.1038/s41467-019-11786-6, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. arXiv. Retrieved from https://arxiv.org/abs/1611.03530

Zhao, H., Gallo, O., Frosio, I., and Kautz, J.: Loss Functions for Image Restoration With Neural Networks. IEEE Transactions on Computational Imaging, 3(1), 47–57. DOI: 10.1109/TCI.2016.2644865, 2017.

Zorita, E., Kharin, V., and von Storch, H.: The atmospheric circulation and sea surface temperature in the North Atlantic area in winter: their interaction and relevance for Iberian precipitation, J. Climate, 5, 1097–1108, 1992.

Zhu, X., Zhang, Q., Xu, C.-Y., Sun, P., and Hu, P.: Reconstruction of high spatial resolution surface air temperature data across China: A new geo-intelligent multisource data-based machine learning technique, Sci. Total Environ., 665, 300–313, https://doi.org/10.1016/j.scitotenv.2019.02.077, 2019.

List of Appendix

Appendix 2A

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°.

Appendix 2B

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 2B.1-2B.6 present reconstruction statistics skill for the spatial North Hemisphere mean temperature in the verification period for ideal PPEs based on CESM 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 2B.1. 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.2. Different learning rate using Huber loss, and with the rest parameters fixed as in Table 2B.1

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 2B.3. Different activation functions with the rest parameters fixed as in Table 2B.1

Activation function	*cc*	SD Ratio
ReLU	0.505	0.566
Leaky ReLU	0.462	0.770
ELU	0.529	0.617
PReLU	0.509	0.544

Table 2B.4. Different hidden layer number with the rest parameters fixed as in Table 2B.1

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 2B.5. Different hidden node numbers in each layer with the rest parameters fixed as in Table 2B.1

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 2B.6. With and without dropout layers conditioned on the rest parameters are fixed as in Table 2B.1

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

Appendix 2C

Appendix 2C 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.

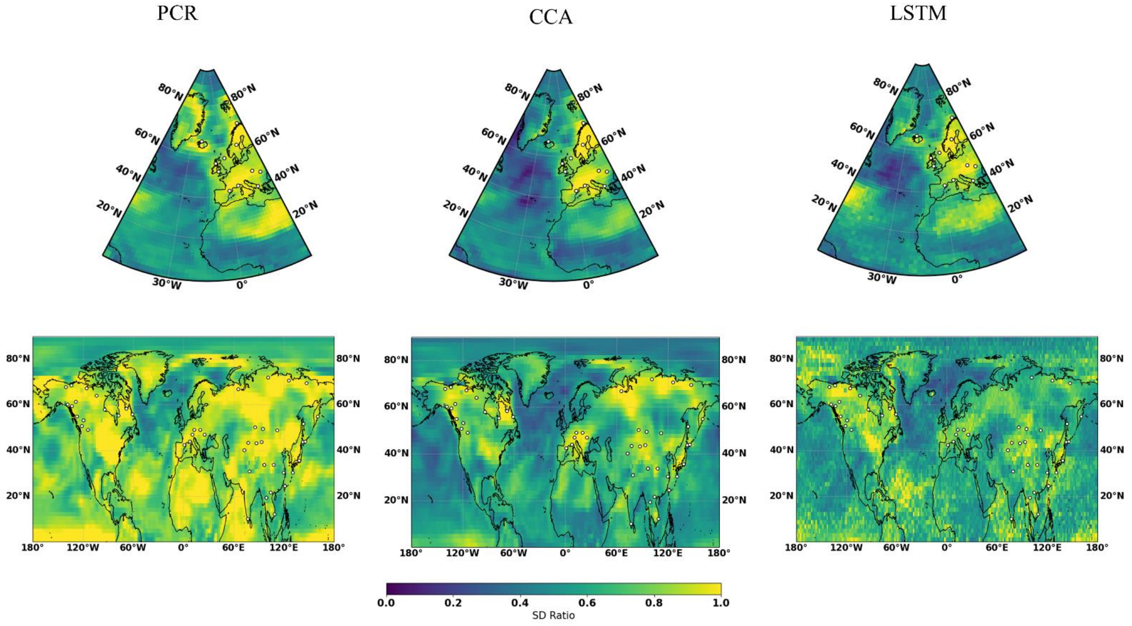

Figure 2C. 1: 30year filtered SD ratio pattern using Ideal-PPEs based on MPI model over validation period 850-1899 for NAE (upper row) and NH (lower row)

Appendix 3A

The ESN model parameters tested and selected in our reconstruction experiments.

Table 3A.1 ESN Parameters:

Details of ESN architecture	
Number of recurrent neurons in hidden reservoir layer	535
Activation function	Relu
Bias scaling factor (Scales the input bias of the activation)	0.1
Distribution for random weight matrices	Central unit normal
Ridge regression coefficient	1e-3
Spectral radius (Scales the recurrent weight matrix)	0.9

Appendix 4A

Loss functions of the generator and discriminator

The Unet-like GAN is established to train two defined network models: Generator (*G*(*Z*)) and Discriminator (*D*). *G*(*Z*) is established to reconstruct SSH maps with full information, where *Z* denotes the input SSH maps with only several tidal gauge point information available. Generated SSH maps from *G*(*Z*) together with real SSH maps from the numerical model are fed into the discriminator (*D*), and the outputs of *D* are a dense scalar that represents the probability of discriminating whether the input maps are from the real input map (which is from the numerical model). The stopping criterion of model training is that the discriminator cannot distinguish whether the generated SSH maps are from real maps. The GAN loss functions are defined as in Goodfellow et al. (2014):

$$L_D^{GAN} = E[\log(D(x))] + E[\log(1 - D(G(z)))] \quad (4A.1)$$

$$L_G^{GAN} = E[\log(D(G(z)))] \quad (4A.2)$$

where L_G is the generator loss and L_D is the discriminator loss. The GAN model parameters are trained and updated based on the following minimization and maximization method:

$$\min\max V(G,D) = E_{x\sim P_{data}(x)}[\log(D(x))] + E_{Z\sim P_z(z)}[\log(1 - D(G(z)))] \quad (4A.3)$$

Here, $P_{\text{data}}(x)$ is the real ssh map data distribution, taking real data sample *x* from $P_{\text{data}}(x)$. While $P_z(z)$ is the Input(z) (as in following figure) ssh data distribution, sampling *z* from $P_z(z)$. *E* is expectation operator. Based on the original DCGAN loss function, we introduce also $L1$ as pixel wise reconstruction loss and $L2$ as content loss into the Generator for accomplishing our experiments, the Discriminator loss is the same as original GAN. The overall loss for the Generator is

$$L_1 = \|G(z) - x\| \quad (4A.4)$$

$$L_2 = \|G(z) - x\|_2^2 \quad (4A.5)$$

$$L_G^{GAN} = E[\log(D(G(z)))] + L1 + L2 \quad (4A.6)$$

While for our LF SSH surge reconstruction model, Generator and Discriminator loss is keep as the same with SSH reconstruction model. Besides, an additional pixel wise reconstruction loss La is introduces into the Generator part

$$L_a = \left\|G_{O_c}(z) - x\right\| \quad (4A.7)$$

$$L_G^{GAN} = E[\log(D(G(z)))] + L1 + L2 + L_a \quad (4A.8)$$

Here, the $G_{Oc}(z)$ represents the generated coarse sea surface LF maps data sample. And *x* is the same as above denotes the real numerical sample data.

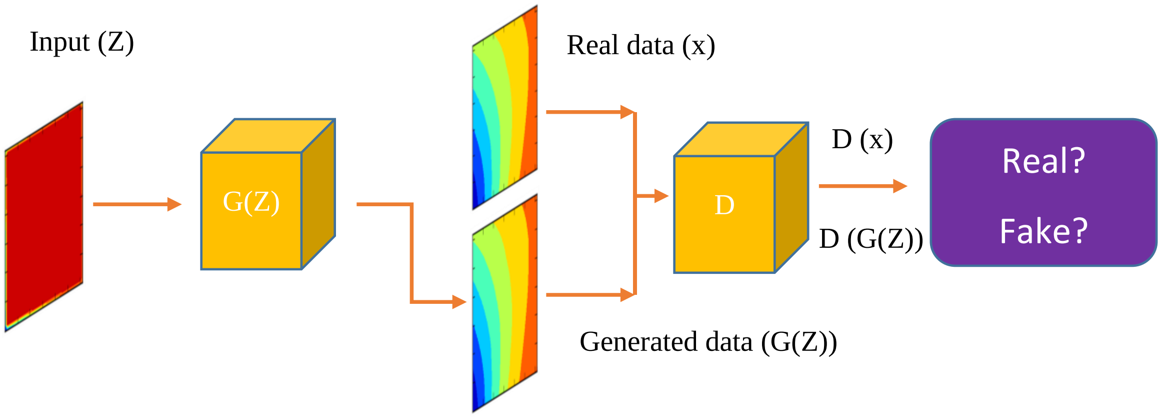

Figure 4A.1. Schematic representation of reconstruction of sea level variability using Generative Adversary networks system. Input (Z) is the SSH map with only tidal gauges point data available, G(Z) denotes the Generator module for mapping Input (Z) to Generated data by samples G(Z). Real data (x) is the full SSH data set from the numerical model, D represents the Discriminator module for discriminating the real data (x) and the Generated data(G(Z)) as Real or Fake samples

Acknowledgements:

I would like to express my sincerest gratitude to my supervisor Prof. Dr. Corinna Schrum. She is not only an academic supervisor who always provides help, support for my PhD research work, but also a tutor for my life. Whenever, I got difficulties both in my research work and in my routine life, she is always there and always ready for providing strong support. She taught me why we should do a PhD study, and also how can we do it successfully. She saved me from many unexpected disasters. So I would say, without strong support and help from her, I cannot finish my PhD study here in Germany, and I cannot even survive from difficulties in my work and life during the 5 years stay in Germany.

I would like to say thanks to my co-supervisor Dr. Eduardo Zorita and Dr. David Greenberg. When I was in the greatest need of help, they stand out and give me whatever they could provide. As senior scientists, they always give me professional guidance, and whenever I have questions or doubts on my thesis process, they are always there and ready for answering any questions. I really appreciate that they are my co-supervisors.

I would also want to say thanks to my SICSS panel chair Prof. Dr. Uwe Schneider, and my panel member Dr. Sebastian Wagner. We have many interesting and instructive discussions in my every panel meetings together with Corinna and Eduardo. They help me to ensure my thesis plan is well on the way, and they can always give insightful recommendations for the improvement of my thesis. It is my great honor to have you all as my strong backing.

Sincere gratitude to my group leader Dr. Birgit Hünicke, she is always there and always ready to fight for me whenever I was stuck in my routine work and study life. Since I am foreigner to pursue my PhD study in Germany, many unexpected things could be imagined. She helps me to overcome almost all these unexpected difficulties step by step, and she gives me strong spiritual supports to ensure me have a positive attitude to face these difficulties. I want to say thank you so much Birgit.

I would also like to say thanks to all my group members in KSI. No matter whom I ask for questions, they always give me prompt feedback and answers in a timely manner. In addition, many thanks to Prof. Dr. Emil Stanev and to my previous group KSD, thanks them for providing me with instructions to adjust myself to adapt to a foreign life when I first came to Hereon.

Sincere gratitude to every Hereon colleagues who ever give me firm help and support, and special thanks to our Chinese colleagues in Hereon. They all selflessly provide me with any help during my PhD study. They make me feel I am in a big family and they give me confidence that I could conquer any difficulties. Most important thing is that they help me to believe life and work is beautiful, hope and sunshine is immortal eternity.

I would like to thank the MPI-ESM modelling groups participating in the CMIP5 initiative for providing their data and the CESM modelling group for making their data available. This work is funded by the China Scholarship Council (no. 201806570017).

List of publications:

Zhang, Z., Stanev, E. V., and Grayek, S.: Reconstruction of the Basin-Wide Sea-Level Variability in the North Sea Using Coastal Data and Generative Adversarial Networks, J. Geophys. Res.-Oceans, 125, e2020JC016402, https://doi.org/10.1029/2020JC016402, 2020.

Zhang, Z., Wagner, S., Klockmann, M., and Zorita, E.: Evaluation of statistical climate reconstruction methods based on pseudoproxy experiments using linear and machine learning methods, Clim. Past Discuss. [preprint], https://doi.org/10.5194/cp-2022-5, *minor revision*, 2022.

Zhang, Z., Wagner, S., and Zorita, E.: Evaluation of the Bilinear Long-Short-Term-Memory and Echo State Network machine learning methods to reconstruct the Northern Hemisphere summer temperature, *(to be) submitted to Climate of the Past*.

Eidesstattliche Versicherung | Declaration on Oath

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

|

I hereby declare upon oath that I have written the present dissertation independently and have not used further resources and aids than those stated.

Ort, den | City, date Hamburg, 01.09.2022 Unterschrift | Signature

Ze guo Zhang