Past climate studies with optimized networks using artificial intelligence

Estudios del clima pasado con redes optimizadas usando inteligencia artificial

Tesis que presenta

Fernando Jaume Santero

para optar al título de

Doctor

Directores:

Dr. Ricardo García Herrera
Dr. David Barriopedro Cepero
Dr. Natalia Calvo Fernández

Departamento de Física de la Tierra y Astrofísica
Facultad de Ciencias Físicas
Universidad Complutense de Madrid
Instituto de Geociencias, CSIC-UCM

Madrid, 2021

UNIVERSIDAD COMPLUTENSE MADRID

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña. Fernando Jaume Santero,
estudiante en el Programa de Doctorado en Física,
de la Facultad de Ciencias Físicas de la Universidad Complutense de Madrid, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Estudios del clima pasado con redes optimizadas usando inteligencia artificial

Past climate studies with optimized networks using artificial intelligence

y dirigida por: Ricardo García Herrera, David Barriopedro Cepero y Natalia Calvo Fernández

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita.

Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada de conformidad con el ordenamiento jurídico vigente.

En Madrid, a 12 de marzo de 2021

Fdo.: Fernando Jaume Santero

Digitally signed by Fernando Jaume Santero
Date: 2021.03.12 07:35:00 +01'00'

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

- A la memoria de Francisco Jaume Aguiló y Lindsay Pickler,
os echamos de menos. -

There ain't no such thing as a free lunch.

Robert A. Heinlein

Acknowledgments

En primer lugar, tengo que agradecer a mis directores de tesis Ricardo García Herrera, David Barriopedro y Natalia Calvo por darme la oportunidad de realizar este increíble doctorado. Además, me gustaría agradecer a mis compañeros del grupo STREAM (y asociados): José Manuel, Antonio, Froila, Maddalen, Javier, Leopoldo, Sancho, Samuel, Solange, Marina, Verónica, Jacob, Marta, Carlos, Blanca, Álvaro, Adrián y Marie por los buenos momentos que pasamos juntos y las críticas constructivas que me han hecho mejor científico. Gracias a la gente de Canadá (Almudena, Francisco, Arlette, Ludovick, Lydia), a mis directores de máster Hugo Beltrami y Jean Claude Mareschal, al grupo PalMA (en especial a Fidel y Marisa) y a Victor Ocaña, porque sin ellos no me hubiera adentrado en el fascinante mundo de las ciencias de la Tierra. Dar las gracias también a Stefan Brönnimann, Jörg Franke y Raphael Neukom por la estancia que tuve el privilegio de hacer en la bonita ciudad de Berna. Mi reconocimiento al Ministerio de Ciencia e Innovación y al Ministerio de Universidades del Gobierno de España por su apoyo económico a través de la beca FPI (BES-2016-077030) del proyecto PALEOSTRAT (CGL2015-69699-R). Finalmente, me gustaría agradecer a mis amigos de la Universidad de Zaragoza: Sergio, Ignacio Hermoso, Ignacio Hierro, Edu de 1°, y Edu de 2°, a mi colega Nick, a la doctora Jenny Turton, y a todos mis familiares, especialmente a mi madre Marisa, a mi hermana Clara y a Pablo, a mis tíos Teresa, Pedro Pablo y Carlos, a mis primos Victoria, Gabriel, Lisa, Tolo y Miguel, a mi abuela María Teresa, a mi familia canadiense (Jocelyne, Andy, Marie-Ève y Bear) y a la jefa del mundo entero, la doctora Carolyne Pickler. Gracias por apoyarme y confiar en mí, sin vuestro apoyo esta aventura no hubiera sido posible.

List of Acronyms

20CRv3 ..... *NOAA-CIRES-DOE 20th Century Reanalysis version 3*

AH ..... *Azores High*

AI ..... *Artificial Intelligence*

AM ..... *Analogue Method*

AR1 ..... *lag-1 Autoregressive Model*

AR5 ..... *Fifth Assessment Report*

CCA ..... *Canonical Correlation Analysis*

CCSM4 ..... *Community Climate System Model version 4*

CE ..... *Common Era*

CESM ..... *Community Earth System Model*

CESM-LME .. *CESM Last Millennium Ensemble*

CFR ..... *Climate Field Reconstruction*

CMIP6 ..... *Coupled Model Intercomparison Project phase 6*

CRO ..... *Coral Reef Optimization*

CRO-AM ..... *Coral Reef Optimization coupled with the Analogue Method*

CRO-CCA .... *Coral Reef Optimization coupled with the Canonical Correlation Analysis*

CRO-MIN .... *Minimum subset of perfect pseudo-proxies obtained with CRO-AM necessary to outperform the reconstruction skill of the full-proxy network*

CRO-OPT .... *Optimized subset of perfect pseudo-proxies of the PAGES-2k network obtained with CRO-AM that yields the best reconstruction skill*

E-Obs ..... *European grid of Temperature Observations*

EOF ..... *Empirical Orthogonal Function*

FDR ..... *False Discovery Rate*

FSNTOAC ... *Clear-sky net solar flux at top of the atmosphere*

GCM ..... *General Circulation Models*

GHG ..... *Greenhouse Gases*

GISTEMP .... *Goddard Institute for Space Studies Surface Temperature*

GMT ..... *Global Mean Temperature*

HadCRUT .... *Hadley Centre/Climatic Research Unit Temperature*

IL ..... *Icelandic Low*

IPCC ..... *Intergovernmental Panel on Climate Change*

LIA ..... *Little Ice Age*

LMR ..... *Last Millennium Reanalysis*

MCA ..... *Medieval Climate Anomaly*

ML ..... *Machine Learning*

MSE ..... *Mean-Square Error*

NAM ..... *Northern Annual Mode*

NAO ..... *North Atlantic Oscillation*

PAGES-2k .... *Past Global Changes*

PC ..... *Principal Components*

PMIP4 ..... *Paleoclimate Modelling Intercomparison Project phase 4*

RMSE ..... *Root-Mean-Square Error*

SLP ..... *Sea Level Pressure*

SLP-Obs ..... *Set of SLP Observations*

SNR ..... *Signal to Noise Ratio*

SST ..... *Sea Surface Temperatures*

SVD ..... *Singular Value Decomposition*

Summary

The availability of high-quality climate records decreases backwards in time, and the associated increase in uncertainty supports the use of complementary sources of climate information (such as model simulations) to understand the underlying physics of the climate system, as well as its past and future changes. In this Ph.D. thesis we assess the potential of Artificial Intelligence as an additional efficient tool to solve complex problems in the field of climate sciences. We show that these techniques can optimize the information coming from different sets of climate networks such as meteorological stations, historical records, and paleoclimate archives. Being employed to address a plethora of questions, they share issues in terms of incompleteness. Within this framework, we address different problems that are common in the climate community by developing tailored methodologies with the same goal of maximizing the extraction of information from incomplete climate datasets. The developed approaches include metaheuristic algorithms and cluster analyses and will be applied to incomplete datasets that are typically employed for paleo-climate reconstructions and regional climate assessments, respectively.

Recent developments in metaheuristics have shown the efficiency of evolutionary algorithms to solve an extensive set of optimization problems. Specifically, the Coral Reef Optimization algorithm has provided robust solutions to high dimensional problems in atmospheric and climate sciences. Here, we combine this optimization algorithm with different climate field reconstruction methods to find an optimal subset of pseudo-proxies that minimizes

the spatial bias of annual temperature reconstructions induced by the non-homogeneous distribution of currently available records. The results indicate that under certain conditions small subsets of records situated over representative locations can outperform the reconstruction skill of the full network. These locations highlight the importance of high-latitude regions and major teleconnection areas to reconstruct annual global temperature fields for the last millennium and their simulated responses to external forcings and internal variability. However, low frequency temperature variations such as the transition between the Medieval Climate Anomaly and the Little Ice Age are better resolved by pseudo-proxies situated at lower latitudes. According to our idealized experiments a careful selection of proxy locations should be performed depending on the targeted time scale of the phenomenon to be reconstructed.

Moreover, we use the Coral Reef Optimization algorithm coupled with the Analogue Method to obtain a new high-resolution ($1^\circ \times 1^\circ$) reconstruction of monthly North Atlantic Sea Level Pressure fields since 1750. After assigning an optimized set of local weights to a network of land-based instrumental observations, the reconstruction skill is improved, particularly over poorly sampled regions, such as that under the influence of the Azores High. The reconstruction reproduces realistic variations of regional climate patterns such as the North Atlantic Oscillation and the Azores High as compared to other observational-based datasets. The results indicate that recent multi-decadal changes in the winter Azores High intensity have been the highest of the past 250 years, being concurrent with the prominent positive trend of the winter North Atlantic Oscillation from the 1960s to the 1990s. Moreover, differences in instrumental-based historical series of the winter North Atlantic Oscillation are partially explained by disparities on the reconstruction of the Azores High, rather than on the Iceland Low. Our findings also confirm the importance of the summer Azores High for the European climate. In particular, displacements of the Azores High center towards the north-east coincided with extremely warm summers in western Europe, as inferred from independent temperature reconstructions. Overall, the results suggest that

substantial improvements in the characterization of the past North Atlantic atmospheric variability could be achieved by reducing current uncertainties of the Azores High past behaviour.

Finally, a new clustering technique (known as k-gaps) has been designed to improve the clustering retrieved from traditional methods when applied to climate datasets with sparse records and/or incomplete temporal information. This method provides a new approach to cluster non-overlapping and discontinuous time series, therefore exploiting information that otherwise could be eliminated with data homogenization procedures. The method has been applied to European station-based daily temperature series since 1950 and validated with synthetic datasets. The results show that k-gaps performs well for limited networks in terms of both number of records and temporal availability. The algorithm can generate a climatically consistent clusterings similar to those obtained with complete time series, and outperforms other clustering methodologies developed to work with fragmentary information. Therefore, k-gaps can provide a useful tool for regional assessments of long-term trends and the detection of historical extreme events at regional scales.

Publications related to this Thesis

- **Jaume-Santero, F.**, Barriopedro, D., García-Herrera, R., Calvo, N. and Salcedo-Sanz, S. Selection of optimal proxy locations for temperature field reconstructions using evolutionary algorithms. *Sci. Rep.*, vol. **10**(1), ISSN 2045-2322, doi:10.1038/s41598-020-64459-6, 2020.

- Carro-Calvo, L., **Jaume-Santero, F.**, García-Herrera, R. and Salcedo-Sanz, S. k-Gaps: a novel technique for clustering incomplete climatological time series. *Theor. Appl. Climatol.*, ISSN 1434-4483, doi:10.1007/s00704-020-03396-w, 2020.

- **Jaume-Santero, F.**, Barriopedro, D., García-Herrera, R. and Luterbacher, J. Monthly North Atlantic Sea Level Pressure reconstruction back to 1750 CE using Artificial Intelligence optimization. *J. Clim*, (Submitted) 2021.

Other publications

- Cuesta-Valero, F. J., García-García, A., Beltrami, H., Zorita, E. and **Jaume-Santero, F.** Long-term Surface Temperature (LoST) database as a complement for GCM preindustrial simulations. *Clim. Past*, vol. **15**(3), pages 1099-1111, doi:10.5194/cp-15-1099-2019, 2019.

- Salcedo-Sanz, S., García-Herrera, R., Camacho-Gómez, C., Alexandre, E., Carro-Calvo, L. and **Jaume-Santero, F.** Near-optimal selection of representative measuring points for robust temperature field reconstruction with the cro-sl and analogue methods. *Glob. Planet. Change*, vol. **178**, pages 15-34, doi:10.1016/j.gloplacha.2019.04.013, 2019.

- 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*, vol. **16**(3), pages 1061-1074, doi:10.5194/cp-16-1061-2020, 2020.

Resumen

La disponibilidad de datos climáticos decrece exponencialmente a medida que retrocedemos en el tiempo, siendo muchas veces necesario el uso de fuentes complementarias de información (como las simulaciones de modelos de circulación general) para comprender la física subyacente del sistema climático, así como sus cambios pasados y futuros. En esta tesis doctoral evaluamos el potencial de la Inteligencia Artificial como una herramienta eficiente que se puede usar para resolver problemas complejos en la ciencia del clima. Mostramos como estas técnicas pueden maximizar la información proveniente de diferentes conjuntos de redes climáticas, como estaciones meteorológicas, registros históricos y *proxies* paleoclimáticos. Todos ellos comparten un problema similar: son datos incompletos que proporcionan información por un periodo de tiempo limitado. Por lo tanto, hemos abordado diferentes problemas cuyo objetivo común es maximizar la extracción de información de conjuntos de datos incompletos. Los métodos desarrollados incluyen algoritmos metaheurísticos y análisis de conglomerados.

Recientes avances en metaheurística han demostrado la eficiencia de los algoritmos evolutivos para resolver diferentes problemas de optimización. Específicamente, el algoritmo de *Coral Reef Optimization* ha proporcionado soluciones robustas a problemas de alta dimensionalidad en ciencias de la Tierra y del clima. En esta tesis, combinamos este algoritmo de optimización con diferentes métodos de reconstrucción climática para minimizar el sesgo espacial inducido por la distribución no homogénea de datos paleoclimáticos. Los resultados indican que subconjuntos de series situadas en zonas claves

pueden mejorar la reconstrucción obtenida con la red completa de datos en paleo-reconstrucciones del último milenio. Se ha evidenciado la importancia de las altas latitudes y áreas de teleconexión para reconstruir campos de temperatura global anual y sus respuestas a los forzamientos externos y variabilidad interna del sistema. Sin embargo, las variaciones de temperatura a largo plazo, como la transición entre la Anomalía Climática Medieval y la Pequeña Edad de Hielo, se resuelven mejor con registros situados en latitudes más bajas. Nuestros experimentos indican que se debe realizar una selección de ubicaciones representativas en función de la escala del fenómeno investigado.

Siguiendo con la misma metodología, utilizamos el *Coral Reef Optimization* junto con el Método de Análogos para obtener una nueva reconstrucción en alta resolución ($1° \times 1°$) de campos mensuales de Presión a Nivel del Mar sobre el Atlántico Norte desde 1750. Estos campos han sido generados a partir de una red optimizada de observaciones terrestres. Se ha comprobado como la técnica de optimización maximiza la robustez de la reconstrucción de los campos de presión en toda la región de estudio, incluso cuando solo hay unas pocas observaciones disponibles, reproduciendo variaciones realistas de los patrones climáticos regionales como la Oscilación del Atlántico Norte, y el Anticiclón de las Azores desde mediados del siglo XVIII. Esta reconstrucción optimizada muestra una tendencia positiva en la intensidad del Anticiclón de las Azores en invierno, siendo la tendencia decenal más alta de los últimos 250 años. Este cambio coincide con la tendencia positiva prominente de la Oscilación del Atlántico Norte de invierno desde la década de 1960 hasta la de 1990. Además, también encontramos que las diferencias en las reconstrucciones de la Oscilación del Atlántico Norte se explican en parte por las diferencias en la reconstrucción del Anticiclón de las Azores, destacando la importancia de reconstruir este sistema persistente de alta presión para reproducir el clima de la región. Esta reconstrucción también muestra que los desplazamientos de este anticiclón hacia el noreste provocaron eventos de calentamiento extremo en Europa Occidental recogidos por fuentes de temperatura independientes. Por tanto, los resultados sugieren que se podría

mejorar sustancialmente la caracterización y estudio de la variabilidad atmosférica del Atlántico Norte en el pasado, reduciendo las incertidumbres a la hora de reconstruir el Anticiclón de las Azores.

Finalmente, hemos desarrollado una nueva técnica de conglomerados (denominada k-gaps) con el objetivo de mejorar los analisis de climas regionales utilizando conjuntos incompletos de datos climáticos. Este método proporciona un nuevo enfoque para agrupar series de tiempo de diferentes longitudes temporales, utilizando la mayor parte de la información recogida en conjuntos de series climáticas, que muchas veces se elimina durante los procesos de homogeneización de datos. El método se ha aplicado a series de temperaturas diarias medidas en estaciones europeas desde 1950 y se ha validado con conjuntos de datos sintéticos. Los resultados muestran que k-gaps es ideal para el agrupamiento de pequeños conjuntos de datos climáticos (con pocas muestras) y con huecos en las series temporales. El algoritmo puede generar una regionalización climáticamente consistente similar a las obtenidas con series de tiempo completas, superando a otras técnicas similares desarrolladas para trabajar con falta de información. Por lo tanto, k-gaps pueden ser una herramienta útil para los análisis regionales de tendencias climáticas a largo plazo y la detección de eventos extremos históricos a escalas regionales.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In 1950, Alan M. Turing wrote for the first time about computing and intelligence (Turing, 1950). Since then, advances in new technologies have led to a revolution in areas such as telecommunications, computer science, and automation that have completely redefined the way we interact with the world (Hodson, 2018). From the way we learn to the way we work, almost every aspect of our lives is influenced by innovative developments that improve our well-being conditions, increase our productivity, or connect us in a vast network that facilitates human communication, providing global access to an (almost) infinite source of knowledge. From an engineering point of view where the world is conceived as a set of problems, new technologies are conceived as optimized solutions to everyday problems.

Optimization is therefore a key component of many fields in social (Rowland, 1946; Ballings et al., 2016), natural (Kell, 2012; Tchemisova et al., 2015), and physical (Bounds, 1987; Vadlamani et al., 2020) sciences. Here is where advanced statistical techniques and fast computational power unite to create what it is nowadays known as AI (*Artificial Intelligence*), a multidisciplinary field with different areas of expertise such as robotics (Rajan and Saffiotti, 2017), machine learning (LeCun et al., 2015), and optimization (Swarnkar and Swarnkar, 2019; Soto et al., 2019), that have shown their proficiency to master high-dimensional and non-linear problems beyond human

capabilities, sometimes without any kind of supervision (Silver et al., 2016, 2018). In this sense, technological advances in the last few decades have transformed past Alan Turing's ideas into practical solutions to problems of the modern world (Kates-Harbeck et al., 2019; Ho, 2020; Jin et al., 2020).

Nevertheless, modernization has also come at an expensive cost for our planet. Since 1950, the great acceleration of socio-economic trends (e.g. Steffen et al. (2015); world population, energy and water use, transportation, and urban construction, among many others) has exerted an important impact on the natural environment (IPCC, 2014). These activities mainly boosted by fossil-fuel consumption caused a global radiative energy imbalance and major changes in the climate system as shown by many studies (e.g. Hansen et al., 2011; Huber and Knutti, 2012) scrutinized by periodic reports of the IPCC (*Intergovernmental Panel on Climate Change*) (Myhre et al., 2013). It is also clear that these changes in Earth's system have been primarily caused by increasing concentrations of atmospheric GHG (*Greenhouse Gases*) such as carbon dioxide, nitrous oxide, and methane attributed to human activities (Ribes et al., 2017). Climate projections in the AR5 (*Fifth Assessment Report*) of the IPCC obtained from state-of-the-art GCM (*General Circulation Models*) (e.g. Flato et al., 2013; Hausfather et al., 2020) indicate that global mean temperatures will increase up to intolerable levels by the second half of the 21st century if no mitigation policies are taken in the next few years (e.g. Collins et al., 2013; IPCC, 2018). The implications of anthropogenic changes on the climate system are so undeniable for our society (Sanderson and O'Neill, 2020) that *United Nations* has characterized them as one of the most pressing issues of our time.

Within this framework, and taking into account the last advances in computerization, it seems quite reasonable to make use of our acquired technical background to address climate-related problems that can contribute to this challenge through an improved characterization of the complex climate system.

1.1 Artificial Intelligence as a tool for Climate Science

The understanding of our own planet has notoriously increased in the last decades with the implementation of systems that allow for the continuous monitorization of the Earth system (i.e. atmosphere, ocean, land, etc). For instance, the exponential accumulation of information obtained from ocean and land-based stations, observation satellites, and model simulations (Agapiou, 2017) has expanded the ways to broad our knowledge about the dynamics of the climate system, non-linear interactions and responses to external forcings, as well as their roles in shaping the climate conditions at regional and global scales (e.g. IPCC, 2013).

Despite these unquestionable benefits, the generation of large amounts of observational and simulated datasets comes with a problem of big data where expensive storage infrastructures and fast supercomputers are required in advance to perform robust climate analyses (e.g. Schnase et al., 2016). In this regard, ML (*Machine Learning*) techniques have been used as a tool to manage large datasets by implementing massive data processing into daily research (Knüsel, 2019) and tackle complex problems with data-driven approaches (Karpatne et al., 2019; Reichstein, 2019). They combine advanced statistical techniques and fast computational power to generate self-learning systems that are trained with large amounts of data without being programmed for any specific task (e.g. Murphy, 2012). Some of the problems in climate sciences and meteorology that have been addressed using these procedures are: climate sensitivity prediction (Caldwell et al., 2014), statistical downscaling (He et al., 2016), remote sensing (Maxwell et al., 2018), tropical cyclone forecasting (Richman et al., 2017), soil moisture prediction (Prakash et al., 2018), cloud physics representation (Rasp et al., 2018), land use / land change estimations (Aburas et al., 2019) pattern recognition (Boers et al., 2019), climate change adaptation (Biesbroek et al., 2020), and model parameterizations (Gagne II et al., 2020). All these contributions were pos-

sible thanks to plentiful of data available in public and private repositories ready to be analyzed. However, there are areas of scientific research where observational data are scarce and/or can only be retrieved from expensive measuring campaigns and therefore they become a scarce commodity. Such is the case of past climate reconstructions (Masson-Delmotte et al., 2013; Emile-Geay et al., 2017) which rely on instrumental observations and paleoclimate records (proxies), respectively, with limited temporal coverage and uneven distribution over the globe. In these cases, it is not possible to apply data-driven techniques, and solutions must then focus on maximizing the extraction of information from a limited network of records. Solutions to this kind of optimization problems have recently been developed using other branches of AI such as metaheuristic algorithms (Salcedo-Sanz, 2016; Del Ser et al., 2019) and cluster analysis (Kettenring, 2006; Omran et al., 2007; Netzel and Stepinski, 2016).

In computer science, metaheuristic procedures such as evolutionary and genetic algorithms use different search engines to find "good enough" solutions to optimization problems (Yang et al., 2014; Abdel-Basset et al., 2018). They are especially useful with datasets containing incomplete information and high-dimensional combinatorial problems where discrete optimal solutions are required. Although previous studies have already unveiled some of the potential that these techniques have in several fields of geosciences such as hydrology (Yoo and Kim, 2014), soil stabilization (Kashani et al., 2016), and wind power reconstruction (Salcedo-Sanz et al., 2018), they still remain underexploited by the climate community (e.g. Knüsel, 2019; Reichstein, 2019; Kadow et al., 2020).

On the other hand, cluster analyses are unsupervised classification techniques able to organize datasets with heterogeneous information into groups with similar features. This property makes them suitable for the study of different fields in Earth science such as geochemestry (Zhou et al., 2018), geophysics, (Song et al., 2010), seismology (Seydoux et al., 2020), and climatology (Netzel and Stepinski, 2016). However most clustering algorithms

require datasets with complete information, limiting the analysis to records with continuous and simultaneous observations, or forcing data homogeneization (i.e, series are truncated and/or interpolated to avoid gaps). While these are not big issues when there is a fair amount of available data, it might lead to substantial loss of information for small networks of sparse records with limited data availability such as historical observations and paleoclimate archives. This stresses the need to develop new cluster methodologies that maximize the extraction of information from incomplete climate datasets while minimizing the loss of information by homogenization procedures.

1.2 Main objectives and structure of the Thesis

Finding solutions to the aforementioned problems need methodologies that provide some sort of information maximization or error minimization, indicating that they can be defined as optimization problems. Hence, the main goal of this thesis is to develop techniques that maximize the extraction of climate information contained in sample-starved datasets, testing the potential and effectiveness of different AI systems for this task. Although sometimes these techniques are seen as black-box models without any physical background (Rudin and Radin, 2019), previous studies have argued that their predictive skill is not only based on correlations but also causality (Pietsch, 2016). In this thesis we will then show that they can be combined with methodologies commonly used in climate science to solve high-dimensional and non-linear problems. The main goal of this thesis has therefore been divided into two well-defined objectives, aiming to address different problems that are frequently found in observational climate datasets, namely: the improvement of spatially-resolved climate field reconstructions obtained from sets of climate networks, and spatial clustering of datasets with incomplete sets of records. The specific nature of the problem varies with the type of targeted information and the characteristics of the observables, therefore requiring tailored developments, which will be dissected in dedicated chapters, accompanied by their respective backgrounds.

The methods are relevant for the scientific community because climate datasets are obtained from meteorological stations and measuring campaigns whose elevated costs impede to have a full coverage of the study region. This leads to a problem of data scarcity and a non-homogeneous distribution of records that can debase global and regional climate reconstructions, increasing the uncertainty of different history fields the further they go back in time. In this regard, the key resides not only in extracting the maximum information possible from each record, but also in selecting the representative areas that should be measured to maximize the reconstruction skill of a given observable with limited funding, a task that can also be tackled with these techniques.

1.2.1 Thesis structure

This thesis is divided into seven chapters. After the introduction, Chapter 2 describes the datasets and it has been divided into two sections: Data description (Section 2.1) and Data post-processing (Section 2.2). Chapter 3 details the climate and AI tools employed in the subsequent chapters of the thesis, including climate reconstruction methods (Section 3.1), metaheuristic algorithms (Section 3.2), and clustering techniques (Section 3.3) that will be applied to different observational datasets with incomplete information.

The next three chapters are the main core of the thesis. Chapter 4combines metaheuristic algorithms and reconstruction methods to find an optimal subset of locations within a global pseudo-proxy network that reduces the spatial bias of annual temperature reconstructions of the last millennium. The main results of this chapter can be found in Jaume-Santero et al. (2020).

In Chapter 5, a metaheuristic approach is again used to obtain a high-resolution reconstruction of monthly North Atlantic SLP (*Sea Level Pressure*) for the past two and half centuries using optimized networks of land

observations. The main results are under review at the moment of writing this thesis (Jaume-Santero et al., Submitted, 2021).

On the other hand, Chapter 6 describes the development of a new clustering technique (k-gaps) aiming to generate a robust regional analysis of a given observable using climate datasets with incomplete information in space and time. The algorithm is applied to daily European temperature series for the second half of the 20^{th} century, and tested with synthetic series in order to retrieve a regional classification of mean and extreme conditions. Carro-Calvo et al. (2020) contains the most relevant results of this chapter.

Finally, Chapter 7 summarizes the main conclusions extracted from this thesis as well as the outlook of future research.

Chapter 2

Data

In this chapter we detail all datasets employed to obtain the results shown in the thesis. It has been divided into two sections: Data description and Data post-processing. Section 2.1 describes the original datasets utilized in the study. They are classified into four different subsections depending on their data type: Model simulations (Subsection 2.1.1), Observations (Subsection 2.1.2), Reanalysis (Subsection 2.1.3), and Paleoclimate archives (Subsection 2.1.4). However, some climate datasets need to be post-processed so that they can be properly used in the experiments. Therefore, in Section 2.2 we describe the data processing necessary to generate modified versions of the aforementioned datasets such as pseudo proxy records from model simulations matching the locations of real paleoclimate archives (Subsection 2.2.1) and pseudo observations of SLP (Subsection 2.2.2).

On the other hand, in Chapter 6 the validation of new clutering techniques for the study of regional climates required the generation of 500 synthetic datasets with time series of daily summer European temperatures with high levels of missing data irregularly distributed over the region. For readability purposes, the process followed to obtain them can be found in the experiment setting (Section 6.3) of that chapter.

2.1 Data description

Choosing datasets for research purposes is never as straightforward as one could think in a first instance. Due to the exponential increasing capabilities of computerization in the last few decades, together with the vast development of telecommunications, it has never been easier to share, exchange, and download scientific data. Climate products such as reanalysis, climate reconstructions, and simulations are known to get more and more accurate (more realistic) with time, improving our understanding of Earth's climate system. However, quality comes at a price. On one side, increases in temporal and spatial resolutions of climate simulations have led to a high volume of output files (big data problem) which are difficult to manage with normal desktop computers. On the other there is a scarcity of instrumental observations prior to the 20$^{\text{th}}$ century, leading to an increase of uncertainty in the reconstruction of the climate of the past. It is therefore important to combine different types of datasets to provide more robust climate insights. Table 2.1 shows relevant information about the datasets employed in the subsequent chapters.

Table 2.1: List of datasets used in the experiments of the thesis.

Name	**Version**	**Period (CE)**	**Time res.**	**Spatial res.**	**Reference**
Model simulations					
CESM-LME	CESM 1.1.2	850-2005	Monthly	1.9° × 2.5°	Otto-Bliesner (2016)
CCC400	ECHAM5.4	1601-2005	Monthly	2° × 2°	Franke et al. (2017)
MRI-ESM2-0	2.0	850-2014	Daily	1° × 1°	Yukimoto et al. (2019)
Observations					
HadCRUT	4.2	1850-2014	Monthly	5° × 5°	Jones et al. (1999)
SLP-Obs	1.0	1750-2004	Monthly	121 series	Küttel et al. (2010)
E-Obs	14.0	1950-2016	Daily	0.25° × 0.25°	Haylock et al. (2008)
Reanalysis					
LMR	2.0	850-2000	Yearly	4.3° × 5.7°	Tardif et al. (2019)
20CR	3.0	1836-2014	Monthly	1° × 1°	Slivinski et al. (2019)
Paleoclimate archives					
PAGES-2k	2.0.0	0-2000	Varying*	692 series	Emile-Geay et al. (2017)

* Proxy records have temporal resolutions that span from seasonal (e.g. tree-rings) to multidecadal (e.g. borehole temperature profiles).

2.1.1 Model simulations

Climate simulations employed in this thesis have been extracted from different GCM outputs. GCMs are numerical models describing the physical processes within Earth's climate subsystems such as the atmosphere, the ocean, the cryosphere, and the land surface.

In Chaper 4 we employ the history fields from two different simulation ensembles: the CESM-LME (*CESM Last Millennium Ensemble*) and the CCC400. The CESM-LME Project(Otto-Bliesner, 2016) has released 36 last millennium simulations for 850-2005 CE (*Common Era*) from NCAR's CESM (*Community Earth System Model*) 1.1.2 GCM, 13 of them including all transient forcings (solar radiation, volcanic aerosols, greenhouse gases, land use/land cover conditions and orbital parameters). The remaining runs are ensembles of single-forcing simulations for the last millennium with only one transient forcing (either solar, volcanic, greenhouse gases, land use/land cover or ozone) and one control simulation (with forcings fixed at 1850 CE). Annual mean air temperature fields at 2-m (TREFHT) and $1.9° \times 2.5°$ spatial resolution for the 850-2005 CE period of the full-forcing CESM-LME have been used as testbed to carry out the experiments with pseudo-proxies (see Subsection 2.2.1). Moreover, the clear-sky net solar flux at the top of the atmosphere (FSNTOAC, in Wm^{-2}) from the control and single-forcing simulations have also been used to identify the dominant forcing for each year of the last millennium. In addition, SLP fields from the full-forcing simulations are employed to compute the Northern Annual Mode, defined as the first empirical orthogonal function of annual mean area-weighted SLP anomalies for [20-90] °N and 850-2005 CE.

To test the independence of the results with respect to the model employed in Chapter 4, experiments are performed with 2-m air temperature fields from the CCC400, an ECHAM5.4 model ensemble (Bhend et al., 2012; Franke et al., 2017) composed of 30 fully forced members for the period 1601-2005 CE at $2° \times 2°$ spatial resolution.

In Chapter 5, we have used last millennium (850-1849 CE) and historical (1850-2014 CE) outputs of SLP from the MRI-ESM2-0 model (Yukimoto et al., 2019) to verify that the results of the optimization process are robust with respect to the reference dataset. The Japanese model was selected because its grid resolution is similar to the 20CRv3 reanalysis but it comes from an independent branch of the model genealogy tree, guaranteeing the consistency of the results from different data sources. History fields of $1^\circ \times 1^\circ$ have been extracted from the *r1i1p1f1* realization which has been run following the full-forcing specifications of the CMIP6 (*Coupled Model Intercomparison Project phase 6*) and PMIP4 (*Paleoclimate Modelling Intercomparison Project phase 4*) experiments (Eyring et al., 2016; Jungclaus et al., 2017).

2.1.2 Observations

In Chapter 4 we have used the HadCRUT (*Hadley Centre/Climatic Research Unit Temperature*) 4.2, a monthly global dataset of gridded temperature series obtained after combining SST (*Sea Surface Temperatures*) records from the Hadley Center (HadSST3) and land surface temperatures (CRUTEM4) from the Climatic Research Unit of the University of East Anglia(Jones et al., 1999). The grid has a spatial resolution of $5^\circ \times 5^\circ$ and spans over the period 1850-2014 CE.

The experiments performed in Chapter 5 employ the largest currently available network of quality-checked SLP observations SLP-Obs (*Set of SLP Observations*) used by Luterbacher et al. (2002) and Küttel et al. (2010). It consists of 121 monthly series of SLP with different time lengths within the 1750-2004 CE period, distributed over the east coast of North America, Greenland and Europe. However, after screening we rejected 20 records (most of them situated in weather stations over the Alps) that did not meet the standard quality as shown in Table 2.2. The accepted series are shown in Table B.1.

Table 2.2: List of rejected observations from SLP-Obs database.

Name	**Latitude(DD)**	**Longitude (DD)**	**Start (CE)**	**End (CE)**
Bad Ischl	47.72	13.63	1854	2003
Basel	47.60	7.60	1754	2004
Geneva	46.30	6.10	1767	2004
Graz	47.07	15.45	1836	2004
Gr. St. Bernhard	45.50	7.10	1863	2004
Hohenpeissenb	47.80	11.00	1780	2004
Innsbruck	47.27	11.40	1829	2004
Jerusalem	31.80	35.20	1860	2003
Karlsruhe	49.01	8.39	1869	2004
Klagenfurt	46.70	14.30	1843	2004
Kremsmunster	48.05	14.13	1821	2004
Lugano	46.00	8.60	1863	2004
Munich	48.10	11.70	1824	2004
Neuchatel	46.60	6.60	1863	2004
Santis	47.20	9.20	1882	2004
Salzburg	47.80	13.03	1841	2004
Sonnblick	47.01	12.95	1886	2004
Vienna	48.30	16.40	1774	2004
Zurich	47.37	8.55	1863	2004
Po Plain*	45.12	9.66	1765	2004

* In the same grid-point, Milan has higher correlation (0.96 vs 0.78) with respect to the 20CRv3 reanalysis during the 1836-2004 CE time period, and it is a longer record.

Finally, in Chapter 6 we use the E-Obs (*European grid of Temperature Observations*) (version 14.0) (Haylock et al., 2008). The E-Obs dataset provides daily spatially resolved European field temperatures with a spatial resolution of $0.25° \times 0.25°$. However, due to the presence of missing values, locations with less than 6000 days were removed, as well as time periods when any of the remaining data points presented missing values. Hence, there were enough time series to generate a complete set (in space and time) of 17,452 grid points with 5569 summer day mean temperatures contained between latitudes 35° S and 72° N, and longitudes 20° W and 42° E, for a time range of 66 years since 1950 CE.

2.1.3 Reanalysis

Climate reanalyses combine instrumental observations with climate models to reproduce realistic grids of history fields such as air temperature, pressure and wind fields at different pressure levels, as well as mono-level variables at the surface (e.g. SST and SLP). By prescribing real boundary conditions to the model and assimilating certain records from instrumental observations, historical, and paleoclimate archives, they provide hybrid products with continuous data, full spatial coverage, high-resolution, and physical consistency.

Two reanalyses have been used. Chapter 4 introduces the LMR (*Last Millennium Reanalysis*), a $4.3° \times 5.7°$ proxy-based annual temperature reconstruction for 850-2000 CE based on the assimilation of 2892 paleoclimate records from tree-rings, lake core, ice core, coral and speleothem archives (Hakim et al., 2016; Tardif et al., 2019). The LMR uses an ensemble data filter that estimates a prior state vector from CCSM4 (*Community Climate System Model version 4*) outputs, later modified by the assimilation of proxy records and subsequently calibrated by the GISTEMP (*Goddard Institute for Space Studies Surface Temperature*).

On the other hand, monthly SLP fields from the 20CRv3 (*NOAA-CIRES-DOE 20th Century Reanalysis version 3*) are employed in Chapter 5. The 20CRv3 reanalysis (Slivinski et al., 2019) provides $1° \times 1°$ history fields from 1836 to 2014 CE. The reanalysis assimilates observations of surface pressure and prescribes sea surface temperatures and sea ice distribution to estimate the remaining climate variables which are publicly available and can be downloaded at NOAA's website. Despite the presence of uncertainties in this reanalysis and the SLP-Obs, the combined use of instrumental and reanalyzed observations through AI lens (Barnes et al., 2019) can help to identify inconsistencies in the datasets and maximize the performance of the CFR by exploiting robust relationships between the local series and the large-scale field (see Chapter 5).

2.1.4 Paleoclimate data

Decades of scientific fieldwork have left abundant proxy datasets to reconstruct the climate of the past. They provide information about climate changes at local and regional scales, and have been pooled to derive spatially-resolved climate reconstructions of the last millennium (Crowley, 2000; 2k Consortium, 2013). Due to the increasing availability of high-resolution records, the initiative PAGES-2k (*Past Global Changes*) has scrutinized thousands of temperature-sensitive proxies, releasing a global archive with paleoclimate records for the last two millennia (Emile-Geay et al., 2017). This database has been used to assimilate climate information to reconstruct annual temperature fields (Franke et al., 2020), as well as in the aforementioned LMR (Tardif et al., 2019). Moreover, in Chapter 4, we have used the locations of proxy records contained in the database to obtain the set of optimal proxy locations that best reconstructs 2-meter air temperature fields for the period 850-2005 CE (Jaume-Santero et al., 2020).

2.2 Data post-processing

Some datasets had to be post-processed to obtain the required files for the experiments included in this thesis. For instance, pseudo-proxies (synthetic temperature series) were generated in Chapter 4 by perturbing 2-m air temperatures from GCM simulations (as described in Subsection 2.2.1), emulating the characteristics of real-world paleoclimate archives. These pseudo-proxies served us as a testbed to carry out the experiments under a controlled framework. Within this context, pseudo-observations of SLP with realistic levels of noise and time length were made for the experiments shown in Chapter 5, with the aim of mimicking the features of the SLP-Obs dataset in series of sea level pressure extracted from outputs of the MRI-ESM2-0 model (see Subsection 2.2.2).

2.2.1 Pseudo proxies

Three types of pseudo-proxies have been constructed in Chapter 4 at the locations of the PAGES-2k multi-proxy database from the annual mean 2-meter air temperature of the first CESM-LME full-forcing member and the first member of the CCC400 ensemble. They include perfect proxies and more realistic pseudo-proxies (Smerdon, 2011; Gómez-Navarro et al., 2017; Neukom et al., 2018) derived by adding red noise to the temperature series using AR1 (*lag-1 Autoregressive Model*) autocorrelation. Different levels of SNR (*Signal to Noise Ratio*) have been tested, including values such as infinite (perfect pseudo-proxies), 1 and 0.5. Note that pseudo-proxies are only sensitive to temperature and the observational availability is complete for the entire period.

Following the methodology of Neukom et al. (2018), pseudo proxies, $\mathbf{P}(t)$, have been synthetized for each time-step, t, from the temperatures series, $\mathbf{T}(t)$, using the model provided by Eq. 2.1.

$$\mathbf{P}(t) = \mathbf{T}(t) + \mathbf{n}(t), \tag{2.1}$$

where $\mathbf{n}(t)$ is red noise iteratively defined by Eq. 2.2

$$\mathbf{n}(t) = \gamma \cdot \mathbf{n}(t-1) + \delta \tag{2.2}$$

where γ is the AR1 autocorrelation coefficient for each pseudo-proxy, and δ is generated as white noise with zero mean and variance, $\sigma^2(\delta)$, given by Eq. 2.3.

$$\sigma^2(t) = \frac{(1-\gamma^2) \cdot \sigma_T^2}{\mathrm{SNR}^2} \tag{2.3}$$

where σ_T^2 is the variance for each temperature series. This parameterization implies that noisy pseudo-proxies are auto-correlated, and the noise variance has the same amplitude as the variance of the climate signal.

2.2.2 Pseudo SLP observations

Pseudo-observations of SLP have been generated in Chapter 5 from MRI-ESM2-0 model outputs (Yukimoto et al., 2019) by extracting 101 time series at the locations of SLP-Obs (only those that were accepted after screening). These series have subsequently been perturbed to match the SLP-Obs database in terms of data quality and missing values. Complete model SLP series from 1750 to 2004 CE were therefore truncated one by one to have the same time length than the real observations. Moreover, to account for imprecisions in meteorological instruments, they were perturbed with white noise so that each series had the same correlation with the unperturbed (perfect) series of the model, as the correlation of real observations with the 20CRv3 reanalysis.

Realistic noise was added to the series by defining a SNR following Eq. 2.4

$$\mathrm{SNR} = \sqrt{\frac{r^2}{1 - r^2}} \tag{2.4}$$

where r is the Pearson correlation between a real SLP time series and the SLP of the 20CRv3 reanalysis extracted at the same location as the weather station where that series was recorded. Then pseudo-observations were generated by following the steps described for the pseudo-proxies of Subsection 2.2.1 (Eqs. 2.1, 2.2, and 2.3). Note that as in this case white noise has been added, the autocorrelation coefficient (γ) is zero.

Chapter 3

Methodology

Section 3.1 details the methods employed for the reconstruction of climate fields also known as CFR (*Climate Field Reconstruction*) such as temperature and SLP. CFRs are spatially-resolved reconstructions of climate variables generated from different sources of climate information such as instrumental observations, historical documents, and paleoclimate archives. The two independent CFR techniques employed in the studies carried out in Chapters 4 and 5 are the AM (*Analogue Method*) (Subsection 3.1.1) and the CCA (*Canonical Correlation Analysis*) (Subsection 3.1.2).

Subsequently Section 3.2 introduces the definition and main uses of an evolutionary algorithm known as the CRO (*Coral Reef Optimization*). Evolutionary algorithms (Eiben and Smith, 2015) are a branch of Artificial Intelligence based on the optimization of high dimensional (and non-linear) systems (Knüsel, 2019) using algorithms biologically-inspired by processes that imitate the evolution and survival of best adapted individuals under ever-changing environmental conditions. The iterative technique that we used for the optimization of sampling networks is known as the CRO, an hybrid-type evolutionary algorithm (Salcedo-Sanz, 2017) that simulates the reproduction and evolution of corals within a reef with a limited number of latching spots (more details in Section 3.2). In our case, we coupled this algorithm with the two CFR methods of Section 3.1 to assess whether it

was possible to improve the skill of spatially-resolved reconstructions by optimizing networks of climate records available on public repositories. Two different approaches were followed to tackle this: the first one was performed by finding the optimal subset of representative records that best reconstruct the climate field, and the second was focused on searching for optimal sets of weights that applied over climate records generate the reconstruction with the highest skill possible.

Lastly, a description of clustering techniques and their use on climate research is presented in Section 3.3. These iterative and self-organizing methods associate data objects into sets whose individuals tend to share more similarities among them than with individuals of other sets (Hartigan and Wong, 1979; Phillips, 2002). Although these techniques are nowadays widely employed in climate research, they usually require complete datasets with homogeneous information, forcing the truncation of longer time series with the subsequently loss of information. Hence, due the necessity of maximizing data contained within climate records, a novel clustering method has been developed (Carro-Calvo et al., 2020) to cluster sets with incomplete climatological time series, allowing for the study of past climate variations beyond the capabilities of classical techniques.

3.1 Climate Field Reconstruction methods

There are multiple reconstruction methods used within the paleoclimate community to spatially resolve climate history fields. Although most of them generate robust reconstructions, they use different approaches, showing significant differences in terms of computational performance. We have selected two CFR techniques among many other reconstruction methods for their high reconstruction skill and fast performance, being the latter essential for the general purpose of this thesis: finding optimal solutions to high dimensional problems. Note that optimization of any sort requires to process the information as fast as possible, especially in soft-computing where evolutionary

algorithms have to generate millions of solutions and check their suitability in a limited amount of time. Here we show how the AM (Gómez-Navarro et al., 2017) and CCA (Smerdon et al., 2010) methods work, the two fastest reconstruction techniques on record.

3.1.1 Analogue Method

The AM reconstructs an unknown target field from an available reference dataset by searching analogues, defined as the spatially resolved fields with the largest resemblance to the the target variable over the observing network (Lorenz, 1969; Franke et al., 2011; Gómez-Navarro et al., 2015; Talento et al., 2019; Bothe and Zorita, 2020). As such, it does not require calibration, but a large pool of spatially resolved fields in the reference dataset to guarantee an enough number of good analogues. The proximity of the pooled fields to observations is measured with a distance metric. To account for the diversity of large-scale fields that are compatible with the distribution of records over the limited network, a minimum number N of analogues is predefined, whose average yields the reconstructed field. AM reconstructions often use constant N values (the best N analogues of the observable for each time step) but there is no objective criterion to define an optimal number of analogues, and therefore the choice is based on the reconstruction skill.

In Chapter 4, global annual temperature fields for the 850-2005 CE period have been reconstructed using the AM, which provides spatially resolved global fields of temperature from the limited sample of pseudo-proxy records, matching the locations from the PAGES-2k archive. In this case, the target field to reconstruct is the annual mean air temperature at 2 meters (TREFHT) of the first CESM-LME full-forcing member. For each year, we sorted (from best to worst) analogues of the target simulation from a pool formed by the remaining 12 last millennium members of the full-forcing ensemble. Best analogues are the years of the pool with minimum RMSE (*Root-Mean-Square Error*) for the pseudo-proxy locations. For each target year, the global temperature patterns of the best analogues are averaged at each grid point and

taken as the reconstructed global field. The correlation between the reconstructed and target fields increases with the number of analogues employed but, at the same time, the variability ratio decreases as seen in Fig. 3.1. Accordingly, we retained the three best analogues of each year, which is similar to those used in previous studies (Gómez-Navarro et al., 2017).

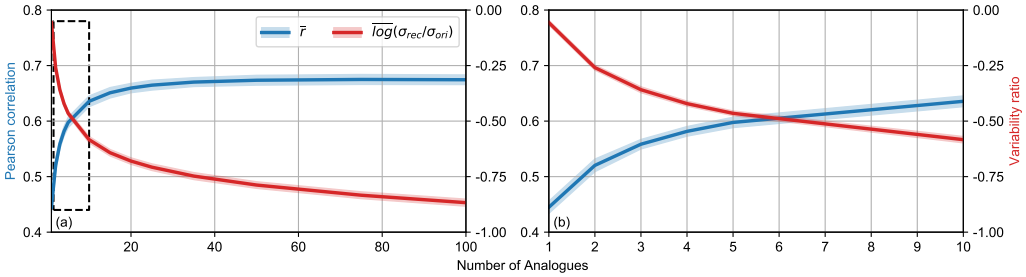

Figure 3.1: Pearson correlation (blue) and variability ratio (red) for AM reconstructions of global temperature fields as a function of the number of analogues. For each member of the CESM-LME the remaining 12 full-forcing simulations are used to reconstruct the global temperature fields from the full-proxy PAGES-2k network of perfect pseudo-proxies, using different number of analogues. Shading shows the spread (two standard deviations with respect to the mean values). (a) Correlation and variability ratio for AM reconstructions with 1 to 100 analogues. (b) A zoom of the black dashed square in (a).

The sensitivity of the AM to proxy weighting was also tested by using area-weighted and un-weighted fields before the reconstruction procedure. Similar results were found as seen in Fig. A.1, and hence no area weighting was applied, for coherence with (Gómez-Navarro et al., 2017). The AM was also employed to derive NAM (*Northern Annual Mode*) reconstructions for the first ensemble member from the temperature field over the pseudo-proxy locations. The reconstructed NAM is retrieved separately for each year by randomly picking the SLP field of one of the 100 best analogue years of the target temperature field of that year. This yielded 100 different SLP reconstructions for 850-2005 CE, with their corresponding NAM series.

Similarly, in Chapter 5, for each month of 1750-2004 CE, we selected the N best analogues of the SLP distribution defined by the available observations of the SLP-Obs network (see Section 2.1.2). The reference pool is formed by the SLP fields of the 20CRv3 reanalysis (Slivinski et al., 2019) except the map that we intend to reconstruct. Most suitable analogues are defined as the N maps of the pool with the lowest RMSE between the SLP observations and the corresponding grid point values of the reanalysis. Note that in most cases, best analogues correspond to the actual month intended to reconstruct or the months before and after it (e.g., for December 1900, the best analogue is December 1979, whereas for December 1800, the best analogue is November 1845). The reconstructed SLP field is given by the mean of the best N analogues of each month. The standard deviation across the N best analogues provides a measure of the uncertainty, i.e. how much the large-scale field is constrained by the available set of observations. After testing different N values ranging from 1 to 50, we used the N=10 best analogues of each month. This number was chosen as a balance between the Pearson correlation coefficient, which increases with N, and the variability ratio, which decreases with N (Fig. 3.2).

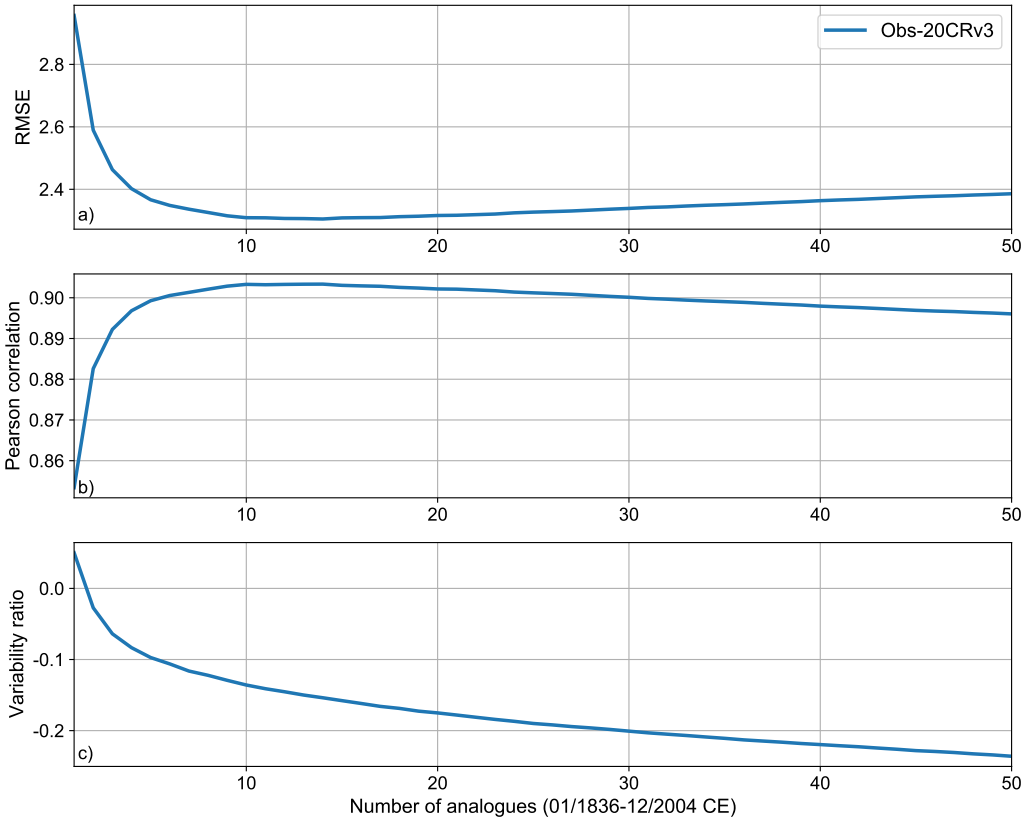

Figure 3.2: Sea level pressure reconstruction skill of the Analogue Method. The performance of the method is shown as a function of the number of selected analogues for (a) the RMSE, (b) the Pearson correlation, and (c) the variability ratio between the reconstructions using observations and the 20CRv3 reanalysis during the 1836-2004 time period.

3.1.2 Canonical Correlation Analysis

The CCA is a mathematical procedure that infers information from cross-covariance matrices (Hotelling, 1936). For instance, if we want to find relationships between different variables, the CCA will generate linear combinations of these variables that maximize the correlations among them. Therefore it can be considered a general test of significance in statistics (Knapp, 1978), and it is commonly used as a dimensionality reduction technique in atmospheric science (Wilks, 2011).

The CCA as described by Smerdon et al. (2010) is employed as CFR method in paleoclimate reconstructions (Neukom et al., 2019) based on maximizing the correlation (specifically the canonical correlation coefficients) of reduced eigenvector spaces from the proxy ($\mathbf{P}$) and calibration ($\mathbf{C}$) datasets (the latter defined as the annual temperature field of the target simulation since 1850 to 2005 CE). All temperature series are standardized by subtracting their mean and divided by their standard deviation. To avoid degenerated eigenvalues (eigenvalues with a value close to 0), both datasets have been truncated (Eqs. 3.1 and 3.2) with a minimum tolerance of 0.01 (eigenvalues below this threshold have been discarded) and a maximum number of 74 eigenvectors (optimal for full-proxy reconstructions).

$$\mathbf{P} = \mathbf{P}^r + \delta_p = \mathbf{U}_p^r \cdot \mathbf{\Lambda}_p^r \cdot \mathbf{V}_p^{r\mathrm{T}} + \delta_p \tag{3.1}$$

$$\mathbf{C} = \mathbf{C}^r + \delta_c = \mathbf{U}_c^r \cdot \mathbf{\Lambda}_c^r \cdot \mathbf{V}_c^{r\mathrm{T}} + \delta_c \tag{3.2}$$

where $\mathbf{U}^r$ and $\mathbf{V}^r$ are matrices composed of orthogonal eigenvectors that diagonalize the reduced proxy ($\mathbf{P}^r$) and calibration ($\mathbf{C}^r$) datasets (where δ is the corresponding residual after the truncation), obtaining their diagonal matrices ($\mathbf{\Lambda}^r$) from SVD (*Singular Value Decomposition*). The matrix of regression coefficients ($\mathbf{B}$) is then generated as in Eq. 3.3,

$$\mathbf{B} = \mathbf{U}_c^r \cdot \mathbf{\Lambda}_c^r \cdot \mathbf{V}_c^{r\mathrm{T}} \cdot \mathbf{V}_p^r \cdot (\mathbf{\Lambda}_p^r)^{-1} \cdot \mathbf{U}_p^{r\mathrm{T}} \tag{3.3}$$

and subsequently decomposed into canonical correlation coefficients (Eq. 3.4) using SVD over $\mathbf{V}_c^{r\mathrm{T}} \cdot \mathbf{V}_p^r$.

$$\mathbf{B} = \mathbf{U}_c^r \cdot \mathbf{\Lambda}_c^r \cdot \mathbf{O}_c \cdot \mathbf{\Lambda}_{\mathrm{cca}} \cdot \mathbf{O}_p^{\mathrm{T}} \cdot (\mathbf{\Lambda}_p^r)^{-1} \cdot \mathbf{U}_p^{r\mathrm{T}} \tag{3.4}$$

where $\mathbf{O}_c$ and $\mathbf{O}_p$ are matrices composed of orthogonal eigenvectors that transform $\mathbf{B}$ into the diagonal matrix $\mathbf{\Lambda}_{\mathrm{cca}}$, whose diagonal elements are the canonical correlation coefficients. Thus, the temperature field reconstruction,

$\mathbf{R}$, is obtained in Eq. 3.5

$$\mathbf{R} = \mathbf{U}_c^r \cdot \mathbf{\Lambda}_c^r \cdot \mathbf{O}_c \cdot \mathbf{\Lambda}_{\text{cca}} \cdot \mathbf{W}_p^{\text{T}} \cdot \mathbf{P} \tag{3.5}$$

where $\mathbf{W}_p$ is the CCA proxy weighting matrix defined as Eq. 3.6

$$\mathbf{W}_p = \mathbf{U}_p^r \cdot (\mathbf{\Lambda}_p^r)^{-1} \cdot \mathbf{O}_p. \tag{3.6}$$

3.2 The Coral Reef Optimization

Evolutionary algorithms are soft-computing techniques inspired by biological processes (Del Ser et al., 2019) such as genetic recombinations and mutations that ensure the survival and evolution of best suited individuals within a natural competitive environment (Eiben and Smith, 2015). These algorithms are designed to provide optimal solutions through the combination (Forrest, 1993) and competition of previously-generated solutions. For high dimensional problems the direct search of the best solution (assuming that it exists and it is unique) is computationally infeasible (Knüsel, 2019). These evolutionary algorithms generate near optimal solutions (Salcedo-Sanz et al., 2018), and tend to outperform other indirect methods when dealing with non-linear problems, such as those of climate.

From all evolutionary algorithms available, we decided to use a version of the CRO with four substrate layers (i.e. search operators). The CRO is a hybrid algorithm (Salcedo-Sanz, 2017) that emulates the living processes of corals and their evolution within an ocean reef. By limiting the number of corals within the reef, the best adapted species will have a higher probability of surviving, promoting the subsequent evolution of best individuals over next generations. Thus, the same way that best adapted coral generations survive over time by transferring their genetic information to their descendants, the CRO applies different recombination procedures (Forrest, 1993; Vrugt and Robinson, 2007) to transfer parts of sub-optimal solutions into new optimal ones generated at each iteration. Mathematically, this is imple-

mented through different techniques (Vrugt and Robinson, 2007) known as search operators (see Subsection 3.2.3) that recombine the set of solutions to iteratively generate better solutions from parts of previous ones. Moreover, to avoid falling in a local minimum, there is also a small probability of spontaneous random perturbations in the set of solutions (known as mutations) to expand the space of solutions.

In climatology, the CRO algorithm has already been used to find sets of locations that best describe spatially resolved climate fields such as wind speed (Salcedo-Sanz et al., 2018) or temperature (Salcedo-Sanz et al., 2019), and their results have been applied to improve solar and wind power forecasts (Salcedo-Sanz et al., 2018) at local scales.

3.2.1 Coral solutions

In the CRO algorithm, corals are defined by two elements: a solution for the predefined problem and a health function that characterizes how good or bad that solution is. For instance, in Chapter 4, we use the CRO algorithm to search for subsets of representative proxy locations of the PAGES-2k network that optimize the area-weighted RMSE between the target field and its reconstruction, obtained from different CFR methods (Section 3.1). Therefore, to solve this optimization problem, corals are composed of a binary array of proxy locations (i.e. the coral solution) with "1" values when such locations were selected and "0" when they were not (Fig. 3.3a). Afterwards, the health function for each coral was calculated as the RMSE between the reconstruction using the selected locations and the target field.

a) Selection of proxy locations (0 rejected, 1 selected):

Record 1	Record 2	Record 3	...	Record n-1	Record n
1	0	1	...	1	0

b) Weather station weighting (from 0 to 1):

Record 1	Record 2	Record 3	...	Record n-1	Record n
0.85	0.22	0.76	...	0.91	0.18

Figure 3.3: Coral solutions. (a) Binary selection of representative proxy locations for the reconstruction of annual 2-m air temperature fields from 850 to 2005 CE. (b) Weighting of weather stations for the reconstruction of SLP fields from monthly observations since 1750 CE.

On the other hand, a more realistic approach is attempted in Chapter 5. It pursues an optimized non-discrete set of weights (ranging from 0 to 1) applied to a real network of SLP observations (SLP-Obs) affected by gaps and observational errors (Fig. 3.3b). The CRO algorithm was run to find optimized weights for the network of SLP records, which are applied during the AM reconstruction. Herein, different corals represent different sets of weights, which measure the degree of large-scale representativeness of the local records, under the given restrictions of the observing network (errors, data availability, etc.). For ideally perfect conditions, weights would only depend on the relationships between local observations and its links with the *ground truth*. However, in the presence of uncertainties contaminating these relationships, weights are also affected by changes in the observing network and observational errors over the reconstructed period. For example, assuming an idealized configuration of equally skillful continuous records, the optimization would assign lower weights to records with larger errors due to inconsistencies with the remaining observations and the large-scale field. This makes optimal weighting necessary to minimize biases induced by uncertainties in the climate network.

3.2.2 The algorithm

Fig. 3.4 illustrates the CRO general steps inspired by the multiple mechanisms of coral reproduction. These mechanisms have been defined in the algorithm as different recombination procedures to generate new solutions. Here we present a brief description of each step (see Salcedo-Sanz (2017) for further details).

Random generation

Coral solutions (as described in Subsection 3.2.1) are randomly synthetized and latched to the reef. Note that the reef has a limited number of spots and corals have to latch on one of them to survive for the next iteration.

External sexual reproduction

Some corals are randomly selected by pairs (i.e. parents) to generate new "baby corals" known as larvae. In this step, new solutions are generated by combining parts of both solutions encoded in the parents. Subsection 3.2.3 describes four different search operators used to embed the genetic information of adult corals into new larvae.

Internal sexual reproduction

The remaining corals (those that have not been reproduced by couples) generate new larvae by copying its solution (genetic information) and applying random perturbations (mutations), so that new solutions are partially different to the original ones.

Asexual reproduction

Asexual reproduction has been included in the algorithm the same way as the internal sexual reproduction. However, only a given percentage of the best solutions will produce new larvae through this method. This is intended to increase the number of better solutions available for reproduction in the next iteration.

Larvae setting

After larvae are formed, they need to latch onto one of the reef's spots in order to become a coral. Two different situations can happen at this moment: the spot is empty or it is occupied by other coral. In the first case the larva is always clinged to the spot. In the second, the latching spot is disputed by competition, and the coral with better health function (Subsection 3.2.1) will gain the battle. For instance, if the health function is defined as the skill of a certain climate reconstruction, the coral with the best skill (i.e. lowest error) will latch onto the spot and survive, while the other will perish. This way, the survival and reproduction of best solutions are guaranteed after successive iterations.

Coral predation

After reproduction and coral settlement, if the stop condition is not met (i.e. a predefined number of iterations has not been reached), a percentage of the worst coral solutions in the reef are preyed, leaving empty spots for future coral generations.

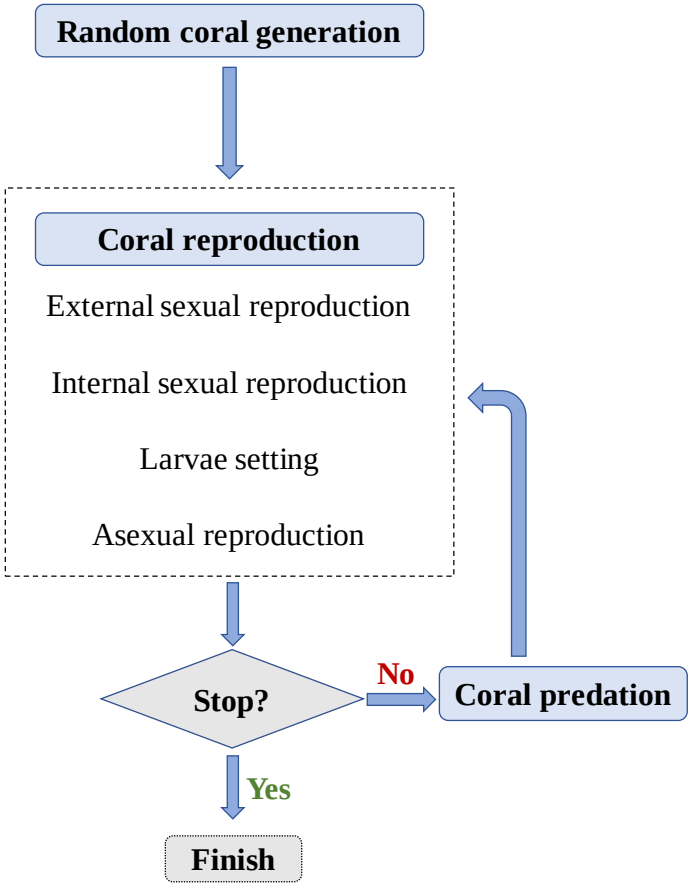

Figure 3.4: CRO flowchart imitating the biological processes of corals within a reef.

3.2.3 Search operators

Advanced versions of the CRO algorithm simulate substrate layers within the reef, implemented in the code as different search operators (Salcedo-Sanz, 2017; Salcedo-Sanz et al., 2019) during the step of external sexual reproduction. We have included four of these operators for the studies presented in Chapters 4 and 5.

One-point crossover

Coral parents generate two new larvae by interchanging their solutions along a randomly chosen point.

Two-point crossover

Coral parents generate two new larvae by interchanging their solutions along two randomly chosen points.

Multi-point crossover

Coral parents generate two new larvae by interchanging their solutions along multiple points.

Differential Evolution

Differential evolution is based on adding differences of two parents into a third coral (Salcedo-Sanz et al., 2019), following Eq. 3.7 for each encoded parameter (x_i) of the solution.

$$x_i^{\text{larva}} = x_i^{\text{coral}_1} + \sigma \cdot (x_i^{\text{coral}_2} - x_i^{\text{coral}_3}) \tag{3.7}$$

where x_i^{larva} is the i^{th} parameter of the new larva solution, $x_i^{\text{coral}_n}$ are the parameters (i) of three coral parents, and σ determines the fraction of the difference between coral_2 and coral_3 that gets transferred into coral_1.

3.3 Clustering techniques

Classical clustering techniques, such as the k-means algorithm (Hartigan and Wong, 1979; Phillips, 2002), have become widespread in the past few years as dimensionality reduction methods are able to extract relevant information from extensive databases (Bernard et al., 2013; Bador et al., 2015; Zhang et al., 2016). In climatology, these methodologies can arrange data according to their internal structure by defining spatial regions for datasets with geolocated climate information (Rao and Srinivas, 2006). Therefore, clustering algorithms have been used for several purposes such as the identification of regional climates (Aliaga et al., 2017), air pollution (Gao et al., 2011; Wang et al., 2015), and ecology (Miele et al., 2014; Cheruvelil et al., 2017).

These classical clustering techniques often require complete datasets, limiting regional analyses to series without time gaps. Unfortunately, most available climate archives (Haylock et al., 2008; Glaser and Riemann, 2009; Emile-Geay et al., 2017) contain missing values which must be properly handled prior to clustering. Most straightforward approaches consist of removing data points (deletion) that do not cover the requested time period (Dixon, 1979), whereas more sophisticated methods intend to estimate missing values (imputation) by means of statistical procedures (Henn et al., 2013). This limitation restricts cluster analyses to periods with complete information, disregarding earlier climate imprints contained in longer time series. In turn, there are only a few methods designed to work with inhomogeneities in climate datasets. Such is the case of k-POD (Chi et al., 2016), an algorithm that instead of relying on deletion and imputation, uses a majorization-minimization algorithm (Lange et al., 2000) to cluster observed data with missing values. It is however difficult to find a method that performs well with sparse climate datasets.

In this section a new clustering method known as the k-gaps algorithm has been defined to classify sets of time series with a significant number of missing values. Its structure is similar to those employed in classical

clustering methods such as the k-means algorithm, but with some key changes that allow for the selection and attachment of records with different temporal lengths. The validation of the method together with its multiple applications in the climate field are in Chapter 6.

3.3.1 Assumptions and Definitions

Let us assume we have a dataset of climate records at different locations, and maybe with different temporal lengths, which together describe the climatology of a specific zone (Europe in our case), for a given period $\mathbf{T} := [n_1, n_2, \cdots, n_T]$, where n_i stands for the time step at which a certain climate variable has been measured, and T represents the total number of time steps. Thus, a given climate record A is defined for a subset of $\mathbf{T}$, and may (or may not) overlap with other climate records included in the dataset, (i.e. incomplete records are considered).

Generally, clustering climate records implies to compute distances between some centroids (in our case, temperature series that are representative of certain regions) and the records, by considering a pre-defined metric (the root-mean-square deviation, for example). However, note that for incomplete datasets, these distances can only be estimated during the time intervals with available information. Therefore, to identify periods where some time series overlap, a vector mask ranging from the oldest to the most recent time step in the dataset has been defined for each time series. These record masks can be defined as indicator functions (Eq. 3.8), filled with "1"s when their associated time series have available data, and "0"s for the remaining period.

$$\text{Mask}_A(n_i) := \begin{cases} 1, & \text{if } n_i \in A \\ 0, & \text{if } n_i \notin A \end{cases} \quad \forall\ n_i \in \mathbf{T} \tag{3.8}$$

For instance, Fig. 3.5 illustrates two randomly-generated records (signals A and B with no specific units) with different temporal lengths, and the subsequent time series obtained when these two records are merged (Fig.

3.5c), for example in a centroid calculation procedure.

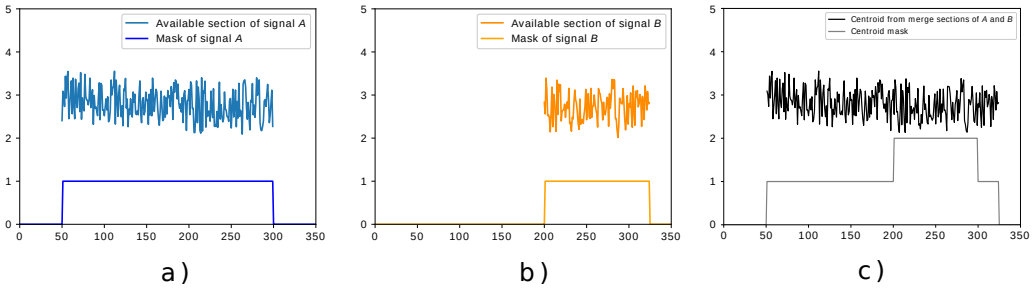

Figure 3.5: Data records considered (a and b), and the resulting series obtained by merging them (c). All the records are presented with their respective masks.

In our case, signals A and B were combined by averaging them in the time range where there are data available in both series (i.e. during the overlapping interval). Otherwise, when only one of the signals is available, its values are included to the merged series. Note that the mask in Fig. 3.5c contains the number of time series with available data at each time step and, although the resulting period is defined from 50 to 325 (i.e. the time interval in which there is at least one time series available), the overlapping interval used to combine them is in-between 200 and 300. Thus the resulting series has data of signal A from time 50 to time 199, the averaged information of signal A and signal B from time 200 to time 300, and the information of signal B from time 301 to time 325. Note that the use of vector masks allows to merge records with different temporal lengths (e.g. temperature observations, or historical archives), and provides information about the number of time series used to calculate a resulting centroid. Moreover, these masks are important because keeping the number of series available at each time step allows us to discern between periods with robust mean values (i.e. time steps with a high number of temperature series available) and those time intervals with data scarcity.

It is noteworthy to mention that clustering techniques can help to iden-

tify regional climates by grouping together instrumental time-series with similar mean temperatures or correlated variability. While the former can be achieved by using the series as they are (absolute values), clusterizing time-series focusing on their variability requires normalizing the series first. For instance, Fig. 3.6 depicts two records with different mean (Fig. 3.6 a and b), but correlated variability (Fig. 3.6c), that would be clustered together if they were normalized.

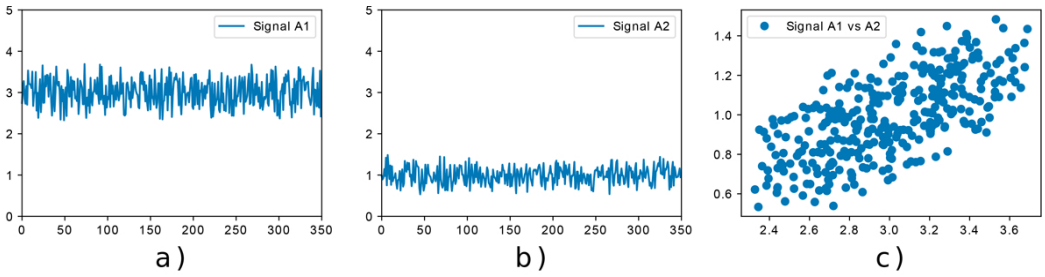

Figure 3.6: Representations of two synthetic records. Series with different mean values (a and b), but with correlated variabilty (c).

The k-gaps algorithm has therefore been designed to cluster time series with similar mean ("basic" mode hereafter) and also with correlated variability ("normalization" mode). While the basic mode is directly applied over temperature records, the normalization mode requires a workaround because homogeneous normalizations cannot be calculated from records with different temporal lengths. This issue has been tackled by applying an adjustment of the climate data to calculate the centroids (Subsection 3.3.3) and a linear fitting to properly reclassify the time series (Subsection 3.3.4).

3.3.2 The k-gaps algorithm

Taken into account these previous definitions, we can now describe the k-gaps algorithm for clustering records of incomplete time series. The general structure of k-gaps is similar to the well known k-means algorithm (Hartigan and Wong, 1979), which is an iterative method whose main purpose is to classify series of data within clusters represented by central vectors known

as *centroids* (C_j). In k-means, these centroids are calculated by averaging the series associated with previous clusters, and subsequently reassigning the records to the nearest centroid for the next iteration. The proposed k-gaps algorithm follows the same idea, but including some extra steps to treat the problematic case of having incomplete records in the database.

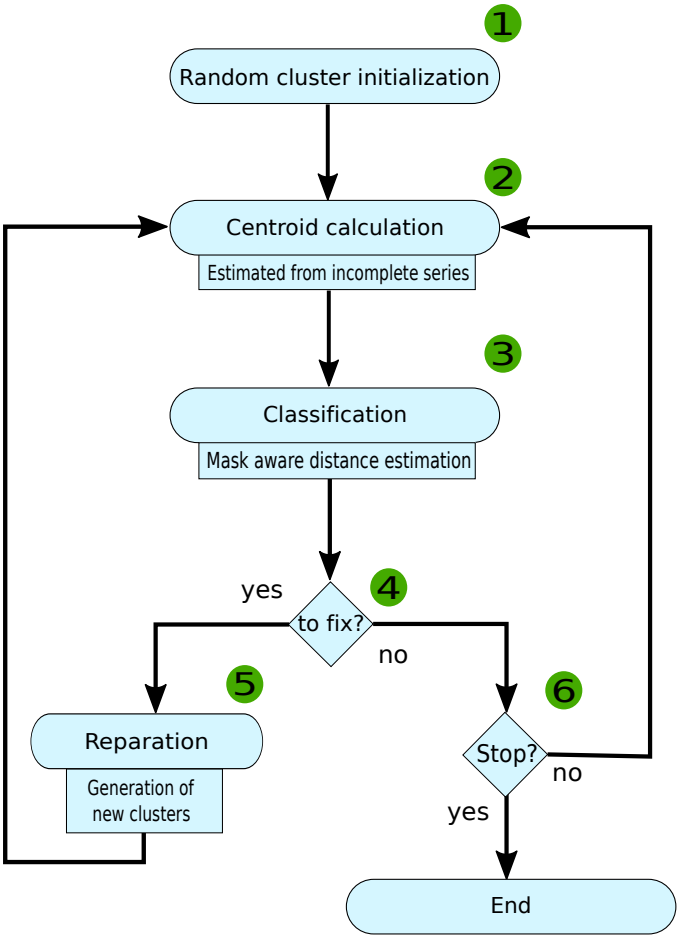

Figure 3.7: k-Gaps flowchart. Circles contain general clustering procedures, and squares describe specific k-gaps operations. The algorithm's conditions are represented as diamonds in the flowchart.

Fig. 3.7 shows a flowchart of the k-gaps implementation, which can be summarized in the following steps:

1. Randomly assign the existing records (time series) over k initial clusters.

2. Proceed to centroid calculation, following the procedure described in Subsection 3.3.3.

3. Reassign records to clusters by similarity with the centroids (Subsection 3.3.4).

4. Count the number of records for each cluster.
   For empty clusters, proceed to step 5. Otherwise, jump to step 6.

5. Generation of new clusters by applying random Gaussian noise to existent groups, and restart from step 2.

6. Check whether the stopping condition has been fulfilled (Subsection 3.3.5). If it is not, go to step 2.

3.3.3 Centroids calculation

In clustering algorithms requiring complete datasets such as the k-means, centroids are usually calculated by averaging the time series associated with a certain cluster at each time step. However, this procedure can debase the robustness of the clustering when applied over incomplete datasets (i.e. sets of records defined for different time periods) by producing unrealistic clusters when some time series do not overlap. This bias will propagate for the subsequent iterations, clustering artificially-mixed regions. To circumvent this problem, centroids have been estimated by only averaging records that overlap for at least a minimum time interval. This minimum overlapping must be estimated in each case depending on the temporal resolution and the number of time steps that characterize the dataset aimed to cluster. After trying different parametrizations, in Chapter 6 we found that the k-gaps

algorithm could not converge for values lower than 90 days, yielding significantly different clusterings over successive runs. This indicates that clusters are barely robust and their results must not be trusted. We tackled this issue by setting a minimum overlap of 90 days. The procedure to calculate the centroids in the algorithm will be described next. Note that it is different depending whether we consider k-gaps in basic mode or in normalization mode:

Procedure for k-gaps basic mode (i.e. clustering records with similar means):

1. Set the longest time series as the new centroid (C_j).

2. Select the records (S_i) with the longest overlap with the centroid ($Ov_{i,j}$) using Equation (3.9).

$$Ov_{i,j} = \sum_{n=0}^{n_T} F_{i,j}(n) \rightarrow F_{i,j}(n) := \begin{cases} 0 & \text{if} \quad \text{Mask}_i(n) \cdot \text{Mask}_j(n) = 0 \\ 1 & \text{if} \quad \text{Mask}_i(n) \cdot \text{Mask}_j(n) > 0 \end{cases} \tag{3.9}$$

   where $j \in [0, k)$ represents a certain cluster (of a total of k clusters), n is the time, n_T is the last time step, Mask_i is the mask associated with S_i, and Mask_j is the centroid mask.

3. Combine the centroid (C_j) and S_i following Equation (3.10):

$$C_j = C_j + S_i \cdot \text{Mask}_i \tag{3.10}$$

   Centroids are firstly calculated by adding at each time step the temperatures of their associated series.

4. Update the centroid mask using Equation (3.11):

$$\text{Mask}_j = \text{Mask}_j + \text{Mask}_i \tag{3.11}$$

5. Get back to Step 2 until all records are checked out or overlapping is below a predefined threshold.

6. Divide the centroids in Step 3 by their respective Mask_j to obtain their average temperature at each time step (note that centroid masks are defined as the number of series available at each time step).

For each cluster in normalization mode (i.e. clustering time series with similar variability):

1. Set the longest series as the new centroid (C_j).

2. Select the time series (S_i) with the longest overlap with the centroid ($Ov_{i,j}$) using Equation (3.9).

3. Adjust the centroid using a linear regression estimated in the overlapping section between the centroid and the series (Equation (3.12))

$$S_i = c_1 \cdot C_j + c_0 \rightarrow S_i' = \frac{S_i - c_0}{c_1} \tag{3.12}$$

   where c_0 and c_1 are two constants known as the intercept and slope of the regression line respectively. Note that while the centroid C_j (which is a time series that determines a certain cluster) is generated by combining normalized series (S_i'), it is adjusted by linear regression to the original series S_i, preventing c_1 from being zero. (The intercept (c_0) is subtracted from the original series (Si) and the result is divided by the slope (c_1), obtaining S_i').

4. Combine the centroid and S_i' following Equation (3.10).

5. Update the centroid mask using Equation (3.11).

6. Divide the centroid by its mask. This process is undertaken for each time series considered, since it is necessary to estimate the centroid to apply Step 3 for the forthcoming record.

7. Back to Step 2 until all records are checked out or overlapping is below a predefined threshold.

3.3.4 Assignment of records to clusters

Clustering algorithms associate records with different clusters by means of a certain metric. In its basic mode, k-gaps assigns each record to the cluster whose MSE (*Mean-Square Error*) is minimum, following Equation (3.13).

$$\mathrm{MSE}(i,j) = Ov_{i,j}^{-1} \sum_{n=0}^{Ov_{i,j}} F_{i,j}(n) \left[S_i(n) - C_j(n)\right]^2 + P \tag{3.13}$$

where $Ov_{i,j}$ is the overlapping length and P is a penalty associated with short overlapping intervals between two time series. Note that the number of superimposed values should be long enough to ensure that the metric used for the classification of time series into clusters (MSE) is significative, because otherwise time series with good fit during a short overlapping interval could be assigned to the wrong cluster (that is why it is important to add the value P in Equation 3.13). Clusterings of this sort are considered barely robust because they show very different classification patterns each time the algorithm is run. Hence, the procedure to find the right value for P is by testing different values until robust clusterings are obtained. We used series of daily temperatures and found that by setting a minimum overlapping period of 30 days and a penalty of 100, the final pattern of the clusterization did not significantly change over successive runs of the k-gaps algorithm.

In the normalization mode, the assignment process is different: in this case linear regressions are computed between time series (S_i) and clusters (C_j) to find the centroid that best fits each record. In our case, cluster assignment is performed by minimization of mean quadratic errors (ϵ_j^2) obtained from residuals of the linear fit estimated using Equation (3.14).

$$S_i = c_1 \cdot C_j + c_0 + \epsilon_j \rightarrow \mathrm{Error}_{i,j} = Ov_{i,j}^{-1} \sum_{n=0}^{Ov_{i,j}} \epsilon_j^2(n) + P \tag{3.14}$$

where c_0 and c_1 are once again the intercept and slope of the linear fit, respectively. Note that the linear regression is performed in the overlapping

section between each centroid and the time series.

3.3.5 Stop conditions

In the classic k-means algorithm, the stop condition is usually reached when convergence is detected (no change of the solution in a number of iterations) or, after a given number of iterations that ensures the algorithm has obtained a "good enough" solution. However, additional criteria are required in the k-gaps to handle incomplete records. For instance, since centroids are based on compositions of uneven time series, they might be affected by adding and removing records from their corresponding clusters, and therefore, total convergence cannot be ensured. Furthermore, looping assignation may occur when some series move cyclically from one cluster to another during consecutive iterations, affecting the convergence of the algorithm. Therefore, to provide a general stop condition for k-gaps, a detector of clustering changes (D) has been defined (Equation (3.15)) as the summation of absolute values obtained from differences between the number of time series in each cluster at a given step (s), and the number of time series in those clusters for the next iteration ($s+1$).

$$D = \sum_{j=0}^{k} |\text{Size}_j(s) - \text{Size}_j(s+1)| \tag{3.15}$$

where k is the total number of cluster, and $\text{Size}_j(s)$ represents the number of records in cluster j during step s.

Hence, when the value of D tends to zero, we assume that the algorithm has converged, and the stop condition has been reached. On the other hand, when D values are repeated over successive iterations, it indicates that the algorithm entered in one of the aforementioned loops, and a halting procedure should be applied, starting again the algorithm with a different initial condition.

Chapter 4

Selection of proxy locations for temperature reconstructions

4.1 Background

Decades of scientific fieldwork have left abundant proxy datasets to reconstruct the climate of the past. They provide information about climate changes at local and regional scales, and have been pooled to derive CFR of the last millennium (Crowley, 2000; 2k Consortium, 2013). They are spatially-resolved reconstructions of climate variables generated from different paleoclimate archives. These reconstructions are affected by several sources of uncertainty (Christiansen and Ljungqvist, 2017) related to the reconstruction method (e.g. dimensionality, dependence on parameters, frequency coherence) (Evans et al., 2014), the underlying proxy observations (Neukom et al., 2018) (e.g. observational error, irregular chronologies, observed resolution), their links with the target field (e.g. multivariate signals, stationarity, spatial and temporal covariance, resolved resolution and seasonality), or the non-uniform spatial distribution of the observing network (Comboul et al., 2015). While most of these uncertainties have been thoroughly studied, spatial biases induced by sparse sampling locations are still not well understood. Due to the limited availability of paleoclimate archives, most proxy records are restricted to land, mainly the middle latitudes of

the Northern Hemisphere, while there are extensive un-sampled regions in the Southern Hemisphere and high latitudes. This unbalanced distribution of paleoclimate records induces a spatial bias in global CFRs that remains poorly quantified.

Circumventing this issue requires, in first place, a wise selection of proxies (Bradley, 1996). Due to the increasing availability of high-resolution records, the initiative PAGES-2k has scrutinized thousands of temperature-sensitive proxies, releasing a global archive with paleoclimate records for the last two millennia (Emile-Geay et al., 2017). This network has been recently exploited to derive global multi-proxy CFRs of annual temperature from a suite of reconstruction methods, including model-based assimilation schemes similar to those employed in modern reanalyses (Hakim et al., 2016; Franke et al., 2017). However, as these proxies are unevenly distributed, the spatial bias is still present. Selecting records strategically situated over key regions that capture the diversity of global patterns and simultaneously minimize the spatial bias of the observing network represents a significant challenge (Evans et al., 1998, 2001). Problems of this kind cannot be directly solved by testing all possible combinations due to their high dimensionality. Efforts to address this issue have only been attempted with sequential approaches (i.e. step-wise solutions based on local search procedures that perform incrementally by adding the proxy records with the best performance). Differently, a branch of artificial intelligence based on soft-computing techniques has recently emerged to solve high dimensional problems (Reichstein, 2019; Knüsel, 2019). For instance, biologically-inspired methods such as evolutionary algorithms (Eiben and Smith, 2015) are global search procedures that outperform sequential approaches in the task of finding optimal solutions to the representative selection problem within large and complex datasets (Salcedo-Sanz et al., 2019; Del Ser et al., 2019).

In this study we deal with the spatial bias in global CFRs of annual temperature arising solely from the non-homogeneous distribution of the currently available network of proxy records for the last millennium. To

better constrain this bias, other sources of uncertainty are avoided by using the CESM-LME (Otto-Bliesner, 2016) as a surrogated reality, where pseudo-proxies (synthetic temperature series from the target simulation) matching the locations of the PAGES-2k archive are artificially generated (see Methods). For these idealized conditions of the PAGES-2k proxy network (complete observational availability over time, univariate signals without observational error, stationary relationships, etc.) we generate global CFRs of annual temperature for the last millennium simulation of the CESM-LME that are biased by the uneven distribution of real proxies. CFR techniques (Gómez-Navarro et al., 2017) are then coupled with an evolutionary algorithm to explore if optimized subsets of PAGES-2k locations can be used instead without sacrificing the reconstruction skill. These pseudo-proxy experiments allow us to address the following questions in the perfectly known model′s world: Can we quantify the spatial bias due to the uneven distribution of records? Do we need all available records of the PAGES-2k network to maximize the skill of global temperature field reconstructions of the last millennium? If not, how many records are required to reconstruct the temperature of the last millennium without degrading the skill? Can we find a subset of PAGES-2k proxy locations that reduces the spatial bias of the full-proxy network?

4.2 Selection of representative locations

Annual temperature global patterns of the first full-forcing CESM-LME member simulation spanning the 850-2005 period of the Common Era (CE) are chosen as the target fields to reconstruct from pseudo-proxies (Smerdon, 2011) at the 569 grid-points matching the locations of the PAGES-2k archive. The accuracy of the CFR is quantified as its RMSE against the spatially-resolved global temperature patterns of the target simulation. Most experiments have been set under idealized conditions by using perfect pseudo-proxies directly assembled from the simulated temperature series of the target simulation. Unlike real reconstructions, our CFRs are only affected by

biases due to the spatial distribution of paleoclimate archives and the CFR methodology. This approach avoids other sources of uncertainty, allowing us to better constrain spatial biases, which can be inferred from changes in the reconstruction skill arising from different rearrangements of the pseudo-proxy network.

To test the sensitivity of the results to the CFR method, two different CFR techniques have been coupled to the CRO algorithm: The AM and the CCA as described in Section 3.1. Note that this coupling brings physical consistence to our method, proving that AI techniques can serve as a tool for climate science without losing the underlying physics of the Earth system. We used the 1850-2005 period of the first member of CESM-LME for the calibration of the CCA (similar results are obtained for shorter calibration periods as shown in Fig. A.2), while the remaining members of the same model ensemble are employed as a pool of analogues to derive the CFR with AM. Our results for the first member have also been tested in other realizations of the same model, a different model, and more realistic datasets, as explained below. To add complexity and more realistic conditions, experiments with additional sources of uncertainty were also performed. They allow us to test the robustness of the results and benchmark the magnitude of the spatial bias against that arising from other sources of uncertainty (the effect of multiple uncertainty sources should not be considered additive, though). In particular, we account for the amplitude of the observational error SNR by adding different levels of red noise to the perfect pseudo-proxies with a lag-1 autoregressive model (see Subection 2.2.1).

The AM-reconstructed global temperature fields using all 569 perfect pseudo-proxies (i.e. SNR = ∞) leads to a measurable RMSE of 0.65 °C, which should be ascribed to uncertainties in both the limited number and distribution of records (spatial bias) and the CFR method.

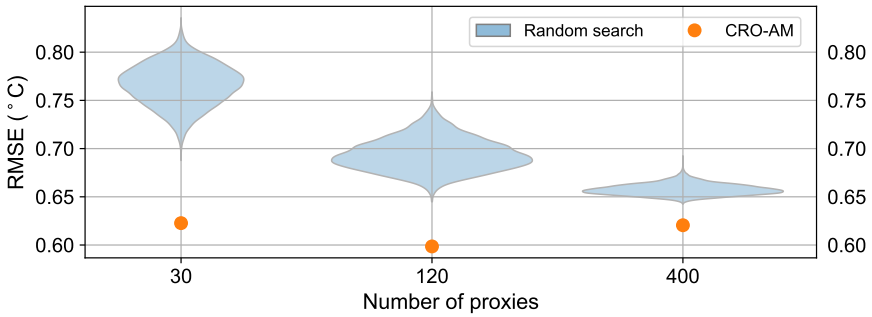

Figure 4.1: RMSE of 850-2005 CE global temperature fields reconstructed with different subsets of perfect pseudo-proxies from the PAGES-2k network. Orange dots represent the RMSE associated with the reconstructions using the optimized subsets of 30, 120, and 400 perfect pseudo-proxies of the PAGES-2k network obtained with the CRO-AM. Blue violins show the RMSE distribution obtained from 10000 reconstructions using different combinations of 30, 120, and 400 randomly selected pseudo-proxies from the PAGES-2k network. RMSE are calculated with respect to the global temperature fields of the target simulation (the first member of the CESM-LME).

To quantify the spatial bias, we derived new CFRs from reduced subsets of N perfect pseudo-proxies constrained by CRO-AM (*Coral Reef Optimization coupled with the Analogue Method*). For all optimized solutions of N records tested, the CRO-AM generates more skillful reconstructions than selecting subsets of N perfect pseudo-proxies at random (Fig. 4.1). Consequently, the improvement obtained with CRO-AM is not related to the reduction of random error, but a meaningful identification of representative locations for the reconstruction of temperature fields. According to Fig. 4.2 (blue curve), a minimum optimized set of 17 perfect pseudo-proxies, CRO-MIN (*Minimum subset of perfect pseudo-proxies obtained with CRO-AM necessary to outperform the reconstruction skill of the full-proxy network*) is enough to obtain global temperature fields with the same RMSE as the full-proxy reconstruction (green dashed line).

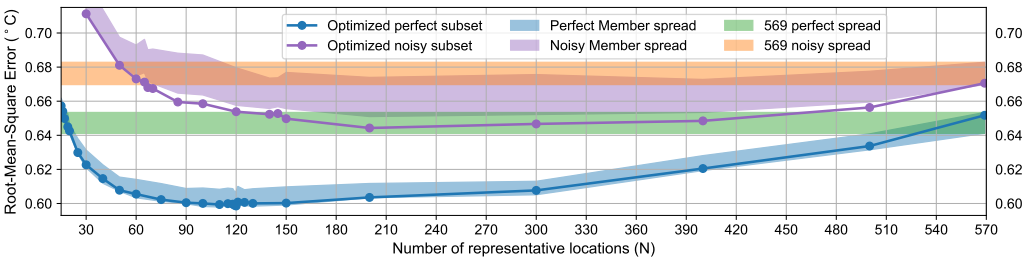

Figure 4.2: RMSE of the 850-2005 CE global temperature fields reconstructed with CRO-AM as a function of the number of selected perfect (blue) and noisy pseudo-proxies (purple) of the PAGES-2k network. The green and orange shades represent the 13-member reconstruction skill spread obtained with all (569) perfect and noisy pseudo-proxies (with SNR of 1) of the PAGES 2-k network. The blue-shaded area represents the spread obtained by using the optimized subset of N pseudo-proxies obtained for the first CESM-LME member to reconstruct the remaining members of the ensemble. The purple-shaded area is the same as the blue-shaded one but for reconstructions using noisy pseudo-proxies with SNR of 1. All shades depict 2 standard deviations with respect to the mean.

Therefore, under ideally perfect conditions, skillful reconstructions can be achieved even when the number of records is reduced up to one order of magnitude with respect to the PAGES-2k network. But, can we reduce further the spatial bias of the full-proxy reconstruction with a subset of well-distributed records larger than CRO-MIN? Our idealized experiments indicate that the RMSE of full-proxy temperature field reconstructions can be reduced up to 0.05 °C (Fig. 4.2) after selecting an optimal set of 120 perfect pseudo-proxies (CRO-OPT). Note that the associated error (0.60 °C) approaches that obtained from a full global grid coverage of perfect pseudo-proxies (0.54 °C). Besides minimizing the RMSE of the CFR for the first member, CRO-OPT (*Optimized subset of perfect pseudo-proxies of the PAGES-2k network obtained with CRO-AM that yields the best reconstruction skill*) locations also yield comparable levels of performance for the other members of the ensemble (Fig. 4.2, shading). This indicates that the selected network is representative across the ensemble, although the compara-

tively higher RMSE in those members points to additional improvements if the CRO-AM were applied separately to each realization. Similar estimates of the spatial bias are also found for CRO-CCA (*Coral Reef Optimization coupled with the Canonical Correlation Analysis*) reconstructions (a RMSE reduction of 0.05 °C for an optimized subset of 120 perfect pseudo-proxies), stressing the robustness of the results with respect to the CFR technique. Indeed, the RMSE variation with the size of the selected network shown in Fig. 4.2 is also reproduced by CCA reconstructions based on the locations selected by the CRO-AM (Fig. 4.3).

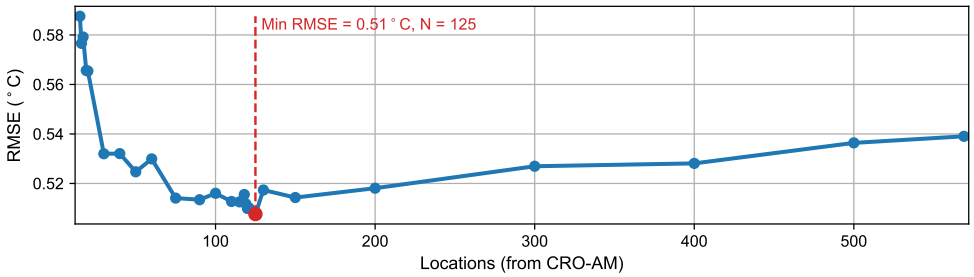

Figure 4.3: RMSE of CCA reconstructions generated with the optimized subsets of perfect pseudo-proxies of the PAGES-2k network selected by the CRO-AM.The red dot and dashed line highlight the minimum RMSE.

Table 4.1: RMSE of global temperature fields for 850-2005 CE (in °C) using CRO-AM reconstructions with N representative AR(1) pseudo-proxies of the PAGES-2k network and different SNR.

N	**SNR** $= \infty$	**SNR** $= 1$	**SNR** $= 0.5$
30	0.62	0.71	0.82
200	0.60	0.64	0.71
569	0.65	0.67	0.72

The reduction of the RMSE with CRO-OPT is not negligible, since it is equivalent to doubling the SNR in the full-proxy reconstruction (Table 4.1). Focusing on the noise, the RMSE derived from noisy pseudo-proxies with a SNR of 1 is only increased by 0.02 °C, whereas a SNR of 0.5 degrades the reconstruction skill by 0.07 °C. In the case of SNR-1 pseudo-proxies, the minimum number of records necessary to reach the same skill as their full-proxy reconstruction increases with respect to that obtained from perfect pseudo-proxies (60 instead of 17 in CRO-MIN). Still, it is possible to find representative subsets of 150 noisy pseudo-proxies that outperform the skill of the complete network of noisy pseudo-proxies (Fig. 4.2, lines). The same conclusion is obtained for pseudo-proxies with more realistic components of random noise (SNR=0.5), although previous experiments have shown that this level of SNR provides pseudo-proxy CFRs with lower skill than real-world reconstructions (Neukom et al., 2018). In addition, the noisy subsets that optimize the CFR of the first member are able to reduce the reconstruction error obtained for the other members of the ensemble from their complete networks of noisy pseudo-proxies (Fig. 4.2, purple shading). For certain conditions (SNR=1 in Fig. 4.2), optimized subsets of noisy pseudo-proxies can even outperform the skill of the full network of perfect pseudo-proxies.

Therefore, with all other sources of uncertainty being absent, spatial biases induced by a non-uniform distribution of high-quality (ideally perfect) proxies can potentially be larger than those obtained from a reduced subset of well-distributed noisy proxies with SNR of 1. Most of the selected locations in CRO-OPT are situated at high latitudes (Fig. 4.4a), stressing the importance of Arctic (Kaufman, 2009; Routson, 2019) and Antarctic (Steig and Neff, 2018) regions. Although there is not a unique solution, the distribution of optimal locations does not vary substantially with the number of selected records (cf. Figs. 4.4a and 4.5), and the CFR method (Fig. 4.5). In this regard, Figure 4.6a also shows that this distribution does not change regardless of the noise associated with the pseudo-proxies even when higher levels of SNR (e.g 0.5) are employed.

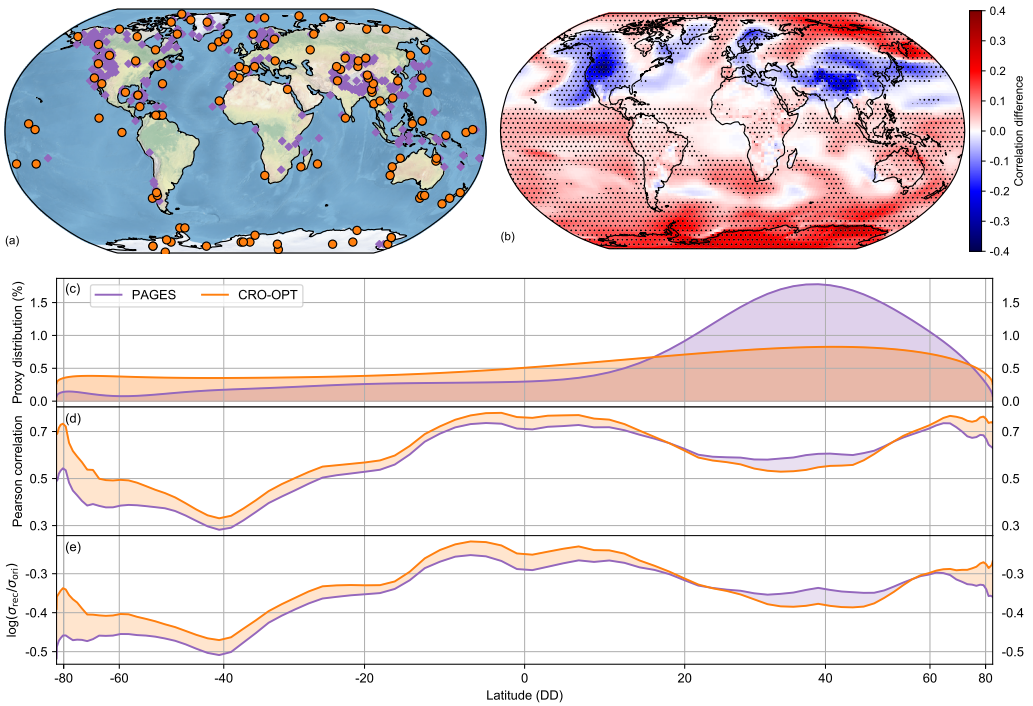

Figure 4.4: Performance of the CRO-AM reconstruction with the optimal subset of PAGES-2k records (CRO-OPT). (a) Spatial distribution of CRO-OPT records (orange dots) obtained from the full PAGES-2k network (purple diamonds). (b) Spatial correlation difference between the temperature reconstructions with CRO-OPT and all perfect pseudo-proxies. Stippling points illustrate significant correlation differences ($p < 0.05$). Kernel density estimation of the (c), Normalized latitudinal distribution of records (in % with respect to the total number of pseudo-proxies) for the CRO-OPT subset (orange) and the full-proxy PAGES-2k network (purple). (d) Latitudinal mean Pearson correlations for the CRO-OPT (orange) and full-proxy (purple) reconstructions. (e) Latitudinal logarithm of the standard deviation ratio for the CRO-OPT (orange) and full-proxy (purple) reconstructions (σ_{rec}) compared with the target simulation (σ_{ori}). The latitudinal axis is proportional to the effective area.

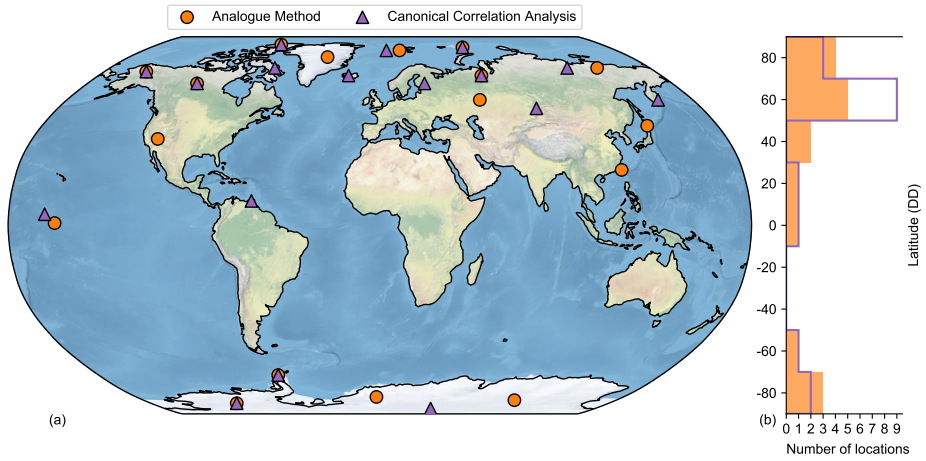

Figure 4.5: Optimized subsets of 17 perfect pseudo-proxies of the PAGES-2k network selected by CRO-AM (CRO-MIN) and CRO-CCA. (a) 2-D and (b) latitudinal distributions of the CRO-MIN locations obtained with CRO-AM (orange dots and shading) and the corresponding subset of perfect pseudo-proxies of the PAGES-2k network (with the same size as CRO-MIN) obtained with CRO-CCA (purple diamonds and line).

Moreover, Figure 4.6b illustrates how this distribution is always better than choosing locations at random, indicating that more robust reconstructions of the past climate are generated when representative locations are selected. The improvement of the CRO-OPT reconstruction is also reflected at local scales (Fig. 4.4b), implying that small changes in the global skill (Table 4.1) hide large regional improvements. The Pearson correlation coefficient of the target simulation with this reconstruction is significantly higher than with the full-proxy reconstruction for almost the entire Southern Hemisphere and the Arctic, and only performs worse in regions where the complete network presents high density of perfect pseudo-proxies. This is consistent with the spatial pattern of the RMSE difference between the CRO-OPT and full-proxy reconstructions (Fig. 4.7). Interestingly, highly sampled regions by the PAGES-2k network tend to coincide with areas of low spatial autocorrelation (Fig. 4.8). As such, the regional details of the temperature field over these regions are better captured by the denser full-proxy network than by CRO-

OPT. However, our experiments indicate that the regional improvement of the full-proxy network is attained with sacrifices in its global performance. This is explained by the fact that improving the reconstruction of small areas of the map comes at the expense of debasing the skill (i.e. lower correlation and higher RMSE) of the entire study region.

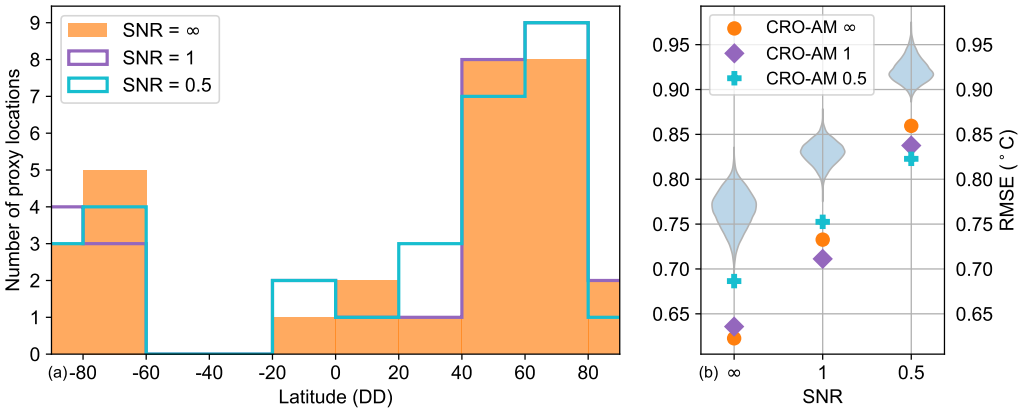

Figure 4.6: Sensitivity of CRO-AM reconstructions to pseudo-proxies with different levels of observational error. (a) Latitudinal distribution of the optimized subsets of 30 locations selected by CRO-AM from a PAGES-2k network of perfect pseudo-proxies (SNR $= \infty$, orange shading), and noisy pseudo-proxies with SNR $= 1$ (purple line) and SNR $= 0.5$ (blue line). (b) RMSE of CRO-AM reconstructions from pseudo-proxies with different SNR. For each type of pseudoproxies, symbols indicate the RMSE of the reconstruction obtained with the optimized subsets of locations found for perfect pseudo-proxies (orange dots), and noisy pseudo-proxies with SNR of 1 (purple diamonds) and 0.5 (blue crosses). Blue violins illustrate the RMSE distributions of 10000 reconstructions obtained from subsets of 30 PAGES-2k locations selected at random.

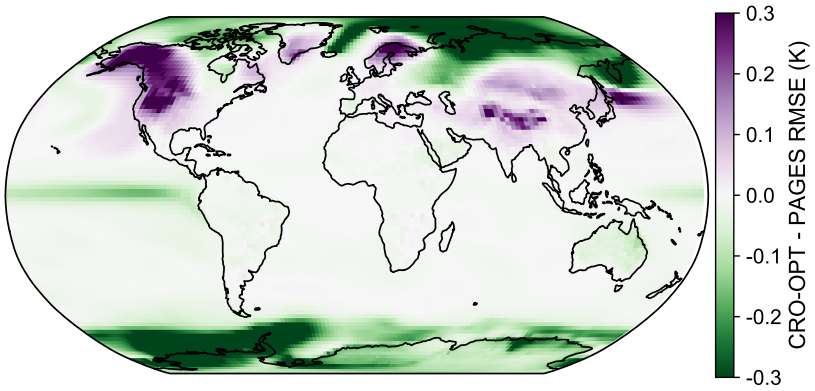

Figure 4.7: RMSE difference between the CRO-OPT and full-proxy reconstructions. Green (purple) color illustrates regions where CRO-OPT yields lower (higher) RMSE than the reconstruction with the full PAGES-2k network of perfect pseudo-proxies. RMSE are calculated with respect to the target field (the first CESM-LME member).

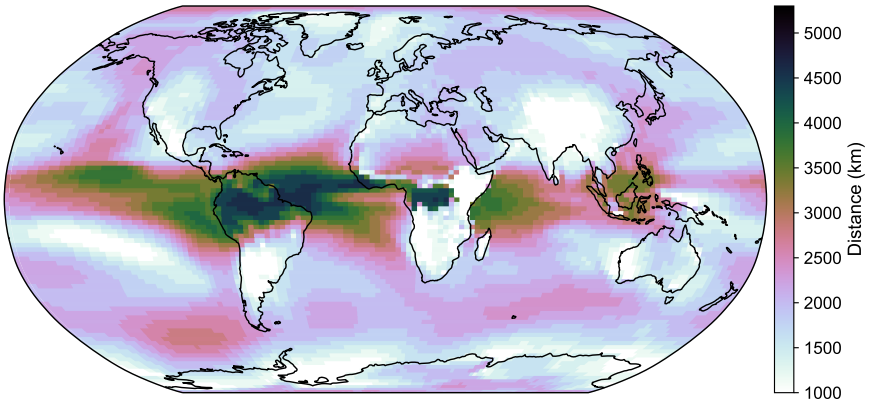

Figure 4.8: Spatial map of e-folding distances of decorrelation for the annual temperature of the first full-forcing CESM-LME member. The distance (in kilometers) for each grid point defines the area of the circle for which the averaged coefficient of determination (R^2) has decayed below e^{-1}.

These results show that spatial clusters of proxy records may debase the skill of spatially-resolved global temperature reconstructions. Note that the CRO-OPT distribution (Fig. 4.4c) is not simply a uniform one, and fewer points are often selected over areas with high spatial autocorrelation (e.g. the tropics). Still, its reconstruction leads to generalized latitudinal improvements in terms of correlation (Fig. 4.4d) and variability (Fig. 4.4e). These strategic locations also retrieve skillful reconstructions of area-weighted GMT (*Global Mean Temperature*) for the last millennium. The GMT of the reconstruction generated with CRO-OPT shows high skill (RMSE of 0.09 and R^2 of 0.88), improving that from the full-proxy reconstruction (RMSE of 0.16 and R^2 of 0.71). Recall that the selected locations represent an optimal subset of the PAGES-2k network for the specific conditions and target field of our idealized experiments. To test the independence of these results from the model employed, the CRO-AM experiment has also been run in the CCC400 model ensemble (Bhend et al., 2012; Franke et al., 2017). The results for an optimized subset of 120 locations (the same number as in CRO-OPT) are depicted in Fig. 4.9.

It is noteworthy to mention that False detection tests (Hu et al., 2017) regarding serial correlation and test multiplicity have been applied for the determination of global statistical significance in correlation maps between the reconstructed and target fields (Fig. 4.4 and Fig. 4.9). Because of the large N, effective degrees of freedom are high, leading to significant correlations at the 95% confidence level for all grid points with significant differences in Figs. 4.4 and 4.9. Furthermore, FDR (*False Discovery Rate*) have been calculated to assess whether there are falsely attributed significant correlations (Hu et al., 2017). In all cases, correlation maps passed the multiplicity test for a FDR of 5%. Note that these tests have been performed for pseudo-proxy experiments under idealized conditions, and the high statistical significance of the results is not necessarily representative of real proxy-based reconstructions, where additional sources of uncertainty are present.

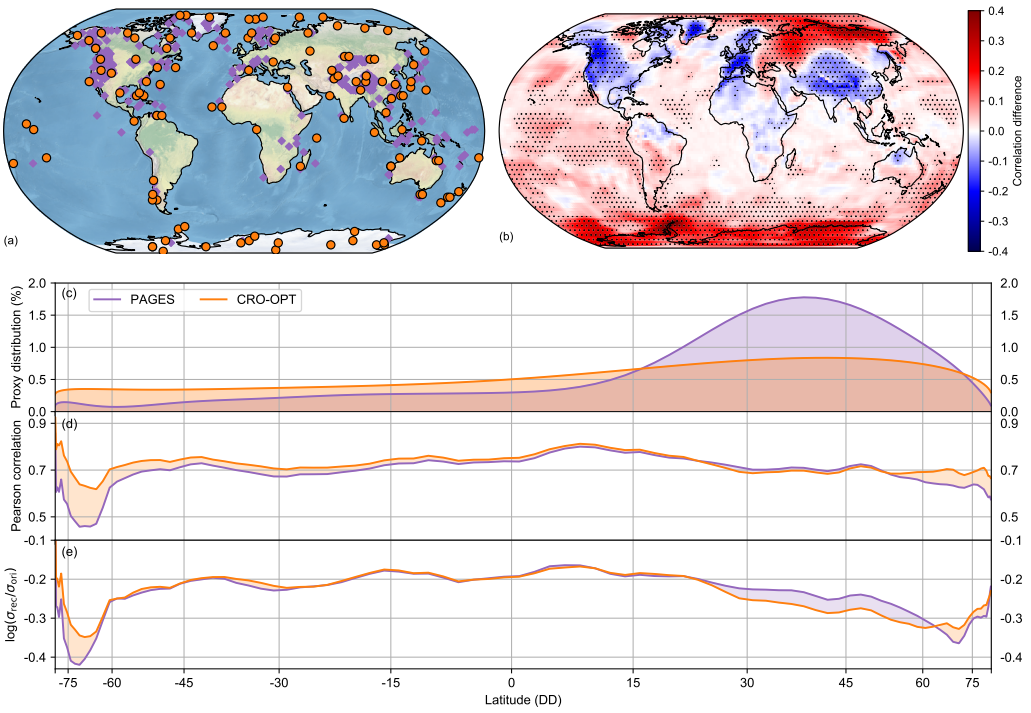

Figure 4.9: As Fig. 4.4 but using the global temperature fields of the CCC400 first ensemble member (1601-2005 CE) as target. The reconstruction has been obtained from the optimized subset of perfect pseudo-proxies of the PAGES-2k network (with the same size as CRO-OPT) selected by CRO-AM in the CCC400 model ensemble.

The spatial distribution of selected locations, the improvement of the correlation at high latitudes and part of the Pacific Ocean, and the latitudinal correlation and variability patterns resemble their CESM-LME counterparts presented in Fig. 4.4. A similar distribution persists in out-of-sample validation tests where the CCC400 ensemble is used as a pool of analogues to reconstruct the first member of the CESM-LME (Fig. 4.10), indicating that the selection of representative locations is not sensitive to the training data used in the process.

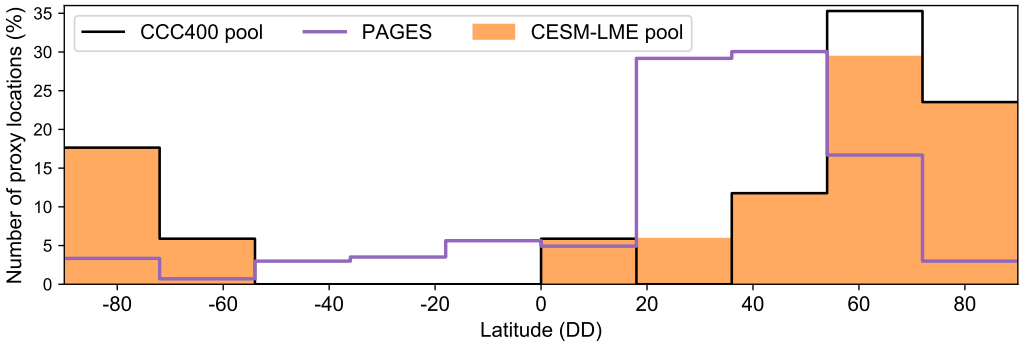

Figure 4.10: Latitudinal distribution of optimized subsets of the PAGES-2k network using different model ensembles as a pool for the CRO-AM reconstruction of the first CESM-LME member. Each subset includes 17 perfect pseudo-proxies obtained with the CRO-AM using as a pool members of the CESM-LME (orange shading) and the CCC400 ensemble (black line). In both cases, the target is the 850-2005 CE global temperature fields of the first member of the CESM-LME. The purple line illustrates the distribution of full-proxy PAGES-2k network.

We have also performed more stringent tests to assess whether the CRO-OPT locations inferred from the CESM-LME are also informative of GMTs in more realistic datasets. The exercise has been applied to instrumental temperature data based on HadCRUT 4.2 (Jones et al., 1999) for the post-industrial period (1850-2008 CE), as well as to proxy-based temperature reconstructions for the last millennium (850-2000 CE) provided by the Last Millennium Reanalysis (LMR, Methods). These products are independent from the CCC400 and CESM-LME, except for the fact that the LMR uses prior state estimates from an earlier version of the CESM atmospheric model. For each of these datasets, we computed first guess GMTs (GMTg hereafter) as the area-weighted mean temperature for two different sets of locations: the CRO-OPT, previously obtained with the CESM-LME, and the complete PAGES-2k network. Note that GMTg are directly obtained from the temperature series at these specific locations, without reconstructing the global temperature fields of the respective dataset. Although this approach is not ideal, and differs from global aggregation strategies typically employed for the

computation of GMTs (area-weighted global means from re-gridded fields), it gains importance in the light of incomplete coverage and time-varying availability of global records.

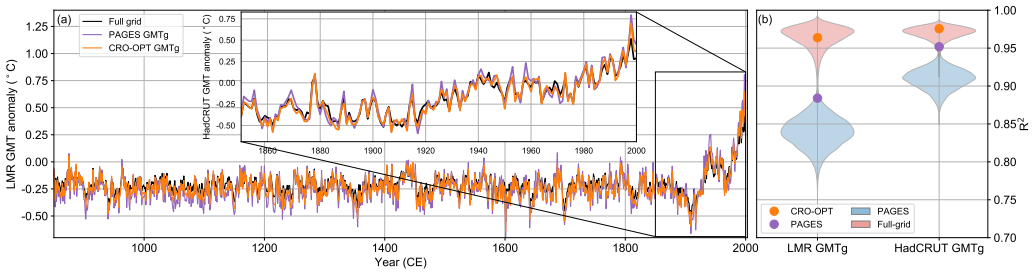

Figure 4.11: Estimates of GMT anomalies (°C) for the last millennium as inferred from selected subsets of the PAGES-2k network. (a) GMT anomalies from the LMR for 850-2000 CE (Inset (a): GMT anomalies from HadCRUT4.2 for 1850-2000 CE). Purple and orange lines show the GMTg of these datasets, defined as the area-weighted temperature mean for the grid points matching the PAGES-2k and CRO-OPT locations, respectively. All anomalies are computed with respect to the 1961-1990 baseline. (b) Coefficient of determination between the time series of GMT and GMTg from PAGES-2k (purple) and CRO-OPT (orange) locations. Violins illustrate the distributions obtained for 10000 subsets (with the same size as CRO-OPT) of randomly-selected locations from the PAGES-2k network (blue) and the full global grid (red).

Figure 4.11 shows how the HadCRUT 4.2 and LMR GMTs are overall consistent with their GMTg. In both cases, the coefficients of determination (R^2) are higher for CRO-OPT than for the complete PAGES-2k network, and they are also significantly higher than selecting random sets of locations from the PAGES-2k archive (Fig. 4.11b). This evidences the representativeness of CRO-OPT in observations and real-world reconstructions.

4.3 Reconstruction of temperature patterns

Despite its improvement, the optimized selection of locations is constrained by the original distribution of proxy records in the PAGES-2k archive, which displays limited coverage and a spatial bias towards mid-latitude land areas of the Northern Hemisphere, with very few proxies over oceanic regions. Therefore, we explored the skill of this reduced subset to capture internal variability and externally forced signals in the global temperature patterns. This also allows us to assess whether the selected locations have physical meaning or simply represent an optimized statistical distribution. El Niño-Southern Oscillation (ENSO) is chosen as an example of the former. The perfect pseudo-proxy CFR with CRO-OPT locations explains more than 80% of the SST variance over El Niño 3.4 region for the pre-industrial period of the target simulation, and even the more constrained CRO-MIN reconstruction captures global ENSO teleconnections reasonably well (Fig. 4.12).

Taking into account that ENSO is the main mode of internal variability on interannual time-scales, it is surprising that only one of the selected locations in CRO-MIN is over the tropical Pacific (Fig. 4.12b). However, the spatial autocorrelation of temperature variations in this region is among the highest of the globe (Fig. 4.8), which arguably reduces the effective number of records required to reproduce them. The ability of this subset to capture ENSO signals can be further explained by the optimized distribution, with some of the selected locations laying in areas strongly affected by ENSO teleconnections, such as the nodes of the Pacific North/South American pattern (Lewis and LeGrande, 2015). This suggests that the CRO-AM tends to prioritize those records strategically situated over major climate teleconnections, which together explain a large fraction of the global temperature variance.

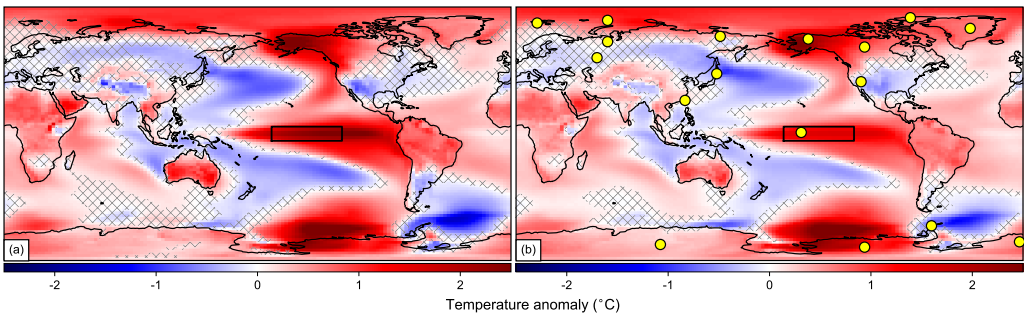

Figure 4.12: Reconstruction skill of internal variability patterns with the CRO-MIN subset of PAGES-2k records. Composite of annual temperature anomalies (°C, with respect to 850-2005 CE) for El Niño events in (a), the target field (the first CESM-LME full-forcing member). (b) The reconstructed field from the CRO-MIN subset of perfect pseudo-proxies of the PAGES-2k network (yellow dots). For each panel, crosses depict non-significant temperature differences at 95% confidence level with respect to its corresponding climatology inferred from a bootstrap of 10000 random samples. El Niño events are defined as years of the target simulation with standardized temperatures above the 95th percentile at El Niño-3.4 region (black square).

Indeed, the CRO-OPT subset can also capture large-scale internal modes of atmospheric circulation variability, such as the Northern Annular Mode (NAM), herein reconstructed from the annual mean sea level pressure (SLP) fields of the best temperature analogue years (Fig. 4.13). The reconstructed NAM series follows the simulated variations in the target run, although with considerable uncertainty, and a tendency to underestimate the target amplitude. The latter suggests that, despite being informative, networks specifically optimized for a given target field do not necessarily represent the optimized solutions for the reconstruction of other fields.

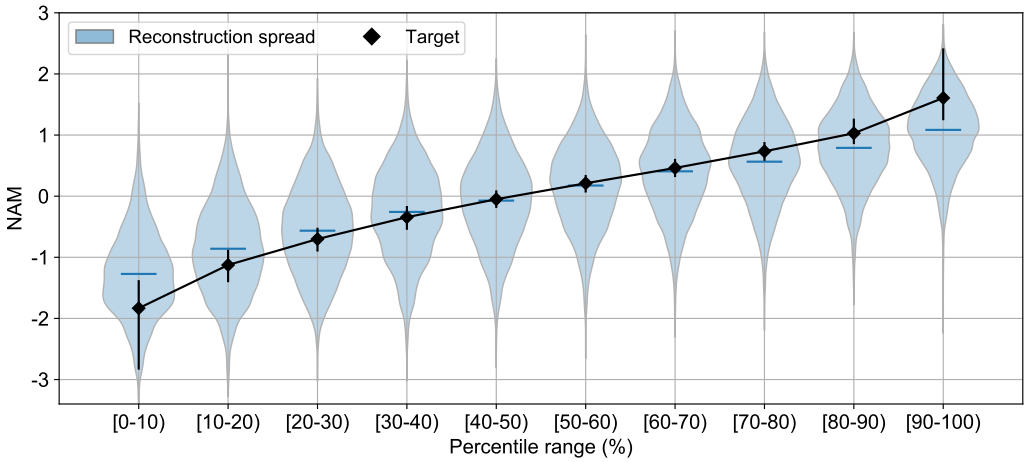

Figure 4.13: Percentile distribution of simulated NAM values in the first CESM-LME member (850-2005 CE) and their corresponding reconstructions from the CRO-OPT subset of the PAGES-2k network. Black diamonds represent the mean simulated NAM for each percentile range, with vertical black lines showing their respective minimum and maximum values. Blue violins show the distribution of the reconstructed NAM values for the same years included in each percentile range and 100 different NAM reconstructions. Mean values of the violin distributions are depicted as horizontal blue lines.

As for the external forcings, the constrained reconstructions from subsets of perfect pseudo-proxies also capture the timing and spatial fingerprint of the simulated cooling response to major volcanic events of the Last Millennium This is seen in Fig. 4.14 where a clear cooling with respect to the mean temperatures of the two previous years appears in the target and reconstructed field from CRO-OPT following the Tambora eruption in 1815. Interestingly, CRO-AM replicates these forced responses by choosing years from the CESM-LME with significant volcanic activity, indicating that the reconstruction method is capturing the physical processes behind the climate response to external forcings.

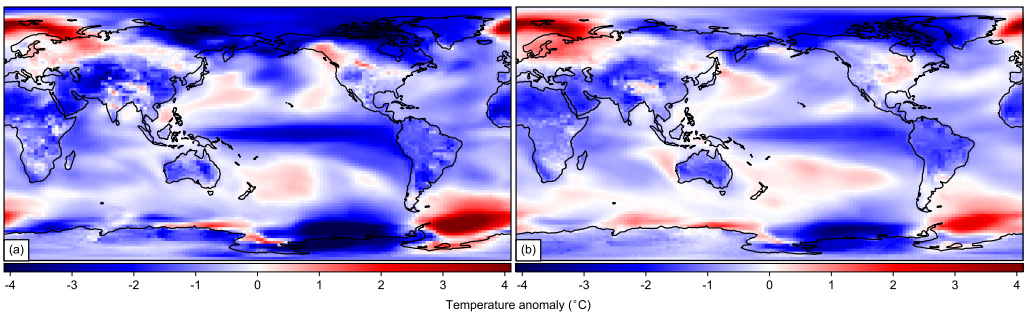

Figure 4.14: Annual temperature anomalies (°C) after Tambora's eruption (1815 CE) in (a) target simulation and (b) CRO-OPT reconstruction. Anomalies are calculated as the difference between the year after the eruption and the mean temperature of the 2 previous years.

This suggests that the CRO-OPT reconstruction can discern the volcanic fingerprint in the global temperature patterns from the internal variability. Therefore, in a more general context, we have assessed if CRO-OPT can detect externally forced responses and attribute them to the responsible forcing. This is done by first assigning each year of the CESM-LME to a dominant factor (i.e. the one with the largest radiative forcing in the top of the atmosphere for that year) by using single forcing simulations of the CESM-LME.

To determine if a particular forcing is dominant for a certain year, we have used the area-weighted mean FSNTOAC (*Clear-sky net solar flux at top of the atmosphere*) from the control simulation and the ensemble of single-forcing simulations for the following available forcings: orbital, solar, volcanic, land use/land cover, greenhouse gas concentrations and ozone. For each year and single-forcing run, we computed the absolute FSNTOAC difference with respect to the 1156-yr mean of the control simulation. By doing this, we estimated the ensemble mean radiative imbalance for each year and forcing. A given year is assigned to an external forcing if two conditions are met: that the forcing has the highest absolute FSNTOAC difference, and this value is above 2 standard deviations of the control mean (to avoid false forcing assignations), which provides a measure of internal variability.

Otherwise, the year is not assigned to any external forcing, this subset comprising years of weak or multiple external forcings, which may obscure the attribution exercise. With this approach, 60% of years on record could not be assigned to a single forcing, whereas 21% and 10% of years were associated with volcanic and solar forcings, respectively. On the other hand, only 2% of years on record have been associated with greenhouse gas emissions (note that this forcing is not relevant before 1850 CE). The remaining forcings, namely land use/land cover (4%), ozone (2%) and orbital (less than 1%), dominated for very few years and are not considered, since their frequency series for the last millennium did not show signals discernible from the internal variability. Note that this attribution method is instantaneous (i.e. it assigns dominant forcings year by year), and not fully independent since forcing years are selected from different simulations of the same model.

Then, for each year of the target simulation we count the detected frequency of each external forcing in the 100 best analogue years selected using CRO-OPT locations and test whether this frequency is significantly larger than that expected by random chance. The results for key forcings of the last millennium (Fig. 4.15) confirm that CRO-OPT is able to detect some forced signals and assign them to the right forcing. This is true for years following large volcanic eruptions (e.g. Tambora 1816) and periods with high volcanic activity (e.g. the 18^{th} century). Similarly, the conspicuous warming of the second half of the 20th century is preferentially reconstructed from years with strong anthropogenic forcing, so that this signal can be attributed to increasing concentration of greenhouse gases. On the contrary, solar signals on annual time scales are not well discernible from the internal variability. Note, however, that it does not mean a lack of solar forcing signals since our approach is instantaneous (based on annual forcings and temperature patterns) and hence it does not take into account lagged or low-frequency responses to external forcings.

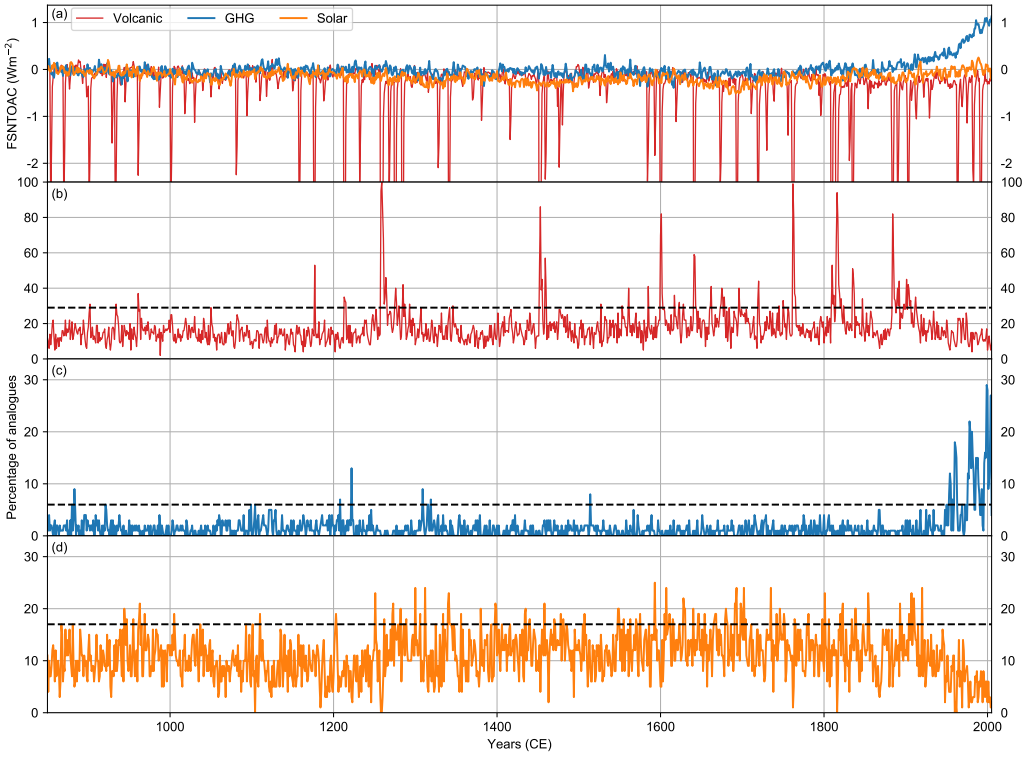

Figure 4.15: Detection of external forcings in the reconstruction with the CRO-OPT subset of the PAGES-2k network. (a) Annual mean clear-sky net solar flux at top of the atmosphere for three single-forcing ensemble simulations. Percentages of the 100 best analogue years selected from CRO-OPT with the same dominant forcing as in the given year of the target simulation. (b) volcanic, (c) greenhouse gases, and (d) solar forcing. Black dashed lines depict the significance thresholds above which there is an instantaneous detection of forced signals attributed to the given forcing.

4.4 Insights on past anomalous periods

We now focus on longer time-scales and explore how well the CRO-AM reconstructions capture key anomalous periods of the past, such as the MCA (*Medieval Climate Anomaly*) (950-1250 CE) (Bradley, 1996) and the LIA (*Little Ice Age*) (1450-1850 CE) (Bradley and Jones, 1993). Although model simulations and proxy-based reconstructions agree on the existence of a global

mean temperature difference between both periods, there are discrepancies in its magnitude and spatial pattern, which are still not well understood (Mann, 2009; Fernández-Donado et al., 2013). Models usually yield a variety of spatial temperature patterns for the MCA-LIA transition as well as weaker differences than proxy-based reconstructions (Fernández-Donado et al., 2013).

Table 4.2: GMT differences between the MCA (950-1250 CE) and the LIA (1450-1850 CE). Area-weighted mean temperatures are calculated globally and for the Northern Hemisphere (NH). The target is the first ensemble member of the CESM-LME. Reconstructions are generated with CRO-AM using perfect pseudo-proxies at the locations of CRO-MIN, CRO-OPT and the full-proxy network of PAGES-2k.

MCA-LIA Temperature (°C)		
Map	**Global**	**NH**
Target	0.19	0.20
CRO-MIN (17)	0.08	0.10
CRO-OPT (120)	0.11	0.13
PAGES-2k (569)	0.09	0.11

The MCA-LIA GMT difference for our target simulation is 0.19 °C, similar to the remaining members of the CESM-LME. However, this value is halved in the reconstructions obtained with CRO-MIN, CRO-OPT and the complete PAGES-2k network, even if we use ideally perfect pseudo-proxies (Table 4.2). The attenuation of the global warming amplitude is further evidenced by the spatial pattern of temperature differences depicted in Fig. 4.16. Note that all regions show milder differences in the reconstruction, especially in the Southern Ocean and part of the Arctic .

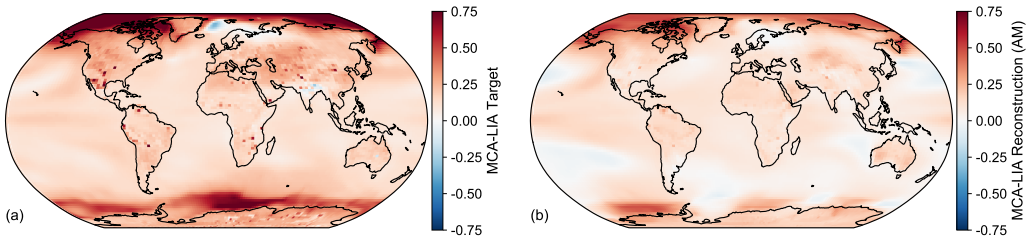

Figure 4.16: Spatial pattern of mean temperature difference (°C) between MCA (950-1250 CE) and LIA (1450-1850 CE) in (a) the target simulation and (b) the CRO-OPT reconstruction.

To deep further into the causes of the MCA-LIA underestimation, we first assess whether they stem from a limited coverage of the current PAGES-2k archive. To address this question, the CRO-AM was run to find the same number of representative locations as in CRO-MIN, but from the full grid of CESM-LME (i.e. without constraining the search to the PAGES-2k network, Free run in Fig. 4.17).

The resulting distribution of selected locations has similarities with CRO-MIN, stressing the relevance of high latitudes (Fig. 4.17b). Therefore, we conclude that climatically representative regions for the last millennium are sufficiently covered by the PAGES-2k archive. This also holds for the MCA and LIA periods, whose reconstruction skills with CRO-MIN are comparable to those obtained for the entire period of the last millennium. Dedicated experiments to find specific subsets of the PAGES-2k network that optimize the CFR for the MCA and LIA separately (MCA and LIA runs in Fig. 4.17) also bear strong resemblance with the distribution obtained in the last millennium experiment (CRO-MIN). Indeed, good analogues of the MCA and LIA global patterns can be found through most of the last millennium, although with reduced probability during exceptionally cold and warm intervals, respectively, including the 20^{th} century (Fig. 4.18). Interestingly, MCA patterns do not resemble those of the 20^{th} century, indicating distinguishable MCA signatures with respect to the organized global warming of the 20^{th} century in the model world.Therefore, the question is if there is any subset

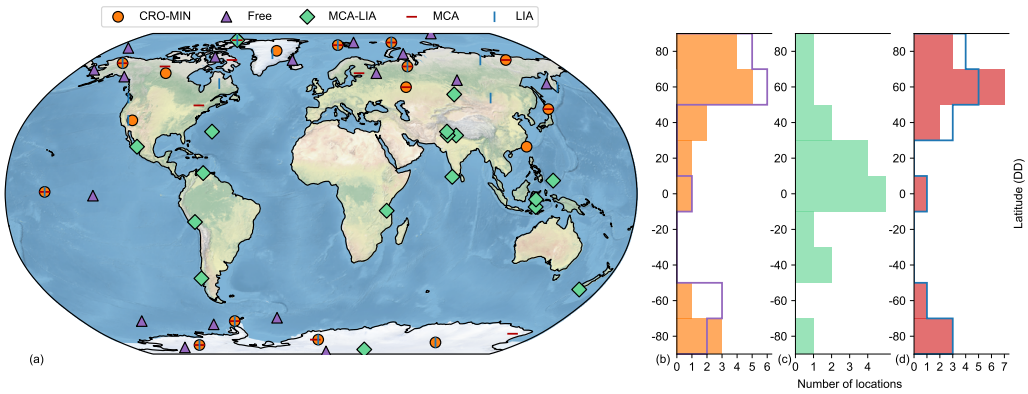

Figure 4.17: Distribution of representative locations for different experiments with perfect pseudo-proxies and the CRO-AM. (a) 2-D and (b-d), latitudinal distribution of optimized subsets of perfect pseudo-proxies (with the same size as CRO-MIN) selected with the CRO-AM for different optimization problems. Optimized reconstruction of the global annual temperature fields of the last millennium from locations constrained to the PAGES-2k network (CRO-MIN, orange) and from an unconstrained selection (Free, purple). Optimized subsets of the PAGES-2k network for the reconstruction of the global annual temperature fields of the MCA (red) and LIA (blue) periods separately, and the spatial pattern of the mean temperature difference between the MCA and LIA (MCA-LIA, green). Latitudinal distribution of (b) CRO-MIN (orange shading) and Free (purple line). (c) MCA-LIA. (d) MCA (red shading) and LIA (blue line).

of PAGES-2k locations that can reproduce the MCA-LIA pattern (i.e. the spatial pattern of temperature differences on multi-centennial scales). This has been accomplished by modifying the health function of the CRO algorithm in order to find subsets of perfect pseudo-proxies that minimize the spatial RMSE of the MCA minus LIA temperature pattern (MCA-LIA run in Fig. 4.17). The results of this experiment reveal that it is possible to find a subset of the same size as CRO-MIN that matches the MCA-LIA pattern with almost the same amplitude as in the target simulation (0.15 °C).

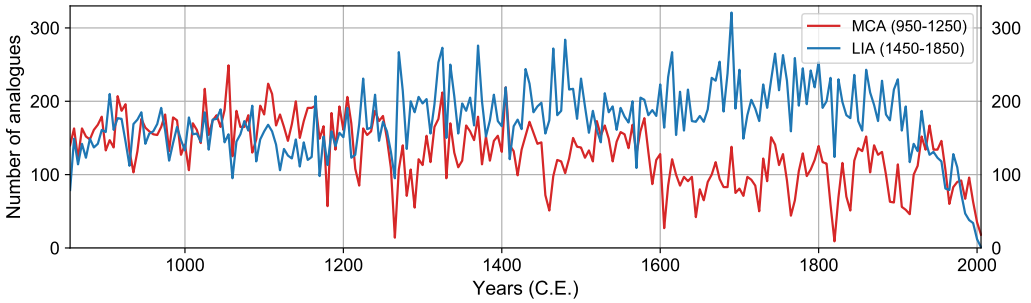

Figure 4.18: Time series with the total frequency of analogues for all years of the MCA (950-1250 CE, red) and LIA (1450-1850 CE, blue) in the CRO-OPT reconstruction. For each year of the MCA and LIA in the target simulation, the 100 best analogues of the annual temperature at the CRO-OPT locations are selected. Their respective years of occurrence are retained and accumulated through the MCA and LIA periods, separately.

However, this improvement in reproducing the MCA-LIA pattern is made by sacrificing the reconstruction skill on annual scales, with the RMSE increasing from 0.65 °C in CRO-MIN to 0.80 °C in the MCA-LIA run. Furthermore, MCA-LIA locations are mainly situated over tropical and extratropical latitudes (Fig. 4.17c), leading to a completely different latitudinal distribution compared to that in the annually-resolved MCA, LIA, CRO-MIN or Free experiments.

These results indicate that perfect pseudo-proxy locations that optimize the reconstruction in the high-frequency do not necessarily bring an improvement in the lower frequencies of the spectrum. To support this statement, we have performed additional experiments where annual temperature fields of the target simulation have been low-pass filtered with 10- and 100-year running windows. The CRO-CCA has been run this time to find optimized subsets of the PAGES-2k network targeting the CFR on each time scale. Figure 4.19 shows how the number of selected records at high latitudes (poleward of 65° N) significantly decreases for lower frequencies.

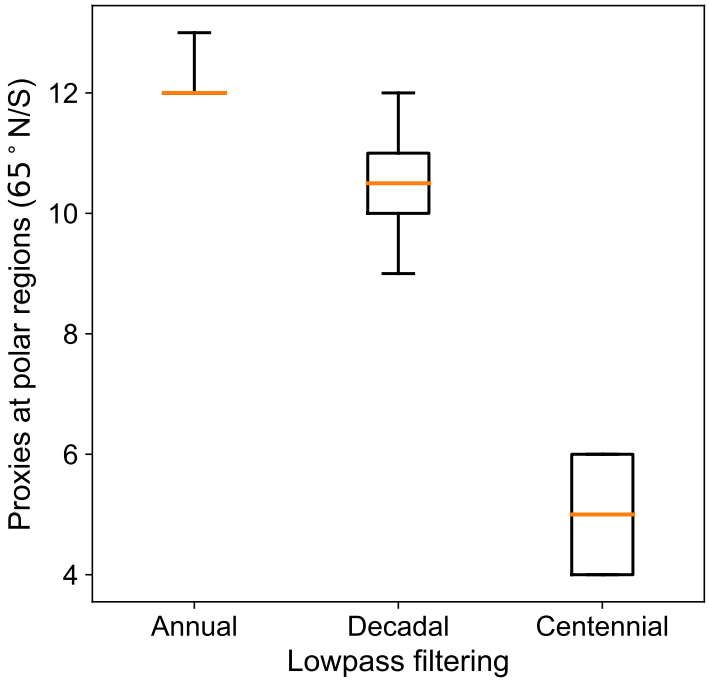

Figure 4.19: Number of records at polar regions (latitudes above 65°N/S) in optimized subsets of 20 perfect pseudo-proxies of the PAGES-2k network for CRO-CCA reconstructions of global temperature fields on different time scales. The CRO-CCA has been run to find subsets of 20 perfect pseudo-proxy locations of the PAGES-2k network that optimize the reconstruction of the global temperature fields of the first CESM-LME member at annual, decadal and centennial time scales by using a 1-, 10- and 100-year low-pass filter smoothing, respectively. The distributions include 200 different optimized solutions from 5 CRO-CCA runs (40 best solutions of each run were kept). Boxes represent the median (orange line) and interquartile ranges of the distribution, with whiskers denoting extreme solutions.

This agrees with the optimized distribution for the reconstruction of the MCA-LIA spatial pattern shown in Fig. 4.17. As the large interannual variability at high latitudes was removed for the low-pass filtered fields, we hypothesize that fewer high-latitude records are needed to reconstruct long term temperature variations. As a consequence, the optimality of the lo-

cations selected for the annual CFR is progressively lost towards the lower frequencies of the spectrum. This is further supported by the power variance spectrum of GMT reconstructions obtained with different subsets of the PAGES-2k network (Fig. 4.20), which shows similar performance of the CRO-OPT and full-proxy reconstructions on longer time scales for all ensemble members.

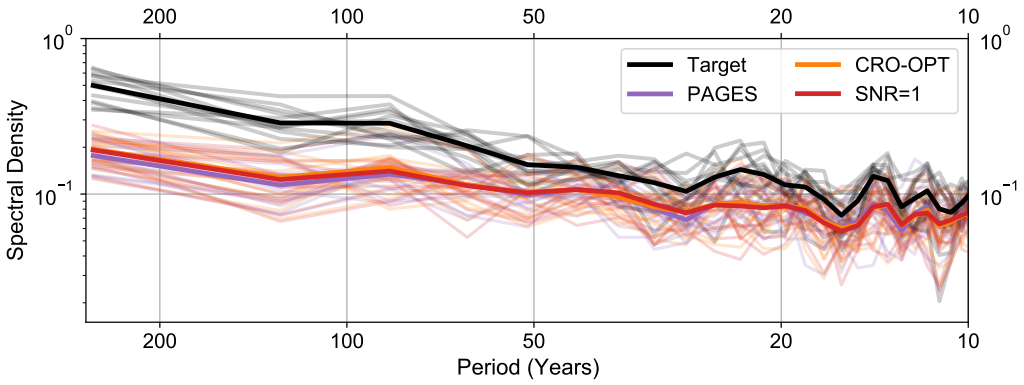

Figure 4.20: Power density spectrum of GMTs series for 850-2005 CE. Color lines show the power spectrum of area-weighted global mean temperature anomalies calculated for the target simulation (black), the CRO-OPT reconstruction (orange), and the CRO-AM reconstructions generated with the full-proxy PAGES-2k network of perfect pseudo-proxies (purple), and an optimized subset of 150 noisy pseudo-proxies with SNR=1 (red). Opaque lines depict the mean spectral density of the ensemble, and transparent lines represent individual spectral densities of all (13) members of the ensemble.

Both networks underestimate the low frequency variations of the true GMT, in agreement with the reduced amplitude of their MCA-LIA reconstructions. Although the ultimate causes are unclear and require dedicated studies in more realistic conditions, our experiments point towards a significant dependence of the optimal subset of PAGES-2k locations on the time scale variations of the target field that are pursued by the CFR.

Keynotes

These are the most important findings of Chapter 4 to take home:

- The bias induced by non-homogeneous distributions is similar to that obtained by adding red noise with the same variance as the climate signal.
- A reduced set of representative proxies generates reconstructions with better skill than using the entire PAGES-2k network.
- There is a significant increase of correlation in reconstructions obtained with proxies at representative locations.
- Annual temperature fields are better reconstructed with locations in high latitudes and teleconnection regions.
- Long-term climate variations are better reconstructed with locations over low and middle latitudes.
- Representative locations depend on the temporal resolution of the target.

Chapter 5

North Atlantic SLP reconstruction since 1750

5.1 Background

Atmospheric conditions are nowadays continuously monitored by systems ranging from land-based meteorological stations to geostationary satellites orbiting the planet. However, the amount of available information from instrumental records drastically decreases the further we go back in time, leading to a problem of data scarcity, particularly acute in the pre-industrial era (before 1850 of the Common Era, CE) and over the oceans (Küttel et al., 2010; Cram et al., 2015; Franke et al., 2017; Brönnimann et al., 2020; Noone et al., 2020). Within this context, new methods that maximize the extraction of information from climate datasets have recently emerged (e.g., Ilyas et al., 2017; Benestad et al., 2019; Carro-Calvo et al., 2020; Kadow et al., 2020; Vaccaro et al., 2021). They represent promising tools to manage large amounts of current information (big data) as well as situations of data scarcity including a better preservation of variability and resolve more accurately the covariance structure for regions with data gaps. In this sense, optimization techniques including evolutionary algorithms (Vrugt and Robinson, 2007; Eiben and Smith, 2015) have been developed in the area of AI to maximize the skill of climate field reconstructions (CFR) as seen in Chapter

4 (Salcedo-Sanz et al., 2019; Jaume-Santero et al., 2020). As it is described in Section 3.2, evolutionary algorithms (Eiben and Smith, 2015) are soft-computing techniques inspired by biological and natural selection processes (Del Ser et al., 2019) which are based on the reproduction and survival of best suited individuals within a competitive environment.

In this chapter, we have coupled a CFR method with the CRO (i.e., the evolutionary algorithm employed in Chapter 4) to obtain optimized high-resolution ($1° \times 1°$) monthly SLP fields over the North Atlantic for 1750-2004 CE from historical land-based observations over Europe, Greenland and North America included in the SLP-Obs dataset (see Subsection 2.1.2). The evolutionary algorithm is designed to find an optimized combination of weights for the observing network that maximizes the reconstruction skill of the SLP field (Section 5.2).This new monthly SLP data set presented herein supersedes the statistically-derived $5° \times 5°$ resolved gridded seasonal SLP dataset of Luterbacher et al. (2002) and Küttel et al. (2010) that cover the eastern North Atlantic Europe and the Mediterranean area back to 1750 CE using terrestrial instrumental pressure series and marine wind information from ship logbooks. In contrast to Luterbacher et al. (2002) and Küttel et al. (2010), we use more station pressure series, provide a higher resolved spatial reconstruction and apply a novel method that can deal more accurately with a low number of station observations, and that can better preserve variability and better resolve the covariance structure, including a more realistic representation of circulation extremes.

The reconstruction performance is compared against other reconstructions obtained using the same set of observations but without optimization (Section 5.3), and allows us to study the extratropical climate variability of the Eastern North Atlantic Ocean and Europe for over the past 255 years focusing on the NAO (*North Atlantic Oscillation*). The NAO is the leading mode of climate variability related to the rearrangement of air mass between subtropical and polar latitudes, and whose alternating phases generate strong changes in surface temperatures, wind, and precipitation over the Atlantic re-

gion and its surroundings (Hurrell and Deser, 2010). Although several NAO indices have been published previously, discrepancies among them (especially present prior to the 19th century where the lack of information is notoriously higher) impairs the study of changes in these phases and their influence on regional climates (e.g., Schmutz et al. (2000); Hernández et al. (2020) and references therein).

In this chapter, we apply the optimized network approach to reconstruct SLP fields back to 1750 CE, that allow to study the change of atmospheric circulation over the Eastern North Atlantic-European area as well as the associated atmospheric action centers such as the AH (*Azores High*) and the Icelandic Low IL (*Icelandic Low*) (Section 5.4). The AH is a persistent subtropical anticyclone situated around the Azores islands that is associated with the descending branch of the Hadley Cell over the Atlantic Ocean (Zishka and Smith, 1980; Davis et al., 1997; Ioannidou and Yau, 2008; Iqbal et al., 2019). In spite of the important role of the AH in the NAO pattern during winter time, and the influence of this high pressure system on European hot temperatures and drier weather in summer (Hasanean, 2004), there are not many high-resolution reconstructions of the AH due to the difficulty of extracting data from that oceanic area. It is therefore crucial to optimize land-based networks of climate observations to maximize the skill over the Atlantic region.

5.2 Optimized networks

Note that the AM in its standalone version (see Subsection 3.1.1) employs the full network of observations, without discrimination among the records, which are considered equally informative (i.e. weights of 1). To test the effect of optimizing the observing network, an additional CRO-AM reconstruction was derived by imposing an optimized set of weights provided by the CRO algorithm (see Section 3.2) during the AM reconstruction. These weights are used in the computation of the metric (i.e., the RMSE), therefore affecting

the selection of the best analogues and the reconstructed field. In that way, the information of the full network is exploited, while acknowledging differences in the predictive skill of individual sites. This approach is novel, and represents an important step with respect to the experiments described in Chapter 4, where a subset of optimal locations was selected (weights equal to 1 or 0), therefore discarding observations that could still provide useful information when the availability of the remaining records is compromised.

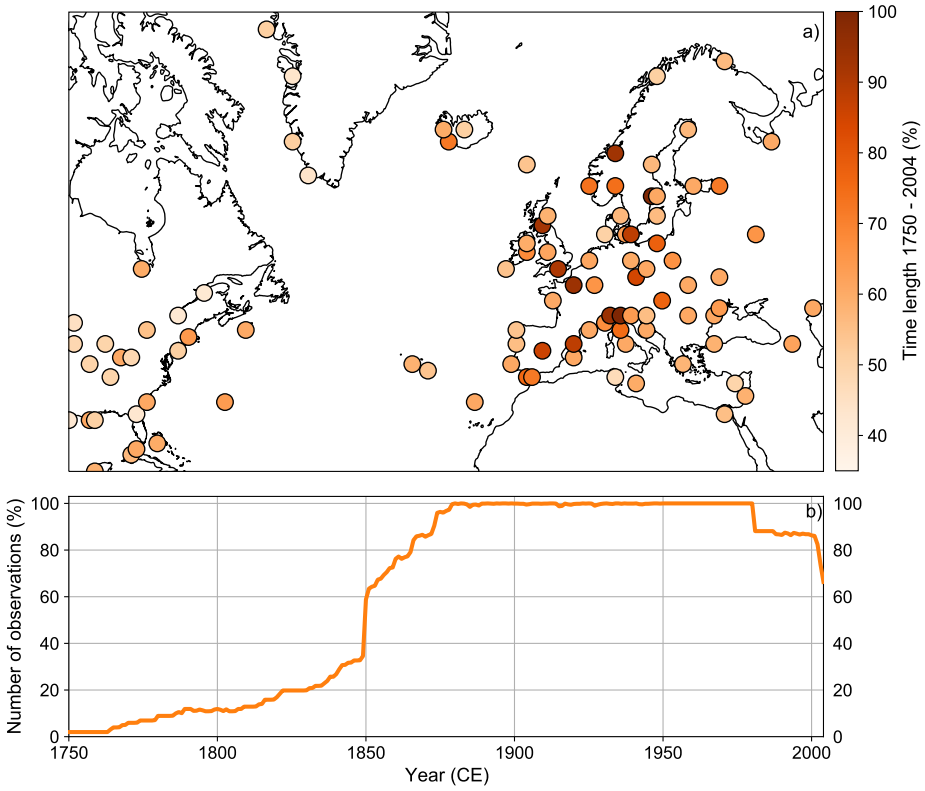

Figure 5.1: Spatio-temporal distribution of SLP observations. (a) Spatial distribution of stations with monthly SLP observations for 1750-2004 CE. Shading shows the percentage of time with available observations over the 1750-2004 CE period, with darker shading indicating longer time series. (b) Evolution of the frequency of observations (in percentage with respect to the total number of stations) for 1750-2004 CE.

Within this framework, weights were optimized for each calendar month of the year by minimizing the area-weighted RMSE of the North Atlantic SLP reconstruction obtained with the AM over the reconstructed period. The performance of the CRO-AM and AM reconstructions, obtained with and without optimized weights, are assessed with respect to the 1836-2004 CE validation period of the 20CRv3 reanalysis, allowing us to quantify the improvements of the optimized reconstruction. While this dataset covers the 1836-2004 CE period of the observations, it does not provide a *ground truth* to optimize the observing network of 1750-1835 CE. Although using the weights of the 1836-2004 CE interval through the entire reconstruction period (1750-2004 CE) is the simplest way to address this issue, it would not take into account changes in the predictive skill of the local records caused by the actual gaps of the earlier period and the pronounced changes in the spatial distribution of the observing network (Fig. 5.1). Therefore, the set of monthly weights were optimized separately for 1750-1835 CE.

To do so, we took the observations of the 1919-2004 CE interval (which comprises the same number of years) and reconstructed their concurrent SLP fields by imposing the same constraints in data availability as in the observing network of the 1750-1835 CE period. In that way, weights can be optimized for a set of records that preserves the spatio-temporal distribution of observations of the earlier period. This set of optimized weights is subsequently applied to the actual observations of 1750-1835 CE to reconstruct the North Atlantic SLP for this earlier period with the AM. This experiment assumes that the relationships between the local records and the large-scale field do not change substantially with time, and that temporal changes in observational errors are not sufficiently large to affect the overall distribution of weights. While these assumptions may not hold for all local records, they were preferred to unrealistic approximations (i.e., no changes in data availability) that are systematically violated by the full network.

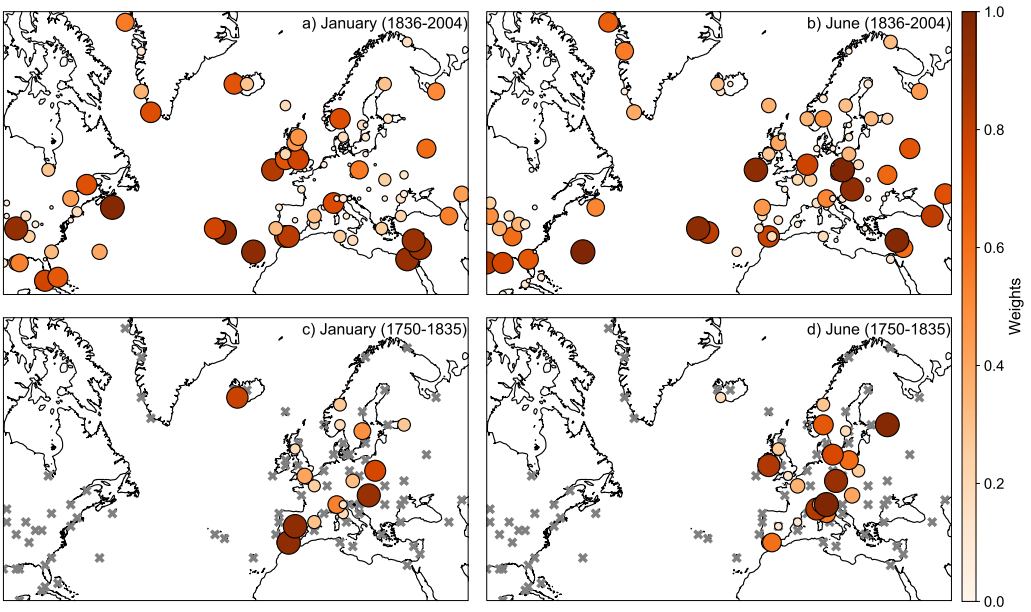

Figure 5.2: Spatial distribution of optimized monthly weights for the observing network obtained with the CRO-AM algorithm. Weights (from 0 to 1) apply to the observing network of (a, c) all Januaries, and (b,d) all Junes of the (a,c) 1750-1835 CE and (b,d) 1836-2004 CE period. The size of the dot is proportional to the magnitude of local weight, which is also indicated by shading. Grey crosses in (c) and (d) represent observations without available information for 1750-1835 CE.

Figure 5.2 shows the spatial distribution of optimized weights for two representative months of the year and the two considered sub-periods: 1836-2004 CE (Fig. 5.2a and b) and 1750-1835 CE (Fig. 5.2c and d). Local weights for the earlier reconstruction period are substantially different from their more recent counterparts. Overall, higher weighting values are assigned to the reduced subset of observations of the 1750-1835 CE period than to the records of the almost complete network of 1836-2004 CE. For instance, ∼85% of the October observations from 1836 to 2004 CE have low weights (below 0.5), whereas this number decreases to ∼70% for the 1750-1835 CE period. Therefore, observations gain representativeness when the lack of information increases, and locations with low weights in a large network can be very informative when considering a reduced subset of the network. In spite

of this, there are no generalized high weights, even in the case of extreme data scarcity. Indeed, for the earlier reconstruction period, only 7 out of 23 records have weights with values above 0.5. Therefore, the CRO algorithm only assigns high values to a few locations, stressing the need for exploiting the information of the full network, particularly when gaps are present. The spatial distribution of weights in the 1836-2004 CE period is preserved for different months of the year, with higher values for latitudes in-between 30°N and 50°N. However, the pattern of weights changes from one month to another for 1750-1835 CE (Fig. 5.2c and d), indicating that a *one-fits-all* pattern of monthly weights can be challenging when there are extensive un-sampled areas. It is difficult to determine whether this seasonal cycle stems from climatological aspects that are not evidenced in larger networks or from peculiarities of the limited network of 1750-1835 CE.

Recall that weights apply to an incomplete network of observations, and hence low weights can be caused by poor instrument calibration, reduced data availability of records (especially in key areas such as the North Atlantic Ocean), an overall weak predictive skill, or redundant information with respect to that provided by the remaining records. Therefore, inferences of physical links among stations (or between local records and the large-scale flow) based on the detailed distribution of local weights within the network can be misleading and should be interpreted with caution. Note also that although changes in data availability were taken into account in the optimization process, weights are time invariant during each reconstructed period (except for the annual cycle). Strictly speaking, weights are expected to change with time, following the configuration of the observing network at any time, and ideally, they should be optimized for each month of the 1750-2004 CE period. However, that approach would be computationally challenging and is not expected to cause large differences in the optimized weights for relatively small changes in the distribution of records, particularly if the number of observations is large. The pronounced changes in the coverage of observations for the earlier reconstruction period leave large un-sampled areas (e.g., the northern part of the North Atlantic and the east

coast of North America) as compared to 1836-2004 CE, justifying a separated optimization.

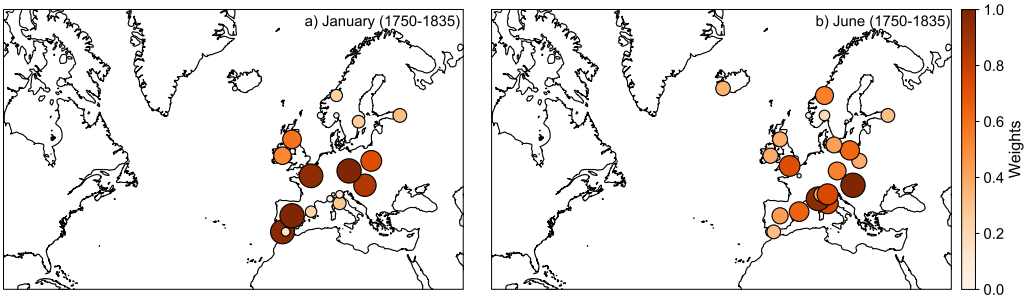

Figure 5.3: As Fig. 5.2 but for the 20CRv3 reanalysis experiment of the CRO-AM reconstruction. Monthly weights for (a) January and (b) June obtained with a perfect (noise free) network of SLP pseudo-observations for the period 1919-2004 CE taken from the 20CRv3 reanalysis with the same gaps as the real observations for the period 1750-1835 CE.

To further test the robustness of our results, we also performed a CRO-AM reconstruction of the 1919-2004 CE SLP fields of the 20CRv3 reanalysis by replacing the observations with the closest reanalysis grid point series and imposing the same spatio-temporal availability as in 1750-1835 CE. Grid point series are not perturbed so that they represent perfect local predictors (SNR $= \infty$) of the 20CRv3 *ground truth*. This subset is also less affected by artifacts (e.g. the mismatch of the spatial scales resolved by the reanalysis and the station-based observations) and is more physically consistent with the large-scale field targeted by the reconstruction. The resulting weights are similar to those found for the observations (Fig. 5.3), stressing the coherence of the station-based and reanalysis grid point series. As the SNR is higher in the reanalysis experiment, the overall agreement also suggests a limited influence of local observational errors on the distribution of weights of the observing network.

Moreover, to verify that the results of the optimization are robust with respect to the datasets employed, a model-based sensitivity experiment was

performed with historical and past1000 simulations from CMIP6-PMIP4 (Eyring et al., 2016; Jungclaus et al., 2017). Both simulations are fully-forced with standard forcing data sets following the specifications included in the input4MIPs documentation (http://goo.gl/r8up31). As the optimization process is time-consuming, a multi-model ensemble was not affordable. Therefore, we used the MRI-ESM2-0 model (Yukimoto et al., 2019), whose spatial resolution (1°) is similar to that of the 20CRv3. Model outputs were also regridded using a bilinear interpolation method to match the resolution of the reanalysis (i.e., grid points at the same latitudes and longitudes). As described in Subsection 2.2.2, we selected the SLP series from 1750 to 1835 of the model grid points that contain the observed records to create pseudo-observations mimicking the characteristics of the observing network (which has the same spatial distribution because of the regridding), and used them to reconstruct the simulated SLP fields of the model. A similar spatial pattern of weights for this model experiment is also found, indicating that the optimization is little affected by the specific realization of internal variability. The model experiment also suggests that model biases in the climatology or fingerprints of external forcings are either small or play a minor role in the optimized weights. These results, and the overall model agreement with the observed distribution of weights for 1750-1835 CE lend support to the approach adopted for inferring the optimized weights of the observing network for that period.

5.3 Skill of the optimized networks

Figure 5.4 summarizes the performance of the CRO-AM optimized networks for 1750-1835 CE and 1836-2004 CE, and compares it to that obtained with the AM only (without weighting). The skill is quantified with the area-weighted RMSE of SLP over the North Atlantic, computed with respect to the corresponding validation period of the 20CRv3 reanalysis employed during the optimization process (1919-2004 CE and 1836-2004 CE, respectively). The optimized network of the 1750-1835 CE period yields area-weighted RMSE below 4 hPa all year round, almost doubling the RMSE retrieved with the much denser network of the 1836-2004 CE period. In both cases, the RMSE displays a pronounced annual cycle with maxima in winter and minima in summer, as expected from the seasonal changes in variability of the North Atlantic atmospheric circulation. The optimized networks have higher skill than their unweighted counterparts for all months of the year. Indeed, for some months the RMSE of the optimized but sparse network of 1750-1835 CE approaches that of the denser unweighted network of 1835-2004 CE, and the weighting improvement is half of that obtained from a 4-fold increase in data density. Note that the improved skill is present for most time-steps of the reconstruction as seen in Fig. 5.5, indicating that the network optimization performs well regardless of the number and distribution of observations.

Although optimized and non-optimized reconstructions have significant correlations ($p < 0.05$) in both periods (Fig. 5.6) with respect to the target field (20CRv3), CRO-AM reconstructions perform significantly better than AM reconstructions for most regions, especially over the North Atlantic Ocean and the Canadian North Pole, where Pearson correlation coefficients increase by up to 0.2 above values obtained without optimization (Fig. 5.4a and b).

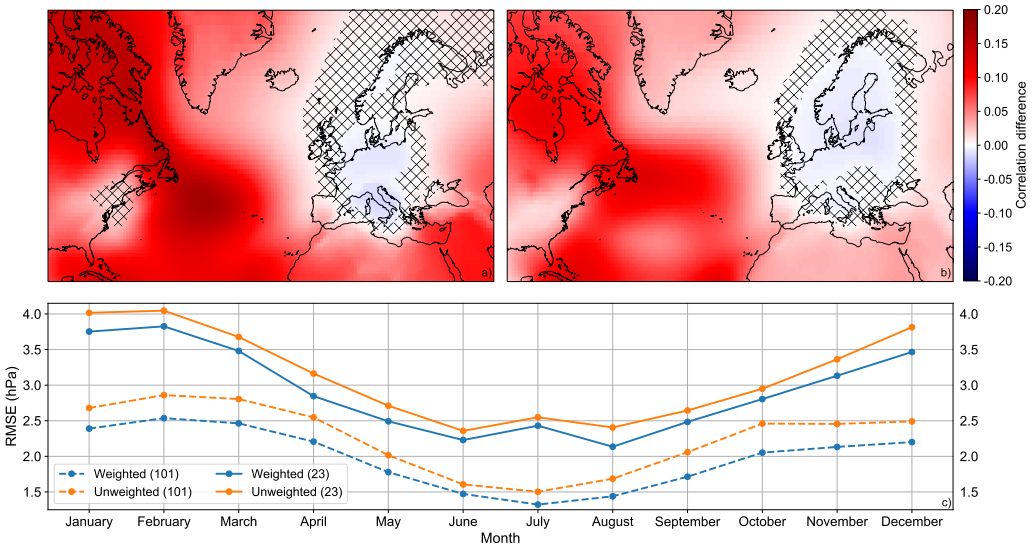

Figure 5.4: Comparison of the SLP reconstruction skill obtained with and without optimization. (Top panels) Difference of performance (Pearson correlation coefficient with the 20CRv3 reanalysis) between the CRO-AM and AM SLP reconstructions generated with the observing network of: (a) 1750-1835 CE; (b) 1836-2004 CE. Crossed regions show non-significant differences ($p>0.05$). (c) Monthly mean evolution of the area-weighted root-mean-square error of the North Atlantic SLP (with respect to 20CRv3) for the CRO-AM (blue) and AM (orange) reconstruction and the observing network of 1750-1835 CE (solid) and 1836-2004 CE (dashed). The performance of the 1750-1835 CE network is evaluated over the 1919-2004 CE period of the reanalysis.

As a matter of fact, the only regions where the reconstruction is not improved by network weighting are those with a higher density of stations (e.g., specific areas of Europe). This indicates that the skill of the AM reconstruction is biased towards well sampled regions, and at the expense of sacrificing its performance over regions with a sparser distribution of observations. Differently, the optimized networks maximize the reconstruction skill of the whole study region, reducing the spatial bias induced by the non-homogeneous and ever-changing distribution of climate observations.

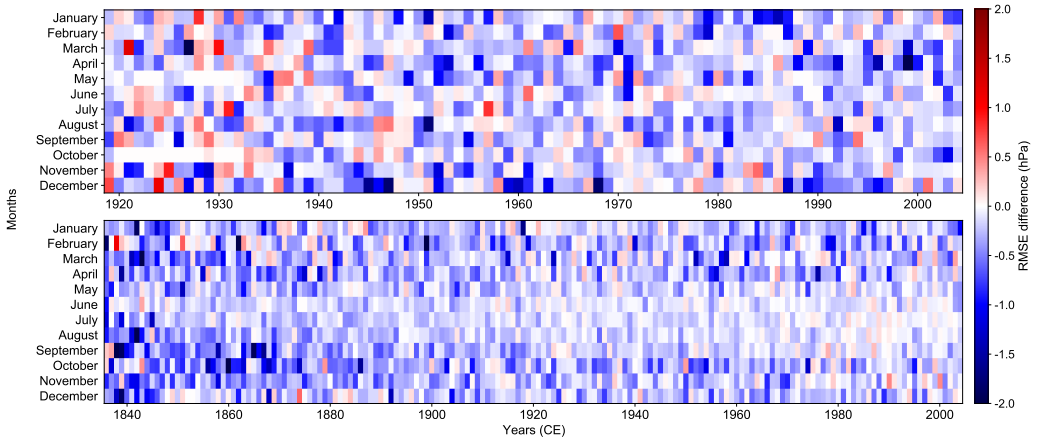

Figure 5.5: Area-weighted RMSE difference between monthly SLP reconstructions generated with and without optimization of the observing network. The area-weighted RMSE of the North Atlantic SLP reconstructions generated with the observing network of 1750-1835 (1836-2004) CE is calculated with respect to the 1919-2004 (1836-2004) CE period of the 20CRv3 reanalysis, and shown in the top (bottom) panel. Blue (red) colors indicate that the optimized reconstruction has lower (higher) RMSE than its non-optimized counterpart.

This is consistent with findings in Chapter 4 where pseudo-proxy reconstructions of global temperature fields from reduced sets of representative locations were improved at the expense of losing skill in over-sampled regions. The optimized reconstruction in the MRI model experiment shows the same pattern of improvement (note the similarity between Fig. 5.7 and Fig. 5.4a), confirming that the reduction of biases in under-sampled regions is a robust feature of the optimized network, and relatively insensitive to high- and low-frequency changes in the background state. Interestingly, and despite the large differences in the distribution of weights for the observing networks of the early and late sub-periods (Fig. 5.2), they bring very similar patterns of improvement (Fig. 5.4). In particular, some of the largest increases in skill are observed over the southern half of the North Atlantic Ocean.

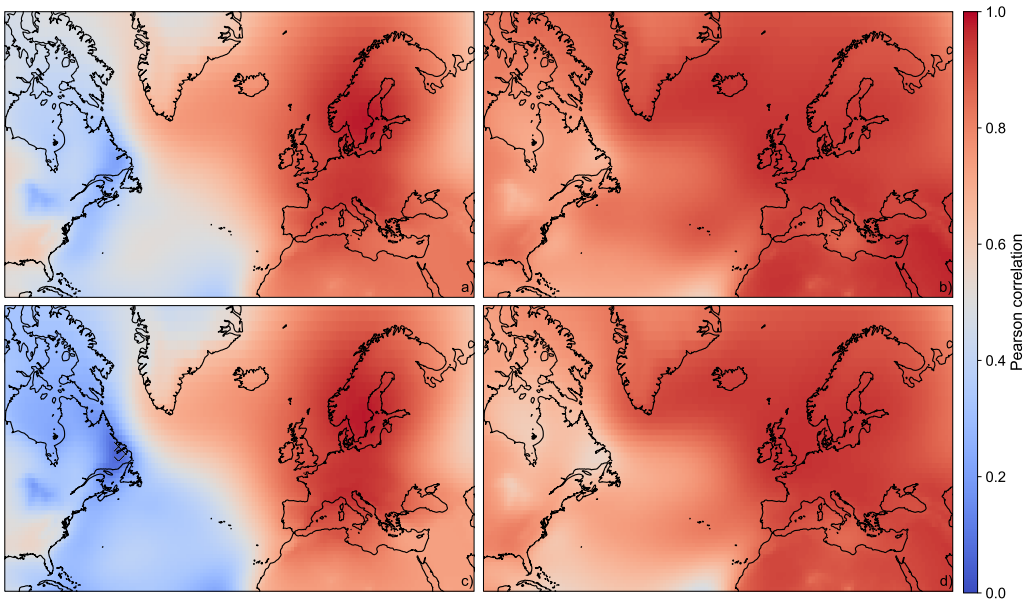

Figure 5.6: Pearson correlation between SLP target fields and their optimized and non-optimized reconstructions. Pearson correlation coefficient with the 20CRv3 reanalysis between the CRO-AM (a and b) and AM (c and d) SLP reconstructions generated with the observing network of: (a and c) 1750-1835 CE; (b and d) 1836-2004 CE. All grid-point correlations are significant for a 95% confidence interval ($p<0.05$).

Being far enough from major continental areas and the well sampled European territories to yield skillful reconstructions, this region has often been disregarded in previous reconstructions. However, regional SLP variations in this area are of paramount importance for the climate of Europe and North and Central America by modulating the southern lobe of the NAO in winter and the intensity and location of the Azores-Bermuda Subtropical High in summer (Davis et al., 1997; Portis et al., 2001). The following sections focus on these seasonal aspects of the atmospheric circulation.

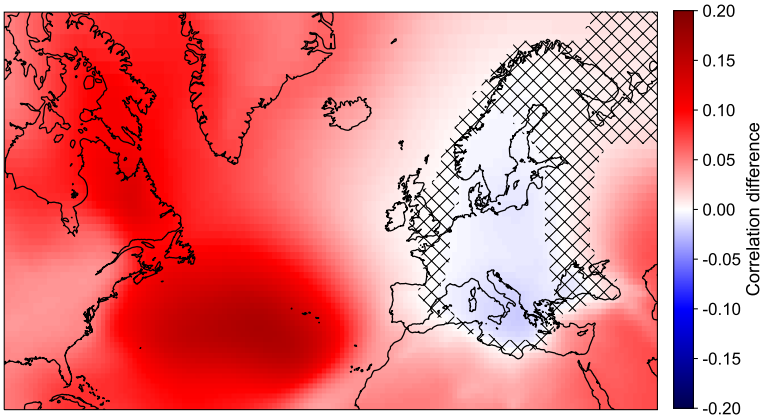

Figure 5.7: As the top panels of Fig. 5.4 but for the SLP pseudo-reconstructions (1750-1835 CE) of the MRI-ESM2-0 model. Shading shows the difference in performance (Pearson correlation coefficient with the 1750-1835 CE targeted SLP field of the MRI-ESM2-0 model) between the SLP pseudo-reconstructions generated with and without optimization of the pseudo-observing network (the simulated SLP series at the grid points matching the locations and availability of the observing network for 1750-1835 CE).

5.4 Climate variability and the Azores High

After demonstrating the added value of network weighting for the validation period of the CRO-AM reconstructions, we generated the SLP field reconstructions (CRO-SLP, hereafter) of the 1750-2004 CE period with the observations of the optimized networks for 1750-1835 and 1836-2004 CE. They yield more than 250 years of high-resolution monthly SLP fields over the North Atlantic, allowing us to study the major components of the North Atlantic atmospheric circulation that govern the climate conditions of its surroundings (i.e., Europe, Greenland, and the east coast of North America). Such is the case of the winter NAO (Hurrell and Deser, 2010) and the East Atlantic (EA) pattern (Barnston and Livezey, 1987; Mellado-Cano et al., 2019), the first and second leading modes of climate variability over the region. By using the CRO-SLP reconstruction from 1751 to 2004 CE, and the 20CRv3 reanalysis from 1836 to 2014, we derived the winter (DJF) in-

dices of the NAO and EA for both datasets, defined as the first and second PC (*Principal Components*) of standardized SLP fields over [95°W - 50°E] and [20 - 73]°N. While the first EOF (*Empirical Orthogonal Function*) (associated with the winter NAO) explains 49% of the total variance, the second EOF (associated with the EA pattern) represents the 19.5%.

The same process was employed to generate the EA index from the seasonal SLP $5° \times 5°$ reconstruction of Küttel et al. (2010), which is provided for 1750-1886 CE and as an extension of the HadSLP2 (Allan and Ansell, 2006). Deriving a PC-based NAO index in Küttel et al. (2010) was hampered by its limited spatial coverage (the reconstruction stops at 40°W), and hence it was better obtained as the standardized SLP difference between Azores and Iceland, these regions being defined as the mean of the four closest grid points on the $5° \times 5°$ grid (Luterbacher et al., 2001). Our indices have also been compared against other NAO and EA instrumental-based indices (see Table 5.1) from previous studies (Jones et al., 1997; Luterbacher et al., 2001; Comas-Bru and Hernández, 2018), obtaining statistically significant correlations ($p < 0.05$) with all of them, as shown in Table 5.2.

Overall, the correlations are significantly higher for the NAO than for the EA index, arguably due to the degraded skill of the CRO-SLP reconstruction over Europe further back in time. This would affect the node of the EA index as well as the European node in the dipolar definitions of the EA. Note that, despite the diversity of data and methodologies employed in the definition of these indices, in all cases the CRO-SLP NAO and EA indices yield higher correlations than their counterparts obtained from the reconstruction of Küttel et al. (2010) that used wind records from ship logbooks over the ocean, in addition to many of the land-based observations. Although the comparison across indices must be taken with caution, and higher correlations do not necessarily involve better reconstructions, these results suggest that optimized networks of land-based observations might eventually outperform non-optimized networks including land and ocean records.

Table 5.1: Definition of NAO and EA indices. Time series have been obtained from the sources indicated below (when provided) or calculated from the spatially resolved fields as detailed in the second column. All indices have been re-standardized with respect to 1951-2000 CE.

Index	Definition	Period
	NAO	
Jones et al. (1997)	Normalized pressure difference between Azores and Iceland	1825-2014
Luterbacher et al. (2001)	Standardized SLP difference between Azores and Iceland, each region defined as the mean of 4 grid points on a 5° x 5° grid	1750-2001
Küttel et al. (2010)	Standardized SLP difference between Azores and Iceland, each region defined the mean of 4 grid points on a 5° x 5° grid	1750-2004
Hurrell and Deser (2010)	First principal component of standardized SLP anomalies (20°-80°N, 90°W-40°E)	1899-2019
	EA	
Küttel et al. (2010)	Second principal component of standardized SLP anomalies (20°-70°N, 40°W-50°E)	1750-2004
Comas-Bru and Hernández (2018)	Composite of EA series generated from historical records at Bergen Florida and Valentia, and five reanalyses	1852-2014

As the CRO-SLP reconstruction brings the largest improvement over the AH pressure system, we performed a more detailed assessment of this subtropical high for 1750-2004 CE. In CRO-SLP (Fig. 5.8a, b), the AH is readily identified from the seasonally averaged SLP fields of all winters and summers of the 1751-2004 CE period. Spatial patterns of the AH show significant seasonal differences, exhibiting a wider high pressure center across the Atlantic for summer, and a weaker system for winter as described in previous findings (Davis et al., 1997; Wanner et al., 1997; Portis et al., 2001; Küttel et al., 2010). The 1750-2004 CE evolution of the seasonal AH is depicted in Fig. 5.8c and d. Its intensity has been defined as the maximum $5° \times 5°$ mean SLP within the [20 - 55]°N and [10 - 70]°W domain. These criteria were chosen to facilitate the identification of the AH center and avoid misdetections, but the results are relatively robust to small changes in the selected domain. There are seasonal differences in the interannual evolution of the AH pressure

Table 5.2: Pearson correlation coefficients of winter NAO and EA indices. Correlations have been calculated for the overlapping interval of each pair of indices within the 1751-1886 CE period (to avoid chucks in some of the series that were filled or extended with observations from other datasets). Coefficients in bold are statistically significant at the 95% confidence interval. Information about the NAO and EA indices can be found in Table 5.1. All indices are standardized with respect to 1951-2000 CE.

	CRO-SLP	Küttel et al. (2010)
NAO		
Jones et al. (1997) (1825-1886)	**0.92**	**0.77**
Luterbacher et al. (2001) (1751 - 1886)	**0.75**	**0.60**
EA		
Comas-Bru and Hernández (2018) (1852-1886)	**0.73**	0.28

system, with summers yielding quite stable pressures around 1024 hPa, and winters displaying larger variability on interannual and longer time scales, including a long-term trend towards the end of the analyzed period.

To place this trend in the context of the last 250 years, we have computed trends of the winter AH intensity for running windows of 50 years from 1751 to 2004 CE (Fig. 5.9) with CRO-SLP, from 1837 to 2013 CE with 20CRv3, and from 1899 to 2019 CE with NCAR SLP (Hurrell et al., 2020). In winter, decadal trends are comparatively smaller and show no large variations from 1751 to 1900 CE, being followed by an increase during the mid-1920s, a decrease during the 1940s, and a sharp increase during the 1960s onwards, in agreement with Davis et al. (1997) and Hasanean (2004). The last change is concurrent with the prominent positive trend of the winter NAO from the 1960s to the 1990s (Pinto and Raible, 2012). Associated impacts of the recent AH strengthening have already been reported (Falarz, 2019). The CRO-SLP captures this trend and further reveals that it is unprecedented

since 1750 CE, with the last 50 years exhibiting the largest intensification of the winter AH pressure system (0.55±0.19 hPa/10y in 2002) of the last two and half centuries. In contrast, a recent intensification of the AH center is not observed during summer, although some studies using stream function as a diagnostic have reported a strengthening and westward movement of its western ridge over North America (Li et al., 2012).

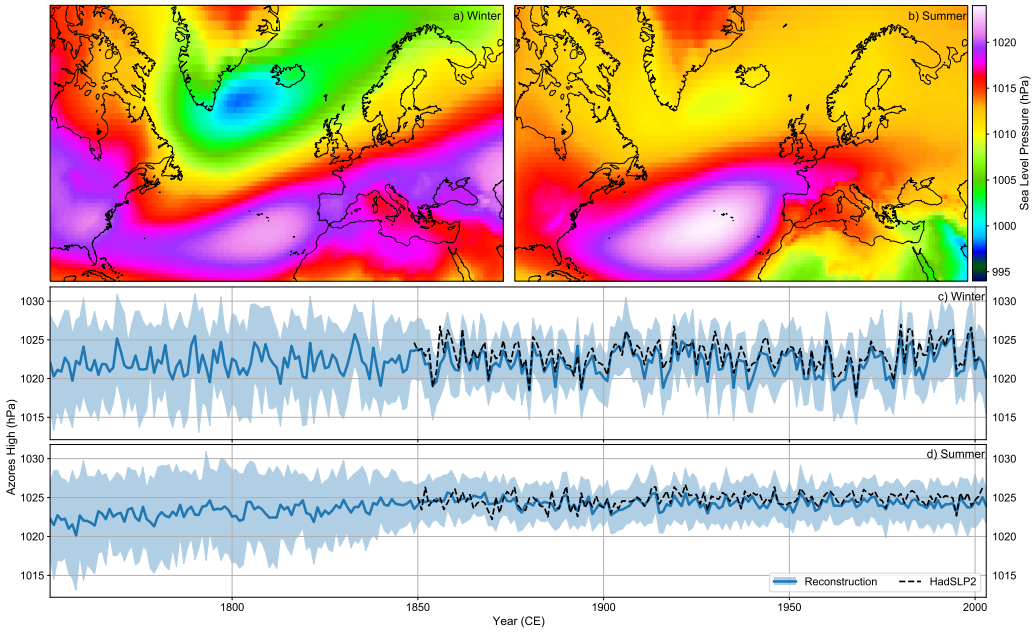

Figure 5.8: SLP variability over the Azores High region from 1750 to 2004 CE. Climatological (1750-2004 CE) mean SLP (shading, in hPa) obtained with the optimized reconstruction for: (a) winter (DJF) and (b) summer (JJA). Seasonal mean time series (1750-2004 CE) of the intensity of the (c) winter and (d) summer Azores High (blue lines, in hPa). Shading shows the uncertainty range calculated as two standard deviations over the $5° \times 5°$ area where the maximum of SLP is located. Dashed lines illustrate the Azores High intensity for the HadSLP2 for the 1850-2004 CE period.

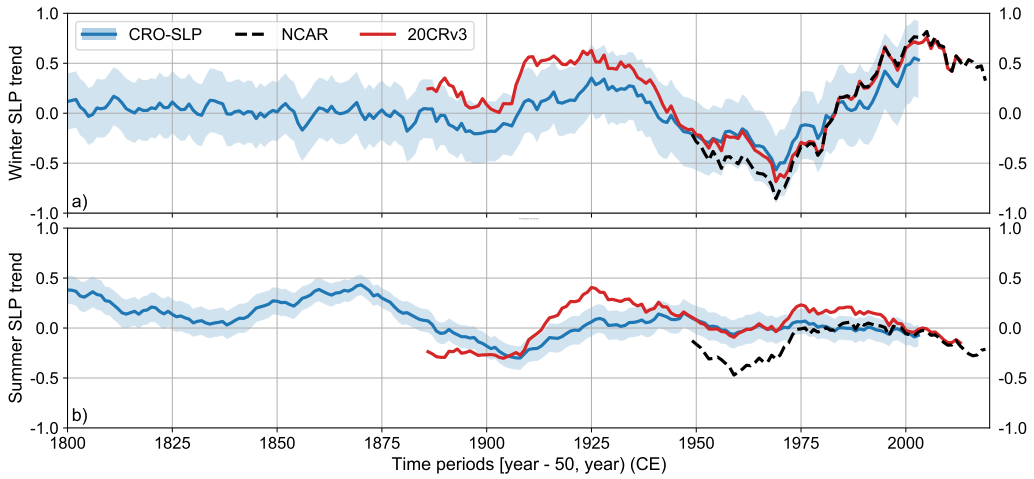

Figure 5.9: Decadal SLP trends of the Azores High pressure center intensity. Decadal linear trends of the (a) winter and (b) summer SLP (blue line, in hPa per decade) obtained from the CRO-SLP reconstruction (blue line), the 20CRv3 reanalysis (red line), and the NCAR SLP dataset (dashed line), over the Azores High center for 50-year running windows from 1750 to 2019 CE. Blue shading illustrates the uncertainty range of CRO-SLP as two standard deviations with respect to the mean.

The most striking feature of the long-term evolution of the summer AH is an overall intensification since the second half of the 18th century and during most of the 19th century (Fig. 5.8d and 5.9b). However, this trend coincides with the period of largest uncertainties, and hence the overall AH weakening towards the beginning of the analyzed period may result from limitations of the observing network (e.g., analogue fields poorly constrained by the scarcity of observations). Moreover, we have also assessed the contribution of AH variations to the historical evolution of the NAO. To do so, the NAO index was decomposed as the standardized sum of its AH and IL components. They have been obtained separately as the projection of the winter SLP anomalies onto the grid points where the NAO pattern was strictly positive (AH) and negative (IL), respectively. The first EOFs associated with both projections describe the 55.4% of the total variance over the AH region and the 49.7% over the IL region.

The sum of these AH and IL indices has a correlation (R^2) of 0.99 and a RMSE of 0.11 with respect to the original CRO-SLP NAO index used in Table 5.2 for the 1751-2004 CE period. This linear behavior allows us to quantify the AH and IL contributions to the NAO of each winter, and discern the dominant component through the 1750-2004 CE period. The leading contributor is easily identified from the absolute values of the AH and IL indices. Fig. 5.10a shows the time series of |AH| minus |IL|, being this difference negative if the IL was predominant for a certain year and positive if the AH was the dominant one. Although AH and IL indices are strongly anti-correlated, there are few years with almost equal contributions to the NAO (e.g., |AH| - |IL| $\in$ (-0.1, 0.1) for 6.7% of winters during the 1751-2004 time period). Although AH and IL dominant years tend to alternate without a clear pattern of long-term trends, the time series displays some low-frequency variability, with e.g. the AH (IL) dominating the NAO over the second half of the 18^{th} century (the first half of the 19^{th} century), which does not translate into concurrent variations in the sign of the NAO index (Fig. A.3). Some of these periods are more evident at the beginning of the series, and may be affected by larger uncertainties of the CRO-SLP reconstruction at that time (Fig. 5.8).

Overall, Fig. 5.10a illustrates that differences in the reconstruction skill of the AH and IL would cause time-varying uncertainties with an impact on the magnitude and even the sign of the NAO (note that 45% of the winters have absolute differences larger than 0.5 standard deviations). To further address this issue, we have calculated the spread of the historical NAO time series across the different NAO indices defined in Table 5.1 (the PC-based NAO from the 20CRv3 reanalysis has also been included). The analysis has been restricted to 1900-2004 CE, since the number of available indices decreases backwards. Interestingly, the evolution of the NAO spread from 1900 to 2004 CE tends to follow that of the |AH| - |IL| difference in AH dominant years (|AH| - |IL|>0), indicating that the more dominant the AH was with respect to the IL, the higher the differences between NAO reconstructions.

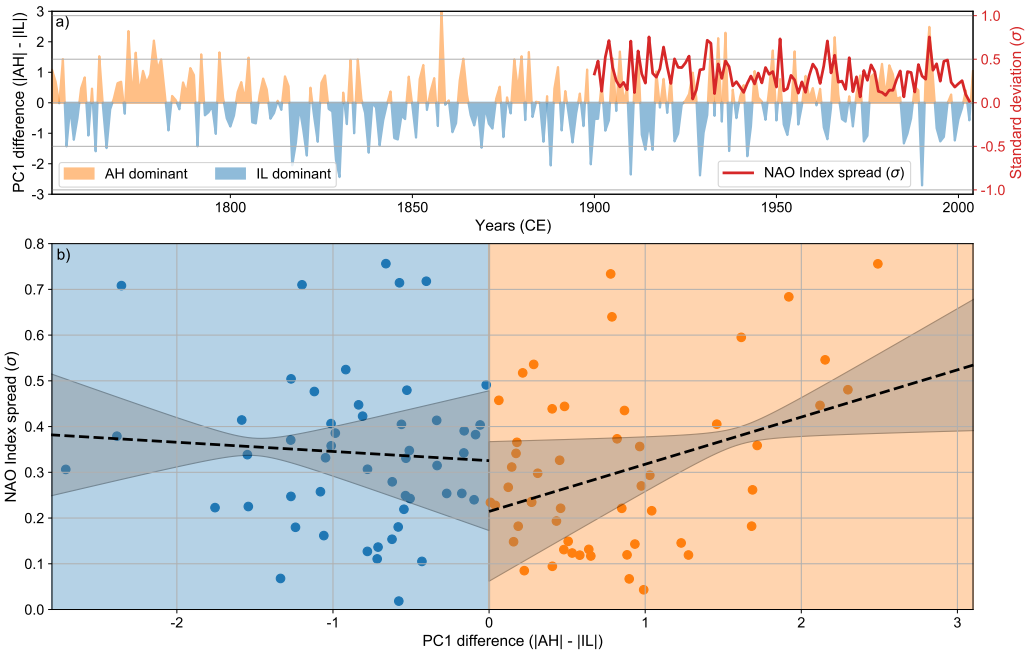

Figure 5.10: Contribution of the Azores High and Iceland Low to the winter NAO. (a) Time series (1751-2004 CE) of the winter |AH|-|IL| index, denoting the unbalanced contribution of the Azores High and Iceland Low anomalies to the winter NAO. Positive (orange) and negative (blue) values show winters with a leading influence of the Azores High and Iceland Low, respectively. The red line represents the spread (standard deviation) of NAO indices for each winter of the 1900-2004 CE period, calculated from a suite of instrumental-based NAO indices standardized with respect to 1951-2000 CE (Table 5.1). (b) Scatter plot (1850-2004 CE) of the spread of NAO indices for winters dominated by the Iceland Low (blue section) and the Azores High (orange section). Dashed lines represent separate linear regressions for each dominant component. Grey shading shows the 95% confidence interval of the linear fits. All series have been standardized with respect to the 1951-2000 CE baseline.

Indeed, the Pearson correlation coefficient between the NAO spread and |AH| - |IL| series is 0.24, but increases to 0.36 ($p<0.01$) for AH dominant years. This is better illustrated in Fig. 5.10b where there is a positive trend in

NAO spread for AH dominant years, and no significant correlation for IL dominant periods. Accordingly, NAO indices tend to show better agreement in years dominated by the IL and higher discrepancies for years when the NAO was largely determined by AH anomalies. Part of the NAO spread is expected to arise from differences in the NAO definition. However, we still find significant correlations when one NAO index is dropped from the spread, with a minimum correlation of 0.21 if the 20CRv3 is not included and a maximum correlation of 0.42 if Jones et al. (1997) is not considered (see Table B.2). Therefore, uncertainties in the AH represent an important source of disagreement for instrumental NAO series. As these NAO indices are obtained from station-based observations or instrumental SLP fields, the result points to different levels of performance in these datasets to capture the winter AH pressure system. This stresses the added value of the CRO-SLP reconstruction, which brings a significant increase in the SLP skill over the AH region (Fig. 5.2), and of optimization as a way to overcome potential shortcomings affecting instrumental datasets.

In summer, the AH becomes detached from the summer NAO and increases its areal extent and intensity. Although interannual changes in intensity are relatively small (Fig. 5.8d), variations in location or extension can be pronounced and affect the climate conditions of the surrounding continental regions in subtropical and mid latitudes. Thanks to the high resolution of CRO-SLP (Iles et al., 2020), it has been possible to trace the AH center from 1750 to 2004 CE, defined as the central location of the $5° \times 5°$ box with maximum averaged SLP, among those within the [20 - 55]°N and [10 - 70]°W domain. The results indicate that the center of action has not experienced long-term changes, being usually situated within [34 - 39]°N and [26 - 39]°W. Despite the relatively stable locations of the summer AH over the 250 years, we found some pronounced excursions. The largest one corresponds to an extreme shift towards the north-east (43°N and 18°W) in summer of 1783 CE (Fig. 5.11).

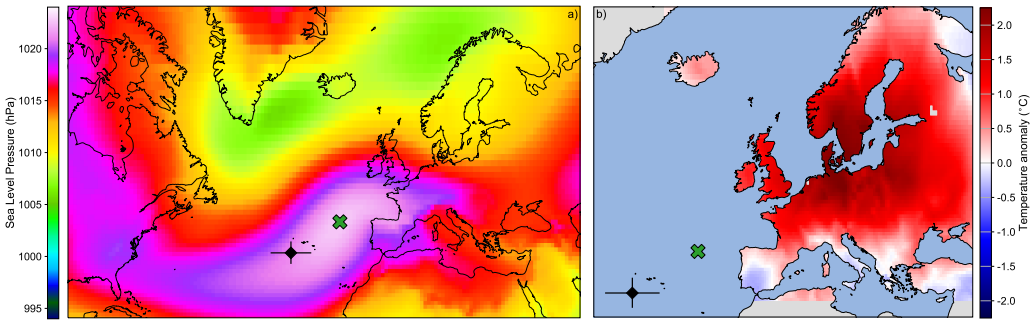

Figure 5.11: Azores High shift in the extremely warm summer of 1783 CE. (a) Summer mean SLP (shading, in hPa) for 1783 CE obtained from the optimized reconstruction. (b) Summer mean temperature anomalies (in °C, with respect to 1500-2002 CE) for 1783 CE. Black diamonds with error bars show the climatological (1750-2002 CE) location (mean and two standard deviations) of the Azores High center. Green crosses represent the center of the Azores High for the summer of 1783 CE.

This year is remembered by the great dry fog in Europe (Stothers, 1996; Thordarson and Self, 2003; Schmidt et al., 2012) after the Laki eruption (Iceland) in June, and the ∼3°C cooling during the following winters. In contrast, reconstructed temperatures (Luterbacher et al., 2004) for the summer of 1783 CE show an European-mean warming of ∼3°C, being particularly pronounced in western Europe. Previous studies have already acknowledged the difficulty of GCMs to reproduce such warming event as a fast response to the volcanic forcing (Zambri et al., 2019), and have rather associated this extreme summer to persistent atmospheric blocking conditions, more likely caused by internal variability. Our results only partially support this hypothesis. While blocking events often cause extremely warm conditions in Europe (Barriopedro et al., 2011), they typically occur in northern latitudes of the continent and are rarely accompanied by anomalies in the summer AH such as those revealed by the CRO-SLP reconstruction. The latter are more typically associated with meridional excursions of subtropical air masses towards western Europe, which can cause simultaneous extreme conditions over a large range of latitudes (e.g., Sousa et al. (2018, 2019); the 2003 mega-heatwave, or the more recent 2019 European heatwave). Consistently, Fig.

5.11a shows how the AH pattern obtained from CRO-SLP was abnormally elongated towards the north-east during that summer, resulting on higher-than-normal SLP values over western Europe that are in good agreement with the warming inferred from independent temperature reconstructions. The meridional excursion of the summer AH is among the largest ones in our 250 year-long record, which could explain why extreme temperatures reached unusual poleward latitudes, exceeding the record-breaking values of the 2019 warm air intrusion reported so far (Sousa et al., 2019).

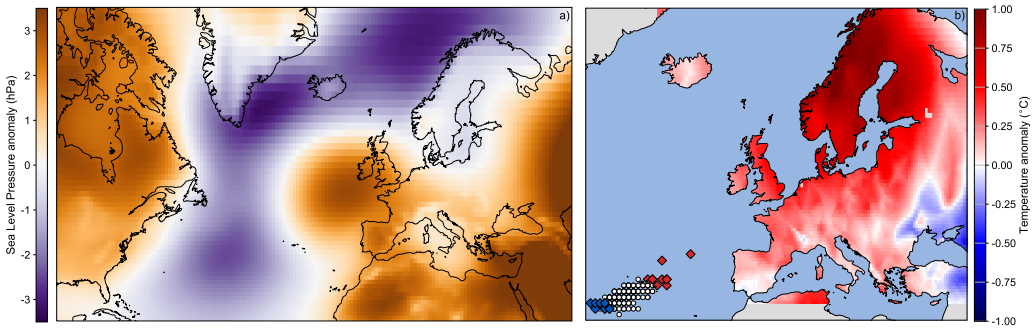

Figure 5.12: SLP and temperature difference between the top ten north-easternmost minus top ten southwesternmost summer AH centers from 1750 to 2002 CE. (a) Summer mean SLP difference obtained from CRO-SLP. (b) Summer mean temperature anomalies (in °C, with respect to 1500-2002 CE). White dots show the location of the Azores High center. Red (blue) diamonds represent the center of the Azores High for the top ten northeasternmost (southwesternmost) summers.

Similar patterns are obtained when comparing mean SLP and temperature fields from summers with AH centers situated at the top ten northeasternmost vs southwesternmost locations for the 1750-2002 CE period (Fig. 5.12), confirming the general increase in European temperatures (especially in Northern Europe) associated with displacements of this high pressure system. Future projections also indicate an intensification, poleward shift and expansion of the summer AH, particularly towards the north-west and secondarily the north-east (He et al., 2017; Cherchi et al., 2018). Although

significant trends in the latter are not detected yet, Figs. 5.11 and 5.12 may represent an example of European summers that are still to come.

Keynotes

These are the most important findings of Chapter 5:

- Evolutionary algorithms maximize the reconstruction skill of SLP fields over the study region.
- General improvement in reconstructions are at the expense of sacrificing skill at over-sampled locations which is compensated by larger improvements over representative regions.
- The reconstruction of SLP over the North Atlantic allows to study the NAO and its associated action centers.
- There is a positive winter AH intensity trend above 0.5 hPa per decade during the second half of the 20^{th} century.
- Differences in reconstructions of the NAO are partially explained by disparities on the reconstruction of the AH.
- Displacements of the AH center towards the north-east caused extreme warming events in Western Europe during the summer of 1783 CE.

Chapter 6

Clustering incomplete climatological time series

6.1 Background

Marked variations in regional climate patterns arise as a response to persistent changes of the climate system. Identifying these patterns is therefore fundamental for a better understanding of past climate changes at local and regional scales. Thus, with increasing computational power, the number of classification methodologies providing robust characterizations of regional climates has quickly escalated in the climate community, becoming a common tool for the study of past climatic patterns (Abatzoglou et al., 2009; Perdinan and Winkler, 2015; Horton et al., 2015). This chapter is therefore intended to validate the k-gaps algorithm, a novel technique to obtain robust clusterings from sample-starved datasets using all the information contained in their records. k-Gaps has already been described in Section 3.3 and the next sections will show some of its potential uses in the study of past climate events.

6.2 Ideal case with complete information

To obtain the "ideal" case scenario where complete data are available for regionalization, two k-means clusterings have been performed (Fig. 6.1) using the EObs gridded dataset of daily temperatures described in Section 2.1.2 . Thus, absolute temperatures have been used to provide the clusterization of areas with similar temperature means (k-gaps basic mode), whereas standardized temperatures have been employed for the clustering of records with correlated variability (the normalization mode).

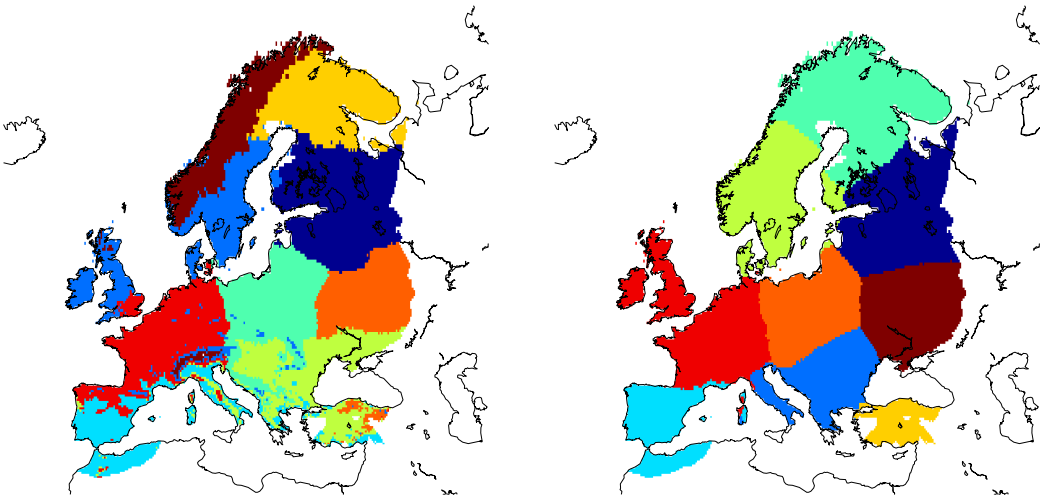

Figure 6.1: K-means clusterization of the E-Obs grid of daily summer temperatures since 1950 to 2016 (k-means$^{[EObs]}$) using absolute (a) and standardized (b) temperatures.

Note that, while the former (Fig. 6.1a) associates Northern regions with high altitude locations such as the Alps (Rubel et al., 2017) (as expected due to their similar lower mean temperatures), the latter (Fig. 6.1b) yields coherent patterns in terms of regional climate variability. Interestingly, the regionalization of Fig. 6.1a has certain similarities with the Köppen-Geiger classification (Köppen, 1884), especially in the aforementioned association of

high altitude temperatures with northernmost climates. However, it is not possible to directly compare against these Köppen-Geiger climate classifications because their classification system also takes into account precipitation patterns that are beyond the scope of this study. The regions obtained using k-means with the entire grid of European temperatures (k-means$^{[EObs]}$, hereafter) will be assigned as target to test the skill of k-gaps and other clustering techniques by comparison. The closer the regionalization to k-mean$^{[EObs]}$ regions, the higher the skill of the method.

6.3 Experiment setting

To assess the robustness of different clustering methods, synthetic datasets are generated from the complete E-Obs temperature grid using three main parameters: the number of synthetic records (i.e. datasets have different number of time series), their spatial distribution (i.e. temperatures are extracted from different locations for each dataset), and their time length (i.e. missing information is different for each dataset). These parameters can be interrelated since sampling networks and measuring campaigns have usually been conducted by organizations at regional scales. So most sampling collections are regionally related in time, leading to changes in the spatial coverage of the zone at different time periods. Synthetic data are therefore generated using Gaussian models (Fig. 6.2) that imitate the distribution and time intervals of real measuring campaigns. These models are spatially conformed by marked centers where there is a higher concentration of records, and a sparser coverage of their surroundings. On the other hand, different temporal lengths for synthetic time series are obtained by randomly altering the start and end times (Inset of Fig. 6.2) of complete temporal records from the E-Obs temperature dataset.

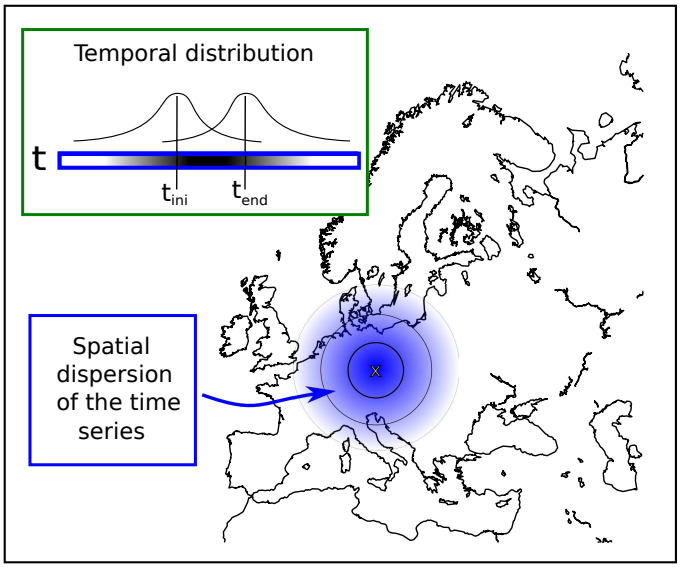

Figure 6.2: General representation of the spatial distribution of a Gaussian model (blue shade) employed to generate sample-starved datasets (with incomplete temperature series) where "×" demarcates the center. Darker blues indicate a higher concentration of time series, whereas lighter blues depict fewer time series. Inset: Time lengths of synthetic series generated with random variations of predefined initial (t_{ini}) and final (t_{end}) days.

Hence, 500 synthetic datasets of temperature records were generated to assess the performance of k-gaps, combining 20 Gaussian models with different centers, distributions, and time periods. Each one of them presents less than 1100 time series, containing from 55% to 90% of missing values within their chronologies (Fig. 6.3). These reduced number of records and data missing level are similar to real historical (Küttel et al., 2010) and paleoclimate (Emile-Geay et al., 2017) databases, and will serve as a test bed to validate the method for future studies.

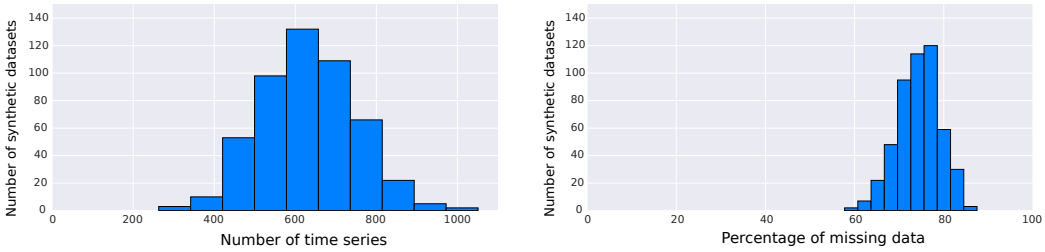

Figure 6.3: Distributions of the number of time series per dataset (left) and level of missing values related to the temporal length of the records (right) associated with 500 synthetic datasets.

For example, Fig. 6.4 illustrates one of these case studies. It is composed of 815 time series with a higher density of temperature records over Central Europe, while large territories of Southern Europe such as Italy, Spain and Portugal remain almost uncovered. Moreover, the number of available records is not constant and decreases back in time since 1980, restricting the information of past temperature changes. These features are different for each synthetic dataset and, therefore, provide a good framework to assess the robustness of clustering algorithms.

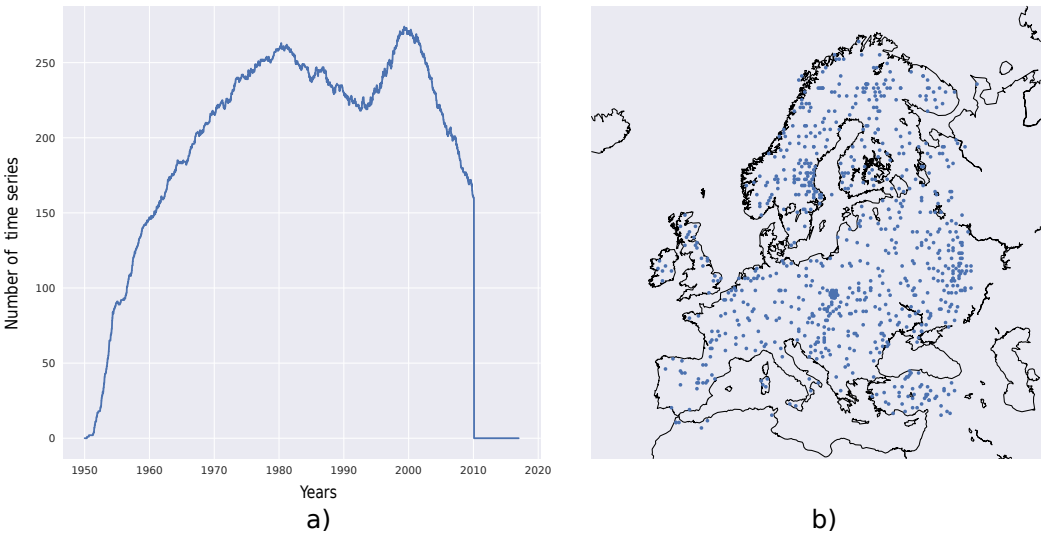

Figure 6.4: Temporal (a) and spatial (b) distribution of a sample study composed of 20 Gaussian models centered at different points in Europe.

6.4 Performance assessment

To assess the k-gaps performance, a statistical analysis has been carried out using the adjusted Rand Index (de Vargas and Bedregal, 2013), a metric that is used in data analysis to assess the similarity between clusterings. The index is constructed to detect whether two clusters obtained from different methods have the same associated time series, indicating how much those cluster look alike. In this sense, it can be seen as an accuracy measure, as well as a comparison test for clustering methods. In our case, the Rand Index was employed to compare the k-gaps results with the ideally perfect k-means$^{[EObs]}$ clustering. The index ranges from 0 to 1, where a value of 0 means that all data points correspond to completely different clusters and value of 1 would indicate that both clusterings are the same.

The performance of k-gaps has been compared with two other clustering techniques: the k-POD algorithm (Chi et al., 2016), and the k-means algorithm (from now onwards k-means$^{[rs]}$, where rs stands for "reduced set", indicating that it is only applied over a few time series, in contraposition with k-means$^{[EObs]}$ which uses the entire grid of temperatures). Note that, whereas k-gaps and k-POD clusterize incomplete time series (i.e. series with different temporal lengths that lead to gaps of missing information for some time intervals), k-means$^{[rs]}$ clusterized the same 500 datasets but with complete records (i.e. series covering the entire time period since 1950 CE). Thus, the comparison of k-gaps and k-POD with k-means$^{[rs]}$ provides information about the skill of these techniques to reproduce robust spatial patterns with fragmentary temporal data. Table 6.1 shows the adjusted Rand-Index means and standard deviations associated with these three methods for both modes. k-Gaps exhibited good results for most datasets with index values close enough to k-means$^{[rs]}$'s indices, outperforming the skill of k-POD to cluster uneven time series.

Table 6.1: Adjusted Rand-Index means of 500 synthetic case studies within 95% confidence interval for three clustering techniques.

Mode	k-POD	k-gaps	k-means[rs]
Basic	0.12±0.09	0.47±0.17	0.56±0.18
Normalized	0.13±0.11	0.54±0.24	0.61±0.19

k-Gaps clusterings obtained higher Rand-indices than k-POD counterparts for all synthetic datasets. Moreover, all clustering techniques yielded higher indices once the time series were normalized, indicating that these methods achieve better skill when records are clusterized in terms of their climate variability. Note that the capability to generate robust clusterings with these methodologies depends on two main factors: the temporal lengths of the time series, and their respective locations.

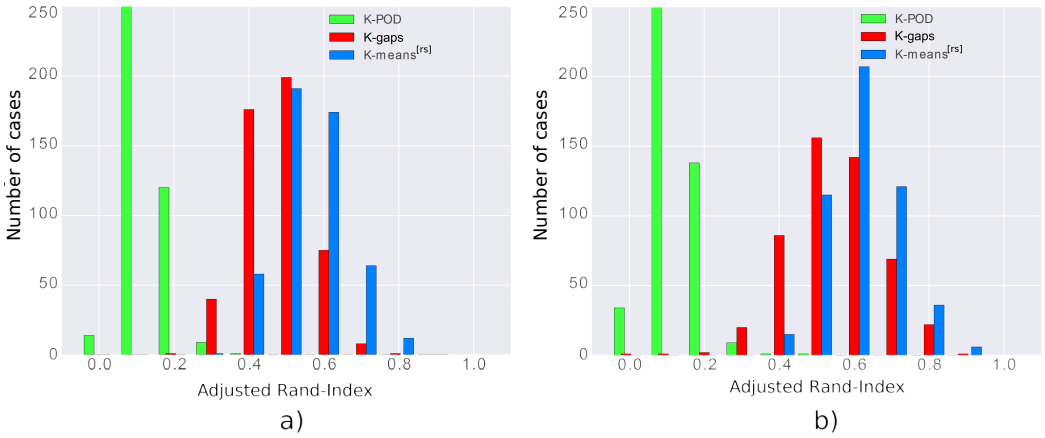

Figure 6.5: Comparison of clustering techniques using adjusted Rand-Index for absolute (a) and normalized (b) temperatures. The adjusted Rand-Index is calculated with respect to the clustering obtained by applying k-means to the complete E-Obs gridded dataset.

To see which factor is predominant, let us examine Fig. 6.5, where k-means[rs] index distributions (blue bars) illustrate the impact of sparse sampling locations on the skill to reproduce perfect clusterings (remember that k-means[rs] uses time series with the same length whereas k-gaps does not). In this regard, performance differences between k-gaps (red bars) and k-means[rs] indices can be explained due to the loss of temporal information (time series with gaps). As a matter of fact, k-gaps and k-means[rs] performances are quite similar (i.e. their Rand Index distributions obtained from 500 synthetic datasets overlap), indicating that clusterings are more sensitive to the distribution of sampling locations rather than to differences in the temporal length of records. The skill of the method has also been assessed as a function of the number of records used in the clustering process. Table 6.2 shows Rand indices for 3 different-sized datasets without any clear correspondence between size and performance (for instance, P404 is the biggest dataset with 875 time series, and its clustering has the worst performance in the normalized mode), suggesting that the efficiency of k-gaps does not strongly rely on the number of time series employed.

Table 6.2: Adjusted Rand-Index for k-gaps clusters with 3 synthetic datasets selected from the pool of 500 datasets described in Subsection 6.3. Note that each dataset is composed of a different number of time series (N), and each time series has lost, on average, 80% of the climate information for the period 1950-2016 CE.

Dataset	**N**	**Missing data (%)**	**Basic**	**Normalized**
P191	815	87.9	0.52	0.69
P280	623	87.9	0.38	0.69
P404	875	83.6	0.51	0.38

Note that in these examples, the average amount of missing data per series is above 80%, and the loss of information (i.e. the distribution of missing data) is different for each dataset as shown in Fig. 6.6.

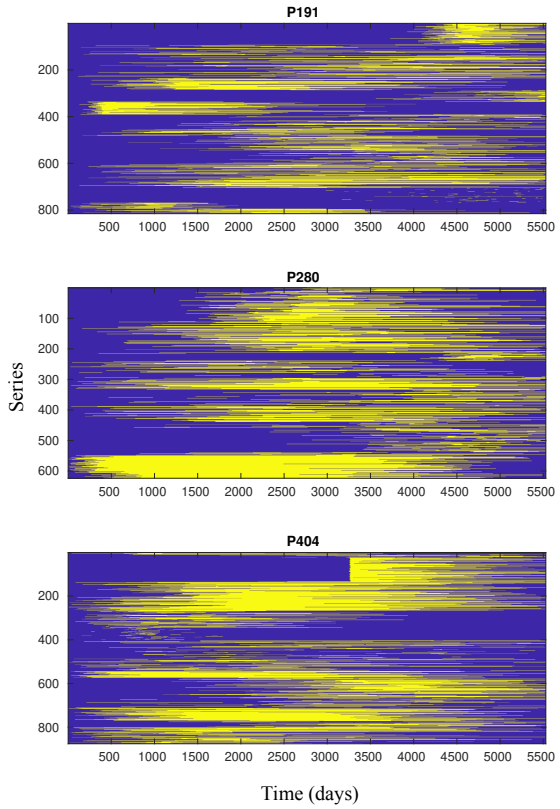

Figure 6.6: Distribution of missing data for each one of the series included in datasets P191, P280, and P404. Days without temperature values are depicted in blue, and available daily temperatures are shown in yellow.

This implies that the algorithm is able to cluster time series that are almost empty, as long as the entire period of study is covered by the sum of the different series included in the dataset. It is also shown that different performances can be obtained for the same dataset depending on the clustering mode (e.g. dataset P280 in Table 6.2 shows a lower Rand index in Basic mode than in Normalized mode), which indicates that a robust clustering with absolute temperature series requires different locations than clusterings with normalized series. This is consistent with the fact that regions with similar mean temperatures (i.e. those selected in Basic mode) are not the same as regions with similar variability (i.e. obtained in Normalized mode), as shown in Fig. 6.1.

To visualize the spatial patterns obtained in a regionalization with the k-gaps algorithm, we have clusterized dataset 191 with an intermediate Rand Index value (its clustering performance in terms of Rand Index is representative of what can be expected from the clustering of the 500 synthetic datasets). Fig. 6.7 displays the resulting clusterings of these series of temperature together with their locations. Note that to facilitate the comparison with Fig. 6.1, once P191 has been clusterized, we have interpolated the clustering in the remaining grid points where temperatures are not available by using the k-nearest neighbours algorithm (Cover and Hart, 1967). The adjusted Rand indices for basic and normalization modes are 0.52 and 0.69, respectively. These values indicate that kgaps has similar performance than k-means (which needs complete temporal information) using sample-starved climate datasets with different temporal lengths.

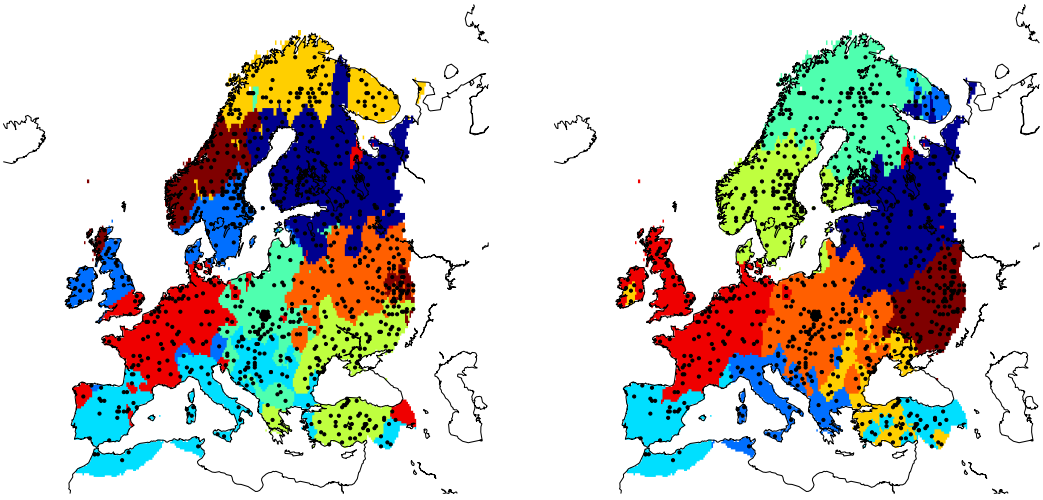

Figure 6.7: K-gaps clusterization of a study case with 815 time series (black dots), for Basic (a) and Normalization (b) modes.

On the other hand, lower Rand indices are related to irregular spatial distributions, where small regions are well characterized by a disproportionated number of climate records, while large extensions of land remain unsampled. Such is the case of dataset P404, chosen as a clustering with low Rand Index (0.38 for normalization mode as seen in Table 6.2), and whose regionalization using the normalization mode is depicted in Fig. 6.8. In this case, the difference between the number of time series in Central Europe and Southern Europe has altered the formation of clusters by associating regions in the Iberian Peninsula with the big aggregation of temperature records in France as seen in Fig. 6.8 (black dots depict the spatial distribution of dataset P404). At the same time, the concentration of points in Ireland produces a new cluster for the British Isles which is not present in k-means[EObs]. This indicates that special attention should be paid to disparities in the coverage of the territory because they play an important role in the final structure of the clustering, and an uneven distribution of time series can debase the analysis of regional climates.

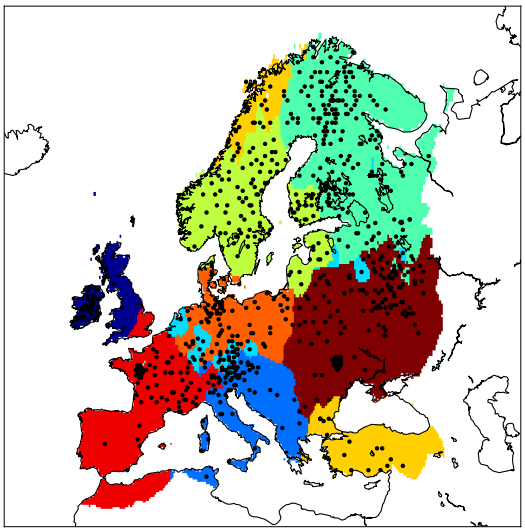

Figure 6.8: K-gaps clusterization for P404 in Table 6.2 using the normalization mode. The dataset is composed of 875 time series (black dots) unevenly distributed over Europe.

However, while non-homogeneous distributions force the splitting and merging of some clusters, some of the spatial patterns from Fig. 6.1 can be identified, proving that realistic information about temperature variability can still be extracted from low quality datasets such as P404.

6.5 Applications on climate studies

The full potential of k-gaps is evidenced in Fig. 6.7 where the resulting spatial patterns for both modes exhibit important similarities with those obtained for k-means$^{[EObs]}$ (Fig. 6.1). This is quite remarkable, because k-gaps is applied over incomplete datasets (i.e. time series with gaps unevenly distributed over Europe) whose size in terms of number of time series is 21 times smaller than the homogeneous grid of temperatures which has complete temperature information for the entire time period (no gaps) in all grid points (complete spatial coverage). It is also noteworthy to mention that although synthetic records are sparsely distributed across Europe, clusters are defined for almost the same regions as for the gridded dataset, confirming the robustness of the spatial clustering. For instance, spatial patterns over Iberia and the United Kingdom are well reconstructed with the method even when there is a lack of sampling records (black dots in Fig. 6.7 represent the locations with temperature series) in the southern part of their territory. This suggests that only a few number of locations are necessary to reproduce most of the climatology within these areas.

Furthermore, k-gaps cluster analyses provide a useful framework for the study of past climate trends and the detection of extreme events at regional scales. In the first case, centroids obtained from k-gaps are time series ranging the entire study period whose variations represent changes in the temperature of climatically different regions. Therefore, the linear regression of these centroids can be utilized to estimate regional temperature trends for periods beyond the scope of complete datasets that have been truncated using homogeneization procedures. For instance, the distribution of temperature trends

obtained from clusterings of 500 synthetic datasets is shown Fig. 6.9. Trend differences between k-means[EObs] and k-gaps are below 0.03 °C/year for 75% of our synthetic datasets, indicating that although there could be small biases induced by the irregular distribution of uneven (and scarce) time series, the temperature trends obtained from k-gaps centroids are similar to those obtained from the clustering of complete gridded datasets (i.e. k-gaps centroids obtained with fragmentary information have similar trends as centroids generated with k-means[EObs]). Therefore, these k-gaps centroids can be used to study changes in past temperature trends from incomplete datasets.

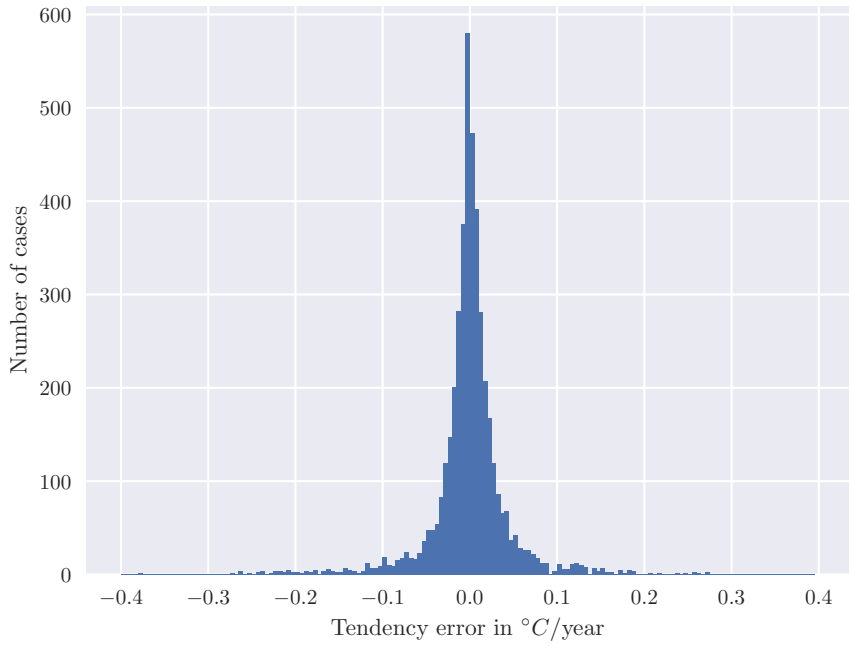

Figure 6.9: Histogram of trend errors estimated from differences between k-gaps centroids and ideal centroids retrieved from k-means[EObs]. Temperature trends have been calculated for 500 synthetic datasets.

On the other hand, as the normalization mode of k-gaps associates time series with similar climate variability, it is possible to detect extreme events in regions where those time series are located. Take for instance series of temperature observations, applying the k-gaps algorithm in normalization mode will allow us to discern regions (delimited by the k-gaps clusters) where temperature anomalies were significantly higher (or lower) with respect to the climatology. This methodology can therefore be used to study when and where extreme temperature events such as heatwaves (or cold spells) took place.

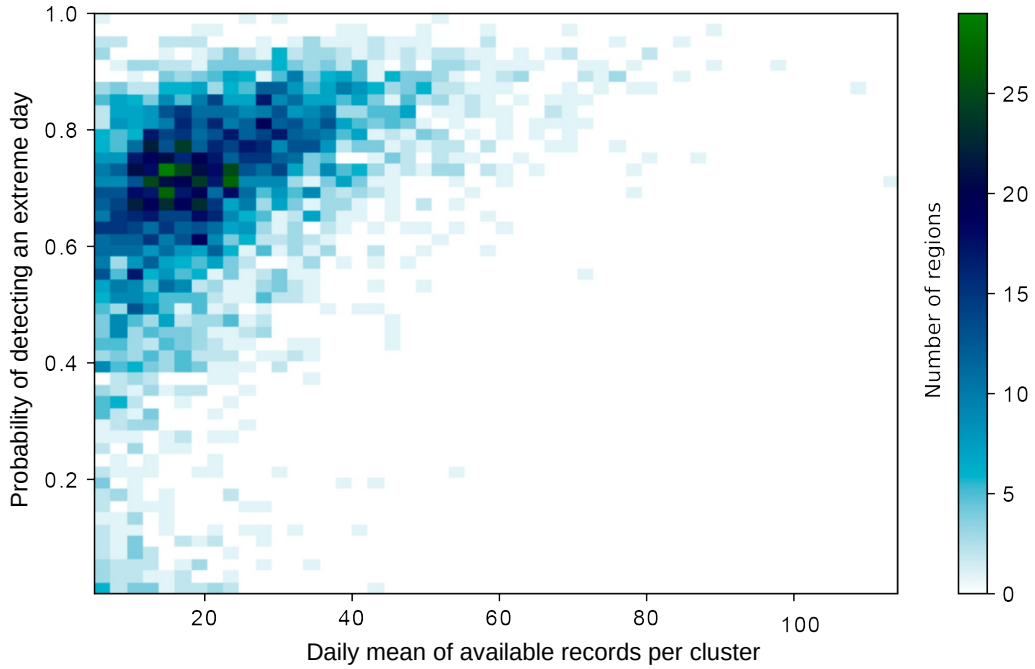

Figure 6.10: Probability of detecting a heatwave within clusters obtained using k-gaps (in normalization mode for 500 synthetic datasets) as a function of the number of time series associated with each cluster. The colorbar represents the number of clusters (regions).

Following Demuzere et al. (2011), we identified a heatwave in a cluster when at least 95% of the time series associated with that cluster reached values above the 95^{th} percentile of temperatures on record. We can therefore assess the probability of detecting heatwaves using k-gaps clusterings by comparing extreme temperature events detected with the 500 synthetic datasets (i.e. datasets with incomplete series and fragmentary information) against extreme events ideally detected with k-means using the complete E-Obs grid of temperatures (i.e. a dataset with complete spatio-temporal information). Note that the probability of detecting extreme temperatures is higher for clusters composed of an elevated number of time series, however Fig. 6.10 shows that the probability of detecting a heatwave using k-gaps clusterings is high enough even when the number of temperature series per day is reduced down to a few tens. For example, clusters with only 20 time series have, on average, between 60% and 80% chance of detecting temperature extremes (remember that we obtain these values by comparison with the extreme events detected using k-means$^{[\text{EObs}]}$ clusters). This probability increases above 90% for clusters with at least 30 time series, indicating that the detection of extremes events within regions with similar climate variability defined by the k-gaps algorithm is possible with sets of just a few time series per cluster.

Keynotes

These are the most important findings of Chapter 6:

- k-Gaps is a new clustering technique for climatological regionalizations.
- It does not require complete timeseries.
- It is able to use the full length of records without the need of filling the gaps.
- It reproduces similar patterns as clusterings obtained with complete time-series.
- Regional trends can be estimated from centroids of k-gaps that span for longer time periods than centroids calculated using classical clustering methods.
- Extreme events can be regionally detected from the clustering of normalized time-series.

Chapter 7

Conclusions and outlook

The use of AI is growing in climate sciences, but applications have mainly focused on big data, in order to synthesize and extract information from massive datasets where data availability is not an issue. Here we have shown that AI procedures such as evolutionary algorithms and clustering techniques can also be useful in equally important Earth science areas, where the opposite problem of data scarcity is a major limiting factor (e.g. paleoclimatology). Perhaps paradoxically, limited data availability also requires solving high dimensional and non-linear problems, as there is a common optimization goal of extracting the maximum useful information. By focusing on specific climate problems (spatially resolved climate field reconstruction or climate regionalization), we show the potential of these new methodologies to optimize the information of incomplete datasets, for a diversity of observing networks in terms of observables (from perfect temperature pseudo-proxies to instrumental observations of dynamical fields), temporal resolution (from annual to daily) and targets (from global to regional). Interestingly, for the cases analyzed here, we came to the same conclusion that more data is not always better. Of course, this might not apply to other networks, arguably those with very few observables or large errors in other sources of uncertainty. However, our results suggest that even in those cases, some level of improvement can be achieved through optimized rearrangements. If the skill gained from the optimization is or not substantial will depend on the characteristics

of the network and the pursued objective, which calls for individual assessments that eventually may require tailored developments of AI procedures to improve their efficiency, as illustrated through the chapters of this thesis.

In Chapter 4, we coupled different reconstruction methods to an evolutionary algorithm in order to reconstruct global temperature fields simulated by model ensembles of the last millennium from a pseudo-proxy network with the same number and distribution of records as in the real world. In doing so, **we have quantified a measurable spatial bias in global temperature reconstructions due to the non-uniform distribution of currently available paleoclimate records** (Emile-Geay et al., 2017). In our idealized experiments, **the field can be reconstructed from an optimized selection of pseudo-proxies from the full network without sacrificing the skill of the reconstruction**. For all experiments performed, the skill of the full-proxy reconstruction can be improved by using subsets of pseudo-proxies strategically situated over representative locations (Jaume-Santero et al., 2020).

The set of optimal locations highlights the importance of polar regions and major teleconnections areas to reconstruct annual global temperature patterns. The optimized network also captures the global responses to major external forcings and modes of internal variability. However, while **annual temperature fields are well described by high-latitude records, low frequency fluctuations such as the MCA-LIA transition are better represented by pseudo-proxies situated at lower latitudes**. As the optimal distribution of records varies with the spectral frequency of the target field, this advises against pooling together proxies that resolve different time scales and encourages further research for weighting records depending on their location and response resolution. Our results are robust to the reconstruction methodology and the ensemble member or model employed, and they hold for more realistic pseudo-proxy experiments and datasets. Still, important assumptions have been made and all uncertainty sources have not been considered, which calls for caution

when extrapolating these results to real reconstructions. This was a major motivation for Chapter 5, where complexity was added to the experiments by using a network of historical instrumental observations with changes in data availability.

Further experiments are encouraged in order to account for additional sources of uncertainty not included in Chapters 4 or 5, such as the non-univariate dependency of proxy records. Considering all sources of uncertainty would require dedicated but in principle feasible implementations in the CRO algorithm, potentially bringing changes in the structure of the optimized networks reported in Chapter 4. For example, more skillful reconstructions could be attempted through optimization functions tackling with the climate signal and response resolution embedded in different paleoclimate archives. Our approach also paves the way for determining valuable regions to carry out future measuring campaigns of proxy records, which can be elucidated by mapping areas with natural paleoclimate archives that are under-represented in current proxy networks. Note that, the CRO algorithm can be very useful to minimize the costs of expensive measuring campaigns because of its suitability to find the minimum number of representative locations to set new meteorological stations and/or to sample proxy records. The algorithm can also be modified to know where it should not be measured, avoiding regions that might present the right observational conditions but are not representative for the aimed task.

In Chapter 5 we made the leap to the real world by using station-based observations of monthly SLP over the North Atlantic from 1750 to 2004 CE. We found that evolutionary algorithms can also outperform non-optimized reconstructions of dynamical fields on regional and monthly scales (Luterbacher et al., 2002; Küttel et al., 2010). Instead of selecting an optimal subset, in this case we tunned the CRO to derive optimal sets of weights that maximize the reconstruction skill of the observing network. The optimization process exploits the information of the full dataset, taking into account changes in data availability, as well as inconsistencies within the network and with the

large-scale field caused by other uncertainties. The relationships learnt by the CRO algorithm are transformed in local weights during the reconstruction of the large-scale SLP field. **The optimized reconstruction improves the performance of reconstructions generated without optimal weighting over almost the entire region** (especially around the North Atlantic Ocean). Additional reconstruction experiments using reanalysis and model data with the same constraints in data availability show that the spatial distribution of weights and the pattern of improvement are robust to the reference dataset and internal variations of the large-scale field targeted by the reconstruction. This is in agreement with the pseudo-reconstructions of Chapter 4, that also demonstrated that the CRO brings similar patterns of improvement for different reconstruction techniques.

According to our results, changes in spatio-temporal data availability substantially affect the representativeness of local observations in the network, therefore arising as an important source of uncertainty in our SLP reconstructions. This result justified a separate optimization of the observing network for the earlier reconstruction period (1750-1835 CE), which displayed marked changes in the number and coverage of observations as compared to the remaining period. Despite major constraints in data availability, the optimized weights for the earlier period bring a pattern of improvement that resembles the one achieved by the denser network of the latter period. **The generalized improvement with respect to the non-optimized reconstruction is reached at the expense of sacrificing some skill in the better-sampled region of Europe. However, this is overcompensated by comparatively larger improvements over the North Atlantic Ocean.** The loss of skill over continental areas can be admissible for our reconstruction as well as other reconstructions of dynamical fields concerned with internal and forced aspects of the large-scale atmospheric circulation, which typically involve changes in major oceanic action centers. Additional improvements could be attempted through optimal subsets tailored to the temporal availability of the observing network (i.e. by deriving weights that have been optimized for each configuration of the observing net-

work during the reconstruction period).

The CRO-SLP reconstruction provides high-resolution monthly SLP fields over the North Atlantic for 1750-2004 CE, from which we derived the longest records of seasonal indices for the main modes of climate variability of the North Atlantic, such as the NAO or the EA pattern, and its main action centers, the AH and the IL (albeit with large uncertainties before the 1830s). In particular, we have focused on the AH, because it covers the region with the largest improvement in the CRO-SLP reconstruction, and different to the NAO, there have been few attempts to reconstruct the AH before the 20$^{\text{th}}$ century, likely due to the scarcity of observations over the southern half of the North Atlantic Ocean. Despite the lack of long-term trends in the summer AH, it can experience substantial interannual changes in location/extension that match with anomalous European conditions inferred from independent temperature reconstructions and might serve as historical analogues of upcoming summers under the projected northern shift and expansion of the summer AH (Cherchi et al., 2018).

Our reconstructions reveal an intensification of the winter AH during the second half of the 20$^{\text{th}}$ century that had no precedents in at least the last 250 years. The strengthening of the winter AH is now declining and is timely with the well-reported positive NAO trend of 1960s-1990s. While different causes have been proposed for this NAO trend, including anthropogenic climate change, several studies show that it is not exceptional as compared to pre-industrial periods before 1650 CE or statistically distinguishable from atmosphere-ocean internal variability (Pinto and Raible, 2012) and references therein). This may also be the case of the trend in the winter AH, whose changes in intensity have been associated with North Atlantic sea surface temperature anomalies (Falarz, 2019). Our SLP reconstruction provides a longer benchmark of pre-industrial conditions to characterize the low-frequency variability of the AH and explore the causes of this recent anomalous behaviour.

We also found that **the spread among NAO indices is larger for winters when the NAO was dominated by the AH.** Although part of these differences may arise from the different definitions of the southern node of the NAO, they seem to be present across indices that follow the same methodology. As such, according to our results, **current discrepancies in instrumental NAO indices would stem more from uncertainties in the AH than in the IL.** This points to limitations of current datasets to capture accurately the historical evolution of the AH, and stresses the need for improved SLP reconstructions over this region. Here is where the CRO algorithm can help to find the right locations to set land-based stations of pressure observations.

On the other hand, in Chapter 6 **we have presented a novel clustering technique for incomplete datasets known as k-gaps** (Carro-Calvo et al., 2020). This algorithm is an iterative technique that allows for the clustering of heterogeneous datasets using most of the information contained in incomplete time series (i.e. with at least 55% of the information unavailable). The method is fully based on the structure of the well-known k-means algorithm, but including different procedures in order to adapt the algorithm to series with gaps and/or different temporal lengths. Thus, **the k-gaps algorithm allows to cluster climate fields whose sampling records are only available for certain time periods** (e.g. as it is typically the case of historical datasets of basic meteorological variables such as temperature, precipitation, wind or SLP). The results show that this classification algorithm performs well with datasets containing high levels of temporal missing values. In the case of daily European temperatures, k-gaps exhibited a good performance with most of the 500 synthetic datasets (with an average of around 80% of missing data per record) employed for its validation, yielding mean and variability patterns similar to those obtained when traditional methods (including k-means) are applied to the originally unbiased records (i.e. complete series). Moreover, consistent climatic clusterings have been achieved for a reduced availability of climate records even when the number of time series were 21 times smaller than the original grid of temperatures.

Therefore, the k-gaps algorithm is well-suited for the analysis of regional climates from datasets with an uneven distribution of records in both space and time (e.g. station-based observations with missing values). On the other hand, additional experiments with synthetic datasets showed no clear correspondence between the number of time series employed in the analysis and the skill of the clustering, **indicating that an adequate spatial distribution of sampling records over the region of interest can be more important than the size of the dataset.** Therefore, and similarly to the results obtained in Chapters 4 and 5 with the CRO, we came to the conclusion that **a dense network is not necessarily better than a subset of well distributed records.**

Furthermore, it has been shown that **k-gaps is well suited for the reconstruction of regional climate trends, and the detection of extreme events at regional scales**, as obtained from time series contained in the clusters. If the clustering is performed in basic mode, the resulting centroids can provide an estimation of the temperature trends of each cluster, whereas if it is done in normalization mode, time series with correlated variability are grouped together, allowing for the detection of extreme events at regional scale. Future applications of the k-gaps algorithm could be applied to other climatically differentiated regions apart from Europe. Other experiments can be carried out to test the robustness of the method with several climate variables such as precipitation and SLP, and other temporal resolutions (e.g. monthly seasonal, annual, etc). Moreover, although further research should be carried out to test its performance under records with significant amounts of noise, the k-gaps algorithm also paves the way for future applications on regional analyses of past climate changes based on historical and paleoclimate archives. Finally, it is also possible to combine the methodologies developed in Chapters 4 and 5 (CRO algorithm) with k-gaps. For example, there are cases (e.g. climate extremes) for which increasing trends may artificially result from increasing data availability. The method would help to detect more robust trends in regional extremes, which would be useful for detection and attribution exercises. A shortcoming that could

be solved with evolutionary algorithms is that k-gaps does not deal with e.g. temporal changes in data quality, which may also cause spurious trends. Further improvements could be achieved by a CRO-based optimized selection of time series that maximize the relationship with the regional field before the application of the k-gaps.

Appendix A

Supplementary Figures

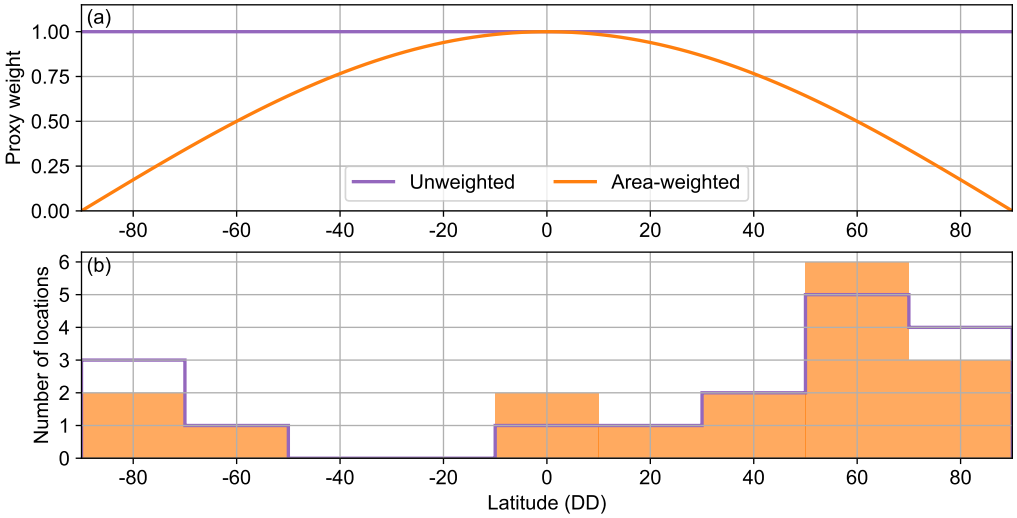

Figure A.1: Sensitivity of the optimization process to area-weighting. (a) Weights assigned to perfect pseudo-proxies as function of their latitude in two experiments of the CRO-AM. (b) Latitudinal distribution of optimized subsets of 17 perfect pseudo-proxies from the PAGES-2k network obtained with area-weighted (orange shading) and unweighted (purple line, CRO-MIN) versions of CRO-AM.

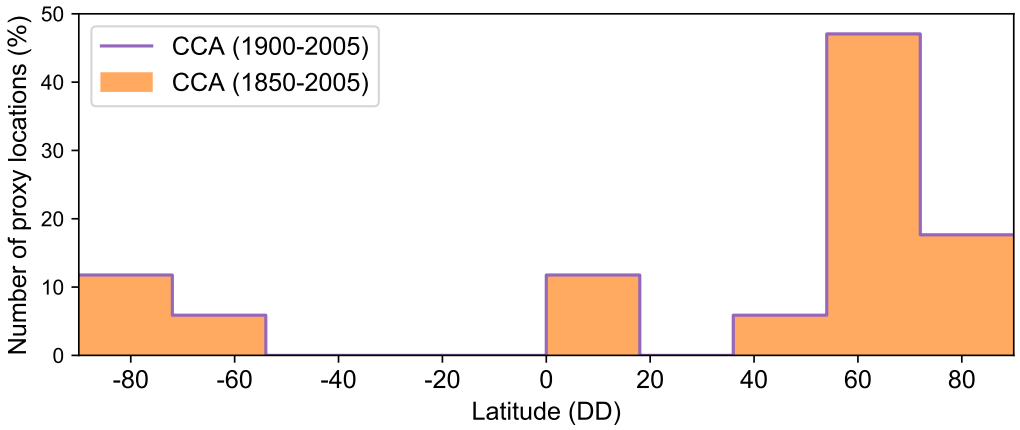

Figure A.2: Latitudinal distributions of 17 representative locations obtained with the CRO-CCA using years from 1850 to 2005 (orange shade) and from 1900 to 2005 as calibration periods.

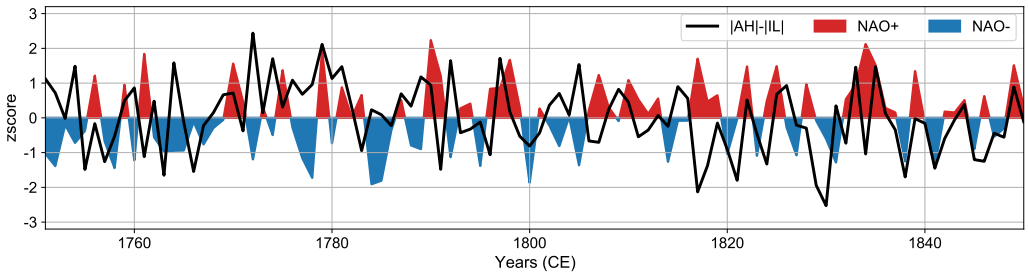

Figure A.3: CRO-SLP NAO(red and blue shade) and |AH|-|IL| (black line) indices from 1751 to 1850 CE. Both series were standardized with respect to the 1751-1850 CE baseline.

Appendix B

Supplementary Tables

Table B.1: List of accepted observations from SLP-Obs database.

Name	**Latitude(DD)**	**Longitude (DD)**	**Start (CE)**	**End (CE)**
Boston	42.36	-71.03	1840	2004
Toronto	43.70	-79.42	1841	1980
Bermuda	32.32	-64.77	1836	2004
Nashville	36.16	-86.78	1854	1980
Cincinatti	39.90	-84.27	1850	2004
Chicago	41.87	-87.62	1853	1980
St Louis	38.63	-90.20	1854	1980
Havana	23.12	-82.38	1858	2002
Jacksonville	30.33	-81.65	1871	1980
Key West	24.55	-81.75	1850	2004
Charleston	32.78	-79.93	1850	2004
New Orleans	29.59	-90.15	1850	2004
New York	40.73	-73.93	1850	1980
Galveston	29.30	-94.79	1873	1980
Father Point	48.52	-68.47	1874	1980
Montreal	45.51	-73.59	1874	1980
Minneapolis	44.99	-93.26	1865	1980

Continued on next page

Table B.1 – *Continued from previous page*

Name	**Latitude(DD)**	**Longitude (DD)**	**Start (CE)**	**End (CE)**
Jacobshavn	69.22	-51.10	1866	1980
Ivigtut	61.21	-48.17	1866	1980
Aberdeen	57.20	-2.22	1855	2004
Akureyri	65.68	-18.08	1873	2004
Archangelsk	64.60	40.50	1850	2004
Armagh	54.40	-6.70	1849	2002
Astrakhan	46.60	48.00	1850	2004
Athens	38.00	23.70	1857	2004
Barcelona	41.20	2.10	1779	2004
Bergen	60.40	5.30	1815	2002
Berlin	52.40	13.10	1850	2004
Bidston	53.40	-3.00	1845	2002
Boothville	29.19	-89.23	1873	2002
Budapest	47.50	19.00	1809	2004
Cadiz	36.50	-6.30	1786	2004
Cairo	30.10	31.40	1856	2003
Copenhague	55.70	12.60	1841	2004
De Bilt	52.10	5.20	1848	2004
Des Moines	41.60	-93.60	1878	2002
Dublin	53.40	-6.30	1830	2004
Edinburgh	56.00	-3.40	1769	2004
Florence	43.80	11.30	1813	2004
Funchal	32.60	-16.90	1849	2004
Gdansk	54.30	18.90	1806	2004
Genoa	44.24	8.55	1832	2004
Gibraltar	36.20	-5.40	1821	2004
Nuuk	64.17	-51.75	1873	2004
Goteborg	57.70	11.99	1859	2004
Haparanda	65.80	24.20	1860	2004
Harnosand	62.38	17.56	1859	2004

Continued on next page

Table B.1 – *Continued from previous page*

Name	**Latitude(DD)**	**Longitude (DD)**	**Start (CE)**	**End (CE)**
Helsinki	60.30	25.00	1850	2004
Istanbul	41.00	29.10	1855	2004
Kiev	50.40	30.50	1849	2004
La Coruna	43.21	-8.24	1866	2004
Lisbon	38.70	-9.20	1849	2004
Lockbourne	39.81	-82.97	1878	2002
London	51.20	-1.00	1773	2004
Lund	55.40	13.10	1779	2004
Luxembourg	49.36	6.70	1837	2004
Lvov	49.80	24.00	1850	2004
Madrid	40.40	-3.70	1785	2004
Malta	35.90	14.50	1851	2004
Marseille	43.50	5.20	1850	2004
Milan	45.50	9.10	1764	2004
Moscow	55.80	37.60	1837	2004
Nantes	47.20	-1.60	1850	2004
Nicosia	35.10	33.20	1866	2003
Nordby	55.52	8.57	1873	2002
Odessa	46.50	30.60	1841	2004
Oporto	41.10	-8.60	1862	2004
Oslo	60.00	10.70	1815	2004
Padua	45.40	11.80	1749	2004
Palma	39.60	2.70	1849	2004
Paris	49.00	2.50	1764	2004
Ponta Delgada	37.80	-25.70	1864	2002
Prag	50.10	14.30	1788	2004
Reykjavik	64.10	-21.80	1820	2004
Rome	41.80	12.20	1849	2003
Sibiu	45.80	24.20	1850	2004
Split	43.30	16.26	1849	2003

Continued on next page

Table B.1 – *Continued from previous page*

Name	**Latitude(DD)**	**Longitude (DD)**	**Start (CE)**	**End (CE)**
Stockholm	59.40	18.10	1849	2004
St Petersburg	60.00	30.30	1821	2004
Stykkisholmur	65.08	-22.73	1849	2003
Sulina	45.15	29.67	1849	2004
Tbilisi	41.43	44.47	1843	2002
Torshavn	62.02	-6.77	1866	2004
Triestre	45.70	13.80	1840	2004
Tromso	69.40	18.56	1873	2004
Trondheim	63.50	10.90	1767	2004
Upernavik	72.47	-56.10	1873	2004
Uppsala	59.90	17.60	1749	2004
Valentia	51.93	-10.25	1865	2004
Vardo	70.40	31.10	1860	2004
Visby	57.38	18.17	1859	2002
Warsaw	52.20	21.00	1835	2004
Wroctaw	51.10	16.88	1849	2004
Zagreb	45.80	16.00	1861	2003
Azores	38.31	-28.38	1850	2004
Beirut	33.54	35.30	1850	2003
Tunis	36.83	10.22	1875	2004
Sable Isla	43.93	-60.02	1850	2004
Moosonee	51.27	-80.65	1850	2004
Merida	20.59	-89.39	1850	2004
Nassau	25.05	-77.47	1850	2004

Table B.2: Pearson correlations of the AH (r_{AH}) and IL (r_{IL}) with the NAO spread calculated as the standard deviation of NAO indices in Table 5.1 (a PC-based NAO Index from 20CRv3 is also included). Different NAO spreads have been calculated from 1900 to 2004 CE by excluding the series in the dropped column.

Dropped	$\mathbf{r}_{AH}$	$\mathbf{r}_{IL}$
None	0.36	-0.07
Küttel et al. (2010)	0.36	-0.03
Luterbacher et al. (2001)	0.30	-0.09
Hurrell and Deser (2010)	0.36	-0.07
20CRv3	0.21	-0.16
Jones et al. (1997)	0.42	-0.03

Bibliography

Abatzoglou, J. T., Redmond, K. T. and Edwards, L. M. Classification of Regional Climate Variability in the State of California. *J. Appl. Meteorol. Climatol.*, vol. 48(8), pages 1527–1541, doi:10.1175/2009JAMC2062.1, 2009.

Abdel-Basset, M., Abdel-Fatah, L. and Sangaiah, A. K. *Chapter 10 - Metaheuristic Algorithms: A Comprehensive Review*, pages 185–231. Intelligent Data-Centric Systems. Academic Press, 2018. ISBN 978-0-12-813314-9.

Aburas, M. M., Ahamad, M. S. S. and Omar, N. Q. Spatio-temporal simulation and prediction of land-use change using conventional and machine learning models: a review. *Environ. Monit. Assess.*, vol. 191(4), page 205, ISSN 1573-2959, doi:10.1007/s10661-019-7330-6, 2019.

Agapiou, A. Remote sensing heritage in a petabyte-scale: satellite data and heritage Earth Engine© applications. *Int. J. Digit. Earth*, vol. 10(1), pages 85–102, ISSN 1753-8947, doi:10.1080/17538947.2016.1250829, 2017.

Aliaga, V. S., Ferrelli, F. and Piccolo, M. C. Regionalization of climate over the Argentine Pampas. *Int. J. Climatol.*, vol. 37, pages 1237–1247, ISSN 1097-0088, doi:10.1002/joc.5079, 2017.

Allan, R. and Ansell, T. A New Globally Complete Monthly Historical Gridded Mean Sea Level Pressure Dataset (HadSLP2): 1850-2004. *J. Clim.*, vol. 19(22), pages 5816–5842, ISSN 0894-8755, doi:10.1175/JCLI3937.1, 2006.

BADOR, M., NAVEAU, P., GILLELAND, E., CASTELLÀ, M. and ARIVELO, T. Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe. *Weather Clim. Extremes*, vol. 9, pages 17 – 24, ISSN 2212-0947, doi:10.1016/j.wace.2015.05.003, 2015.

BALLINGS, M., VAN DEN POEL, D. and M., B. Social media optimization: Identifying an optimal strategy for increasing network size on Facebook. *Omega*, vol. 59, pages 15 – 25, ISSN 0305-0483, doi:10.1016/j.omega.2015.04.017, 2016.

BARNES, E. A., HURRELL, J. W., EBERT-UPHOFF, I., ANDERSON, C. and ANDERSON, D. Viewing Forced Climate Patterns Through an AI Lens. *Geophys. Res. Lett.*, vol. 46(22), pages 13389–13398, ISSN 0094-8276, doi:10.1029/2019GL084944, 2019.

BARNSTON, A. G. and LIVEZEY, R. E. Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns. *Mon. Weather Rev.*, vol. 115(6), pages 1083–1126, doi:10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2, 1987.

BARRIOPEDRO, D., FISCHER, E. M., LUTERBACHER, J., TRIGO, R. M. and GARCÍA-HERRERA, R. The hot summer of 2010: Redrawing the temperature record map of europe. *Science*, vol. 332(6026), page 220, doi:10.1126/science.1201224, 2011.

BENESTAD, R. E., ERLANDSEN, H. B., MEZGHANI, A. and PARDING, K. M. Geographical Distribution of Thermometers Gives the Appearance of Lower Historical Global Warming. *Geophys. Res. Lett.*, vol. 46(13), pages 7654–7662, ISSN 0094-8276, doi:10.1029/2019GL083474, 2019.

BERNARD, E., NAVEAU, P., VRAC, M. and MESTRE, O. Clustering of Maxima: Spatial Dependencies among Heavy Rainfall in France. *J. Clim.*, vol. 26(20), pages 7929–7937, doi:10.1175/JCLI-D-12-00836.1, 2013.

BHEND, J., FRANKE, J., FOLINI, D., WILD, M. and BRÖNNIMANN, S. An ensemble-based approach to climate reconstructions. *Clim. Past*, vol. 8, doi:10.5194/cp-8-963-2012, 2012.

BIESBROEK, R., BADLOE, S. and ATHANASIADIS, I. N. Machine Learning for research on climate change adaptation policy integration: an exploratory UK case study. *Reg. Environ. Change*, vol. 20(3), page 85, ISSN 1436-378X, doi:10.1007/s10113-020-01677-8, 2020.

BOERS, N., GOSWAMI, B., RHEINWALT, A., BOOKHAGEN, B., HOSKINS, B. and KURTHS, J. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, vol. 566(7744), pages 373–377, ISSN 1476-4687, doi:10.1038/s41586-018-0872-x, 2019.

BOTHE, O. and ZORITA, E. Proxy surrogate reconstructions for Europe and the estimation of their uncertainties. *Clim. Past*, vol. 16(1), pages 341–369, doi:10.5194/cp-16-341-2020, 2020.

BOUNDS, D. G. New optimization methods from physics and biology. *Nature*, vol. 329(6136), pages 215–219, ISSN 1476-4687, doi:10.1038/329215a0, 1987.

BRADLEY, R. S. Are there optimum sites for global paleotemperature reconstruction? In *(eds. Jones P.D., Bradley R.S., Jouzel J.) Climatic variations and forcing mechanisms of the last 2000 years. NATO ASI Series (Series I: Global Environmental Change)*, pages 603–624. 1996.

BRADLEY, R. S. and JONES, P. D. 'Little Ice Age' summer temperature variations: their nature and relevance to recent global warming trends. *Holocene*, vol. 3, doi:10.1177/095968369300300409, 1993.

BRÖNNIMANN, S., ALLAN, R., ASHCROFT, L., BAER, S., BARRIENDOS, M., BRÁZDIL, R., BRUGNARA, Y., BRUNET, M., BRUNETTI, M., CHIMANI, B., CORNES, R., DOMÍNGUEZ-CASTRO, F., FILIPIAK, J., FOUNDA, D., HERRERA, R. G., GERGIS, J., GRAB, S., HANNAK, L., HUHTAMAA, H., JACOBSEN, K. S., JONES, P., JOURDAIN, S.,

KISS, A., LIN, K. E., LORREY, A., LUNDSTAD, E., LUTERBACHER, J., MAUELSHAGEN, F., MAUGERI, M., MAUGHAN, N., MOBERG, A., NEUKOM, R., NICHOLSON, S., NOONE, S., NORDLI, Ø., ÓLAFSDÓTTIR, K. B., PEARCE, P. R., PFISTER, L., PRIBYL, K., PRZYBYLAK, R., PUDMENZKY, C., RASOL, D., REICHENBACH, D., REZNÍCKOVÁ, L., RODRIGO, F. S., ROHR, C., SKRYNYK, O., SLONOSKY, V., THORNE, P., VALENTE, M. A., VAQUERO, J. M., WESTCOTTT, N. E., WILLIAMSON, F. and WYSZYNSKI, P. Unlocking Pre-1850 Instrumental Meteorological Records: A Global Inventory. *BAMS*, vol. 100(12), pages ES389–ES413, ISSN 0003-0007, doi:10.1175/BAMS-D-19-0040.1, 2020.

CALDWELL, P. M., BRETHERTON, C. S., ZELINKA, M. D., KLEIN, S. A., SANTER, B. D. and SANDERSON, B. M. Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys. Res. Lett.*, vol. 41(5), pages 1803–1808, ISSN 0094-8276, doi:10.1002/2014GL059205, 2014.

CARRO-CALVO, L., JAUME-SANTERO, F., GARCÍA-HERRERA, R. and SALCEDO-SANZ, S. k-Gaps: a novel technique for clustering incomplete climatological time series. *Theor. Appl. Climatol.*, ISSN 1434-4483, doi:10.1007/s00704-020-03396-w, 2020.

CHERCHI, A., AMBRIZZI, T., BEHERA, S., FREITAS, A. C. V., MORIOKA, Y. and ZHOU, T. The Response of Subtropical Highs to Climate Change. *Curr. Clim.*, vol. 4(4), pages 371–382, ISSN 2198-6061, doi:10.1007/s40641-018-0114-1, 2018.

CHERUVELIL, K. S., YUAN, S., WEBSTER, K. E., TAN, P.-N., LAPIERRE, J.-F., COLLINS, S. M., FERGUS, C. E., SCOTT, C. E., HENRY, E. N., SORANNO, P. A., FILSTRUP, C. T. and WAGNER, T. Creating multithemed ecological regions for macroscale ecology: Testing a flexible, repeatable, and accessible clustering method. *Ecol. Evol.*, vol. 7(9), pages 3046–3058, ISSN 2045-7758, doi:10.1002/ece3.2884, 2017.

Chi, J. T., Chi, E. C. and Baraniuk, R. G. k-POD: A Method for k-Means Clustering of Missing Data. *Am. Stat.*, vol. 70(1), pages 91–99, doi:10.1080/00031305.2015.1086685, 2016.

Christiansen, B. and Ljungqvist, F. C. Challenges and perspectives for large-scale temperature reconstructions of the past two millennia. *Rev. Geophys.*, vol. 55, doi:10.1002/2016rg000521, 2017.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A. J. and Wehner, M. Long-term climate change: Projections, commitments and irreversibility. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed. by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley), pages 1029–1136. Cambridge University Press, Cambridge, UK and New York, USA, 2013.

Comas-Bru, L. and Hernández, A. Reconciling north atlantic climate modes: revised monthly indices for the east atlantic and the scandinavian patterns beyond the 20th century. *Earth Syst. Sci. Data*, vol. 10(4), pages 2329–2344, doi:10.5194/essd-10-2329-2018, 2018.

Comboul, M., Emile-Geay, J., Hakim, G. J. and Evans, M. N. Paleoclimate Sampling as a Sensor Placement Problem. *J. Clim.*, vol. 28, doi:10.1175/jcli-d-14-00802.1, 2015.

2k Consortium, P. Continental-scale temperature variability during the past two millennia. *Nat. Geosci.*, vol. 6, doi:10.1038/ngeo1797, 2013.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, vol. 13(1), pages 21–27, ISSN 0018-9448, doi:10.1109/TIT.1967.1053964, 1967.

CRAM, T. A., COMPO, G. P., YIN, X., ALLAN, R. J., MCCOLL, C., VOSE, R. S., WHITAKER, J. S., MATSUI, N., ASHCROFT, L., AUCHMANN, R., BESSEMOULIN, P., BRANDSMA, T., BROHAN, P., BRUNET, M., COMEAUX, J., CROUTHAMEL, R., GLEASON JR, B. E., GROISMAN, P. Y., HERSBACH, H., JONES, P. D., JÓNSSON, T., JOURDAIN, S., KELLY, G., KNAPP, K. R., KRUGER, A., KUBOTA, H., LENTINI, G., LORREY, A., LOTT, N., LUBKER, S. J., LUTERBACHER, J., MARSHALL, G. J., MAUGERI, M., MOCK, C. J., MOK, H. Y., NORDLI, Ø., RODWELL, M. J., ROSS, T. F., SCHUSTER, D., SRNEC, L., VALENTE, M. A., VIZI, Z., WANG, X. L., WESTCOTT, N., WOOLLEN, J. S. and WORLEY, S. J. The International Surface Pressure Databank version 2. *Geosci. Data J.*, vol. 2(1), pages 31–46, ISSN 2049-6060, doi:10.1002/gdj3.25, 2015.

CROWLEY, T. J. Causes of climate change over the past 1000 years. *Science*, vol. 289, doi:10.1126/science.289.5477.270, 2000.

DAVIS, R. E., HAYDEN, B. P., GAY, D. A., PHILLIPS, W. L. and JONES, G. V. The North Atlantic Subtropical Anticyclone. *J. Clim.*, vol. 10(4), pages 728–744, doi:10.1175/1520-0442(1997)010<0728:TNASA>2.0.CO;2, 1997.

DEL SER, J., OSABA, E., MOLINA, D., YANG, X., SALCEDO-SANZ, S., CAMACHO, D., DAS, S., N. SUGANTHAN, P., COELLO COELLO, C. A. and HERRERA, F. Bio-inspired computation: Where we stand and what's next. *Swarm Evol. Comput.*, vol. 48, pages 220 – 250, ISSN 2210-6502, doi:10.1016/j.swevo.2019.04.008, 2019.

DEMUZERE, M., KASSOMENOS, P. and PHILIPP, A. The COST733 circulation type classification software: an example for surface ozone concentrations in Central Europe. *Theor. Appl. Climatol.*, vol. 105(1), pages 143–166, ISSN 1434-4483, doi:10.1007/s00704-010-0378-4, 2011.

DIXON, J. K. Pattern Recognition with Partly Missing Data. *IEEE Trans. Syst. Man. Cybern. B Cybern.*, vol. 9(10), pages 617–621, ISSN 0018-9472, doi:10.1109/TSMC.1979.4310090, 1979.

EIBEN, A. E. and SMITH, J. From evolutionary computation to the evolution of things. *Nature*, vol. 521, doi:10.1038/nature14544, 2015.

EMILE-GEAY, J., MCKAY, N. P., KAUFMAN, D. S., VON GUNTEN, L., WANG, J., ANCHUKAITIS, K. J., ABRAM, N. J., ADDISON, J. A., CURRAN, M. A., EVANS, M. N., HENLEY, B. J., HAO, Z., MARTRAT, B., MCGREGOR, H. V., NEUKOM, R., PEDERSON, G. T., STENNI, B., THIRUMALAI, K., WERNER, J. P., XU, C., DIVINE, D. V., DIXON, B. C., GERGIS, J., MUNDO, I. A., NAKATSUKA, T., PHIPPS, S. J., ROUTSON, C. C., STEIG, E. J., TIERNEY, J. E., TYLER, J. J., ALLEN, K. J., BERTLER, N. A., BJÖRKLUND, J., CHASE, B. M., CHEN, M.-T., COOK, E., DE JONG, R., DELONG, K. L., DIXON, D. A., EKAYKIN, A. A., ERSEK, V., FILIPSSON, H. L., FRANCUS, P., FREUND, M. B., FREZZOTTI, M., GAIRE, N. P., GAJEWSKI, K., GE, Q., GOOSSE, H., GORNOSTAEVA, A., GROSJEAN, M., HORIUCHI, K., HORMES, A., HUSUM, K., ISAKSSON, E., KANDASAMY, S., KAWAMURA, K., KILBOURNE, K. H., KOÇ, N., LEDUC, G., LINDERHOLM, H. W., LORREY, A. M., MIKHALENKO, V., MORTYN, P. G., MOTOYAMA, H., MOY, A. D., MULVANEY, R., MUNZ, P. M., NASH, D. J., OERTER, H., OPEL, T., ORSI, A. J., OVCHINNIKOV, D. V., PORTER, T. J., ROOP, H. A., SAENGER, C., SANO, M., SAUCHYN, D., SAUNDERS, K. M., SEIDENKRANTZ, M.-S., SEVERI, M., SHAO, X., SICRE, M.-A., SIGL, M., SINCLAIR, K., ST. GEORGE, S., ST. JACQUES, J.-M., THAMBAN, M., KUWAR THAPA, U., THOMAS, E. R., TURNEY, C., UEMURA, R., VIAU, A. E., VLADIMIROVA, D. O., WAHL, E. R., WHITE, J. W., YU, Z., ZINKE, J. and PAGES2K-CONSORTIUM. A global multiproxy database for temperature reconstructions of the Common Era. *Sci. Data*, vol. 4(1), page 170088, ISSN 2052-4463, doi:10.1038/sdata.2017.88, 2017.

EVANS, M. N., A., K., A., C. M. and R., V. Globality and optimality in climate field reconstructions from proxy data. In *(ed. Markgraf V.) Interhemispheric Climate Linkages.*, pages 53–72. Elsevier BV, 2001.

EVANS, M. N., KAPLAN, A. and CANE, M. A. Optimal sites for coral-based

reconstruction of global sea surface temperature. *Paleoceanogr. Paleoclimatol.*, vol. 13, doi:10.1029/98pa02132, 1998.

EVANS, M. N., SMERDON, J. E., KAPLAN, A., TOLWINSKI-WARD, S. E. and GONZÁLEZ-ROUCO, J. F. Climate field reconstruction uncertainty arising from multivariate and nonlinear properties of predictors. *Geophys. Res. Lett.*, vol. 41, doi:10.1002/2014GL062063, 2014.

EYRING, V., BONY, S., MEEHL, G. A., SENIOR, C. A., STEVENS, B., STOUFFER, R. J. and TAYLOR, K. E. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, vol. 9(5), pages 1937–1958, doi:10.5194/gmd-9-1937-2016, 2016.

FALARZ, M. Azores High and Hawaiian High: correlations, trends and shifts (1948–2018). *Theor. Appl. Climatol.*, vol. 138(1-2), pages 417–431, doi:10.1007/s00704-019-02837-5, 2019.

FERNÁNDEZ-DONADO, L., GONZÁLEZ-ROUCO, J. F., RAIBLE, C. C., AMMANN, C. M., BARRIOPEDRO, D., GARCÍA-BUSTAMANTE, E., JUNGCLAUS, J. H., LORENZ, S. J., LUTERBACHER, J., PHIPPS, S. J., SERVONNAT, J., SWINGEDOUW, D., TETT, S. F. B., WAGNER, S., YIOU, P. and ZORITA, E. Large-scale temperature response to external forcing in simulations and reconstructions of the last millennium. *Clim. Past*, vol. 9(1), pages 393–421, doi:10.5194/cp-9-393-2013, 2013.

FLATO, G., MAROTZKE, J., ABIODUN, B., BRACONNOT, P., CHOU, S. C., COLLINS, W., COX, P., DRIOUECH, F., EMORI, S., EYRING, V., FOREST, C., GLECKLER, P., GUILYARDI, E., JAKOB, C., KATTSOV, V., REASON, C. and RUMMUKAINEN, M. Evaluation of Climate Models. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed. by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley), pages 741–866. Cambridge University Press, Cambridge, UK and New York, USA, 2013.

FORREST, S. Genetic algorithms: principles of natural selection applied to computation. *Science*, vol. 261, doi:10.1126/science.8346439, 1993.

FRANKE, J., BRÖNNIMANN, S., BHEND, J. and BRUGNARA, Y. A monthly global paleo-reanalysis of the atmosphere from 1600 to 2005 for studying past climatic variations. *Sci. Data*, vol. 4, doi:10.1038/sdata.2017.76, 2017.

FRANKE, J., GONZÁLEZ-ROUCO, J. F., FRANK, D. and GRAHAM, N. E. 200 years of European temperature variability: insights from and tests of the proxy surrogate reconstruction analog method. *Clim. Dyn.*, vol. 37(1), pages 133–150, ISSN 1432-0894, doi:10.1007/s00382-010-0802-6, 2011.

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*, vol. 16(3), pages 1061–1074, doi:10.5194/cp-16-1061-2020, 2020.

GAGNE II, D. J., CHRISTENSEN, H. M., SUBRAMANIAN, A. C. and MONAHAN, A. H. Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz '96 Model. *J. Adv. Model. Earth Syst.*, vol. 12(3), page e2019MS001896, ISSN 1942-2466, doi:10.1029/2019MS001896, 2020.

GAO, H., CHEN, J., WANG, B., TAN, S. C., LEE, C. M., YAO, X., YAN, H. and SHI, J. A study of air pollution of city clusters. *Atmos. Environ.*, vol. 45(18), pages 3069 – 3077, ISSN 1352-2310, doi:10.1016/j.atmosenv.2011.03.018, 2011.

GLASER, R. and RIEMANN, D. A thousand-year record of temperature variations for Germany and Central Europe based on documentary data. *J. Quat. Sci.*, vol. 24(5), pages 437–449, ISSN 1099-1417, doi:10.1002/jqs.1302, 2009.

GÓMEZ-NAVARRO, J. J., WERNER, J., WAGNER, S., LUTERBACHER, J. and ZORITA, E. Establishing the skill of climate field reconstruction techniques for precipitation with pseudoproxy experiments. *Clim. Dyn.*, vol.

45(5), pages 1395–1413, ISSN 1432-0894, doi:10.1007/s00382-014-2388-x, 2015.

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*, vol. 13, doi:10.5194/cp-13-629-2017, 2017.

HAKIM, G. J., EMILE-GEAY, J., STEIG, E. J., NOONE, D., ANDERSON, D. M., TARDIF, R., STEIGER, N. and PERKINS, W. A. The Last Millennium climate reanalysis project: Framework and first results. *J. Geophys. Res. Atmos.*, vol. 121(12), pages 6745–6764, doi:10.1002/2016JD024751, 2016.

HANSEN, J., SATO, M., KHARECHA, P. and VON SCHUCKMANN, K. Earth's energy imbalance and implications. *Atmos. Chem. Phys.*, vol. 11, pages 13421–13449, doi:10.5194/acp-11-13421-2011, 2011.

HARTIGAN, J. A. and WONG, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. Royal Stat. Soc. Series C*, vol. 28(1), pages 100–108, ISSN 00359254, 14679876, doi:10.2307/2346830, 1979.

HASANEAN, H. M. Variability of the North Atlantic subtropical high and associations with tropical sea-surface temperature. *Int. J. Climatol.*, vol. 24(8), pages 945–957, ISSN 0899-8418, doi:10.1002/joc.1042, 2004.

HAUSFATHER, Z., DRAKE, H. F., ABBOTT, T. and SCHMIDT, G. A. Evaluating the Performance of Past Climate Model Projections. *Geophys. Res. Lett.*, vol. 47(1), page e2019GL085378, ISSN 0094-8276, doi:10.1029/2019GL085378, 2020.

HAYLOCK, M. R., HOFSTRA, N., KLEIN TANK, A. M. G., KLOK, E. J., JONES, P. D. and NEW, M. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *J. Geophys. Res. Atmos.*, vol. 113(D20), ISSN 2156-2202, doi:10.1029/2008JD010201, 2008.

HE, C., WU, B., ZOU, L. and ZHOU, T. Responses of the Summertime Subtropical Anticyclones to Global Warming. *J. Clim.*, vol. 30(16), pages 6465–6479, doi:10.1175/JCLI-D-16-0529.1, 2017.

HE, X., CHANEY, N. W., SCHLEISS, M. and SHEFFIELD, J. Spatial downscaling of precipitation using adaptable random forests. *Water Resour. Res.*, vol. 52(10), pages 8217–8237, ISSN 0043-1397, doi:10.1002/2016WR019034, 2016.

HENN, B., RALEIGH, M. S., FISHER, A. and LUNDQUIST, J. D. A Comparison of Methods for Filling Gaps in Hourly Near-Surface Air Temperature Data. *J. Hydrometeorol.*, vol. 14(3), pages 929–945, doi:10.1175/JHM-D-12-027.1, 2013.

HERNÁNDEZ, A., SÁNCHEZ-LÓPEZ, G., PLA-RABES, S., COMAS-BRU, L., PARNELL, A., CAHILL, N., GEYER, A., TRIGO, R. M. and GIRALT, S. A 2000-year Bayesian NAO reconstruction from the Iberian Peninsula. *Sci. Rep.*, vol. 10(1), page 14961, ISSN 2045-2322, doi:10.1038/s41598-020-71372-5, 2020.

HO, D. Artificial Intelligence in cancer therapy. *Science*, vol. 367(6481), page 982, doi:10.1126/science.aaz3023, 2020.

HODSON, R. Digital Revolution. *Nat. Outlook*, vol. 563(7733), page 1, doi:10.1038/d41586-018-07500-z, 2018.

HORTON, D. E., JOHNSON, N. C., SINGH, D., SWAIN, D. L., RAJARATNAM, B. and DIFFENBAUGH, N. S. Contribution of changes in atmospheric circulation patterns to extreme temperature trends. *Nature*, vol. 522(7557), page 465, doi:10.1038/nature14550, 2015.

HOTELLING, H. Relation between two sets of variates*. *Biometrika*, vol. 28(3-4), pages 321–377, ISSN 0006-3444, doi:10.1093/biomet/28.3-4.321, 1936.

HU, J., EMILE-GEAY, J. and PARTIN, J. Correlation-based interpretations of paleoclimate data – where statistics meet past climates. *Earth Planet. Sci. Lett.*, vol. 459, doi:10.1016/j.epsl.2016.11.048, 2017.

HUBER, M. and KNUTTI, R. Anthropogenic and natural warming inferred from changes in Earth's energy balance. *Nat. Geosci.*, vol. 5(1), pages 31–36, ISSN 1752-0908, doi:10.1038/ngeo1327, 2012.

HURRELL, J., TRENBERTH, K. and NCAR. The Climate Data Guide: NCAR Sea Level Pressure. In *(Eds. National Center for Atmospheric Research Staff) Retrieved from https://climatedataguide.ucar.edu/climate-data/ncar-sea-level-pressure*. 2020.

HURRELL, J. W. and DESER, C. North Atlantic climate variability: The role of the North Atlantic Oscillation. *J. Mar. Syst.*, vol. 78(1), pages 28 – 41, ISSN 0924-7963, doi:10.1016/j.jmarsys.2008.11.026, 2010.

ILES, C. E., VAUTARD, R., STRACHAN, J., JOUSSAUME, S., EGGEN, B. R. and HEWITT, C. D. The benefits of increasing resolution in global and regional climate simulations for European climate extremes. *Geosci. Model Dev.*, vol. 13(11), pages 5583–5607, doi:10.5194/gmd-13-5583-2020, 2020.

ILYAS, M., BRIERLEY, C. M. and GUILLAS, S. Uncertainty in regional temperatures inferred from sparse global observations: Application to a probabilistic classification of El Niño. *Geophys. Res. Lett.*, vol. 44(17), pages 9068–9074, ISSN 0094-8276, doi:10.1002/2017GL074596, 2017.

IOANNIDOU, L. and YAU, M. K. A climatology of the Northern Hemisphere winter anticyclones. *J. Geophys. Res. Atmos.*, vol. 113(D8), doi:10.1029/2007JD008409, 2008.

IPCC. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed. by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley), pages 1–1535. Cambridge University Press, Cambridge, UK and New York, USA, 2013.

IPCC. Summary for policymakers. In *Climate Change 2014: Impacts,Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed. by C. Field, V. R. Barros, D. J. Dokken, K. J. Mach, M. D. Mastrandrea, T. E. Bilir, M. Chatterjee, K. L. Ebi, Y. O. Estrada, R. C. Genova, B. Girma, E. S. Kissel, A. N. Levy, S. MacCracken, P. R. Mastrandrea and L. L. White), pages 1–32. Cambridge University Press, Cambridge, UK and New York, USA, 2014.

IPCC. Global Warming of 1.5°C. In *An IPCC Special Report on the impacts of global warming of 1.5° C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* (ed. by V. Masson-Delmotte, P. Zhai, H. O. Pörtner, D. Roberts, J. Skea, P. R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J. B. R. Matthews, Y. Chen, X. Zhou, M. Gomis, E. Lonnoy, T. Maycock, M. Tignor and T. Waterfield), pages 1–616. Cambridge University Press, Cambridge, UK and New York, USA, 2018.

IQBAL, M. J., REHMAN, S. U., HAMEED, S. and QURESHI, M. A. Changes in Hadley circulation: the Azores high and winter precipitation over tropical northeast Africa. *Theor. Appl. Climatol.*, vol. 137(3-4), pages 2941–2948, doi:10.1007/s00704-019-02765-4, 2019.

JAUME-SANTERO, F., BARRIOPEDRO, D., GARCÍA-HERRERA, R., CALVO, N. and SALCEDO-SANZ, S. Selection of optimal proxy locations for temperature field reconstructions using evolutionary algorithms. *Sci. Rep.*, vol. 10(1), ISSN 2045-2322, doi:10.1038/s41598-020-64459-6, 2020.

JAUME-SANTERO, F., BARRIOPEDRO, D., GARCÍA-HERRERA, R. F. and LUTERBACHER, J. North Atlantic Sea Level Pressure reconstruction back to 1750 CE using optimized networks of observations. *J. Clim.*, Submitted, 2021.

JIN, C., CHEN, W., CAO, Y., XU, Z., TAN, Z., ZHANG, X., DENG, L., ZHENG, C., ZHOU, J., SHI, H. and FENG, J. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nat. Commun.*, vol. 11(1), page 5088, ISSN 2041-1723, doi:10.1038/s41467-020-18685-1, 2020.

JONES, P. D., JONSSON, T. and WHEELER, D. Extension to the North Atlantic oscillation using early instrumental pressure observations from Gibraltar and south-west Iceland. *Int. J. Clim.*, vol. 17(13), pages 1433–1450, doi:10.1002/(SICI)1097-0088(19971115)17:13<1433::AID-JOC203>3.0.CO;2-P, 1997.

JONES, P. D., NEW, M., PARKER, D. E., MARTIN, S. and RIGOR, I. G. Surface air temperature and its variations over the last 150 years. *Rev. Geophys.*, vol. 37, doi:10.1029/1999rg900002, 1999.

JUNGCLAUS, J. H., BARD, E., BARONI, M., BRACONNOT, P., CAO, J., CHINI, L. P., EGOROVA, T., EVANS, M., GONZÁLEZ-ROUCO, J. F., GOOSSE, H., HURTT, G. C., JOOS, F., KAPLAN, J. O., KHODRI, M., KLEIN GOLDEWIJK, K., KRIVOVA, N., LEGRANDE, A. N., LORENZ, S. J., LUTERBACHER, J., MAN, W., MAYCOCK, A. C., MEINSHAUSEN, M., MOBERG, A., MUSCHELER, R., NEHRBASS-AHLES, C., OTTO-BLIESNER, B. I., PHIPPS, S. J., PONGRATZ, J., ROZANOV, E., SCHMIDT, G. A., SCHMIDT, H., SCHMUTZ, W., SCHURER, A., SHAPIRO, A. I., SIGL, M., SMERDON, J. E., SOLANKI, S. K., TIMMRECK, C., TOOHEY, M., USOSKIN, I. G., WAGNER, S., WU, C.-J., YEO, K. L., ZANCHETTIN, D., ZHANG, Q. and ZORITA, E. The PMIP4 contribution to CMIP6 – Part 3: The last millennium, scientific objective, and experimental design for the PMIP4 *past1000*simulations. *Geosci. Model Dev.*, vol. 10(11), pages 4005–4033, doi:10.5194/gmd-10-4005-2017, 2017.

KADOW, C., HALL, D. M. and ULBRICH, U. Artificial intelligence reconstructs missing climate information. *Nat. Geosci.*, vol. 13(6), pages 408–413, ISSN 1752-0908, doi:10.1038/s41561-020-0582-5, 2020.

KARPATNE, A., EBERT-UPHOFF, I., RAVELA, S., BABAIE, H. A. and KUMAR, V. Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Trans. Knowl. Data Eng.*, vol. 31(8), pages 1544–1554, ISSN 1558-2191, doi:10.1109/TKDE.2018.2861006, 2019.

KASHANI, A. R., GANDOMI, A. H. and MOUSAVI, M. Imperialistic Competitive Algorithm: A metaheuristic algorithm for locating the critical slip surface in 2-Dimensional soil slopes. *Geosci. Front.*, vol. 7(1), pages 83–89, ISSN 1674-9871, 2016.

KATES-HARBECK, J., SVYATKOVSKIY, A. and TANG, W. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, vol. 568(7753), pages 526–531, ISSN 1476-4687, doi:10.1038/s41586-019-1116-4, 2019.

KAUFMAN, D. S. Recent warming reverses long-term Arctic cooling. *Science*, vol. 325, doi:10.1126/science.1173983, 2009.

KELL, D. B. Scientific discovery as a combinatorial optimisation problem: how best to navigate the landscape of possible experiments? *BioEssays*, vol. 34(3), pages 236–244, ISSN 1521-1878, doi:10.1002/bies.201100144, 2012.

KETTENRING, J. R. The Practice of Cluster Analysis. *J. Classif.*, vol. 23(1), pages 3–30, ISSN 1432-1343, doi:10.1007/s00357-006-0002-6, 2006.

KNAPP, T. R. Canonical correlation analysis: A general parametric significance-testing system. *Psychol. Bull.*, vol. 85(2), pages 410–416, doi:10.1037/0033-2909.85.2.410, 1978.

KNÜSEL, B. Applying big data beyond small problems in climate research. *Nat. Clim. Change*, vol. 9, doi:10.1038/s41558-019-0404-1, 2019.

KÖPPEN, W. Die Wärmezonen der Erde, nach der Dauer der heissen, gemässigten und kalten Zeit und nach der Wirkung der Wärme auf die organische Welt betrachtet (The thermal zones of the earth according to the duration of hot, moderate and cold periods and to the impact

of heat on the organic world). *Meteorol. Z.*, vol. 1(21), pages 215–226, doi:10.1127/metz/2016/0816, 1884.

KÜTTEL, M., XOPLAKI, E., GALLEGO, D., LUTERBACHER, J., GARCÍA-HERRERA, R., ALLAN, R., BARRIENDOS, M., JONES, P. D., WHEELER, D. and WANNER, H. The importance of ship log data: reconstructing North Atlantic, European and Mediterranean sea level pressure fields back to 1750. *Clim. Dyn.*, vol. 34(7), pages 1115–1128, ISSN 1432-0894, doi:10.1007/s00382-009-0577-9, 2010.

LANGE, K., HUNTER, D. R. and YANG, I. Optimization Transfer Using Surrogate Objective Functions. *J. Comput. Graph. Stat.*, vol. 9(1), pages 1–20, ISSN 10618600, doi:10.2307/1390605, 2000.

LECUN, Y., BENGIO, Y. and HINTON, G. Deep Learning. *Nature*, vol. 521(7553), pages 436–444, ISSN 1476-4687, doi:10.1038/nature14539, 2015.

LEWIS, S. C. and LEGRANDE, A. N. Stability of ENSO and its tropical Pacific teleconnections over the Last Millennium. *Clim. Past*, vol. 11, doi:10.5194/cp-11-1347-2015, 2015.

LI, W., LI, L., TING, M. and LIU, Y. Intensification of Northern Hemisphere subtropical highs in a warming climate. *Nat. Geosci.*, vol. 5(11), pages 830–834, ISSN 1752-0908, doi:10.1038/ngeo1590, 2012.

LORENZ, E. N. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *J. Atmos. Sci.*, vol. 26(4), pages 636–646, doi:10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2, 1969.

LUTERBACHER, J., DIETRICH, D., XOPLAKI, E., GROSJEAN, M. and WANNER, H. European Seasonal and Annual Temperature Variability, Trends, and Extremes Since 1500. *Science*, vol. 303(5663), pages 1499–1503, ISSN 0036-8075, doi:10.1126/science.1093877, 2004.

LUTERBACHER, J., XOPLAKI, E., DIETRICH, D., JONES, P. D., DAVIES, T. D., PORTIS, D., GONZALEZ-ROUCO, J. F., VON STORCH, H., GYAL-

ISTRAS, D., CASTY, C. and WANNER, H. Extending North Atlantic oscillation reconstructions back to 1500. *Atmos. Sci. Lett*, vol. 2(1-4), pages 114–124, doi:10.1006/asle.2002.0047, 2001.

LUTERBACHER, J., XOPLAKI, E., DIETRICH, D., RICKLI, R., JACOBEIT, J., BECK, C., GYALISTRAS, D., SCHMUTZ, C. and WANNER, H. Reconstruction of sea level pressure fields over the Eastern North Atlantic and Europe back to 1500. *Clim. Dyn.*, vol. 18(7), pages 545–561, ISSN 1432-0894, doi:10.1007/s00382-001-0196-6, 2002.

MANN, M. E. Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. *Science*, vol. 326, doi:10.1126/science.1177303, 2009.

MASSON-DELMOTTE, V., SCHULZ, M., ABE-OUCHI, A., BEER, J., GANOPOLSKI, A., ROUCO, J. F. G., JANSEN, E., LAMBECK, K., LUTERBACHER, J., NAISH, T., OSBORN, T., OTTO-BLIESNER, B., QUINN, T., RAMESH, R., ROJAS, M., SHAO, X. and TIMMERMANN, A. Information from Paleoclimate Archives. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed. by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley), pages 383–464. Cambridge University Press, Cambridge, UK and New York, USA, 2013.

MAXWELL, A. E., WARNER, T. A. and FANG, F. Implementation of machine-learning classification in remote sensing: an applied review. *Int. J. Remote Sens.*, vol. 39(9), pages 2784–2817, ISSN 0143-1161, doi:10.1080/01431161.2018.1433343, 2018.

MELLADO-CANO, J., BARRIOPEDRO, D., GARCÍA-HERRERA, R., TRIGO, R. M. and HERNÁNDEZ, A. Examining the North Atlantic Oscillation, East Atlantic Pattern, and Jet Variability since 1685. *J. Clim.*, vol. 32(19), pages 6285–6298, ISSN 0894-8755, doi:10.1175/JCLI-D-19-0135.1, 2019.

MIELE, V., PICARD, F. and DRAY, S. Spatially constrained clustering of ecological networks. *Methods Ecol. Evol.*, vol. 5(8), pages 771–779, ISSN 2041-210X, doi:10.1111/2041-210X.12208, 2014.

MURPHY, K. *Machine Learning: A Probabilistic Perspective*. The MIT Press. ISBN 978-0-262-01802-9, 2012.

MYHRE, G., SHINDELL, D., BRÉON, F. M., COLLINS, W., FUGLESTVEDT, J., HUANG, J., KOCH, D., LAMARQUE, J. F., LEE, D., MENDOZA, B., NAKAJIMA, T., ROBOCK, A., STEPHENS, G., TAKEMURA, T. and ZHANG, H. Anthropogenic and Natural Radiative Forcing. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed. by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley), pages 659–740. Cambridge University Press, Cambridge, UK and New York, USA, 2013.

NETZEL, P. and STEPINSKI, T. On Using a Clustering Approach for Global Climate Classification. *J. Clim.*, vol. 29(9), pages 3387–3401, ISSN 0894-8755, doi:10.1175/JCLI-D-15-0640.1, 2016.

NEUKOM, R., SCHURER, A. P., STEIGER, N. J. and HEGERL, G. C. Possible causes of data model discrepancy in the temperature history of the Last Millennium. *Sci. Rep.*, vol. 8, doi:10.1038/s41598-018-25862-2, 2018.

NEUKOM, R., STEIGER, N., GÓMEZ-NAVARRO, J. J., WANG, J. and WERNER, J. P. No evidence for globally coherent warm and cold periods over the preindustrial Common Era. *Nature*, vol. 571(7766), pages 550–554, ISSN 1476-4687, doi:10.1038/s41586-019-1401-2, 2019.

NOONE, S., ATKINSON, C., BERRY, D. I., DUNN, R. J. H., FREEMAN, E., PEREZ GONZALEZ, I., KENNEDY, J. J., KENT, E. C., KETTLE, A., MCNEILL, S., MENNE, M., STEPHENS, A., THORNE, P. W., TUCKER, W., VOCES, C. and WILLETT, K. M. Progress towards a holistic land

and marine surface meteorological database and a call for additional contributions. *Geosci. Data J.*, ISSN 2049-6060, doi:10.1002/gdj3.109, 2020.

OMRAN, M. G., ENGELBRECHT, A. P. and SALMAN, A. An overview of clustering methods. *Intell. Data Anal.*, vol. 11, pages 583–605, ISSN 1571-4128, doi:10.3233/IDA-2007-11602, 2007.

OTTO-BLIESNER, B. L. Climate Variability and Change since 850 C.E.: An Ensemble Approach with the Community Earth System Model (CESM). *BAMS*, vol. 97, doi:10.1175/bams-d-14-00233.1, 2016.

PERDINAN and WINKLER, J. A. Selection of climate information for regional climate change assessments using regionalization techniques: an example for the Upper Great Lakes Region, USA. *Int. J. Climatol.*, vol. 35(6), pages 1027–1040, doi:10.1002/joc.4036, 2015.

PHILLIPS, S. J. Acceleration of K-Means and Related Clustering Algorithms. In *Algorithm Engineering and Experiments* (ed. by D. M. Mount and C. Stein), pages 166–177. Springer Berlin Heidelberg, Berlin Heidelberg, 2002. ISBN 978-3-540-45643-8.

PIETSCH, W. The Causal Nature of Modeling with Big Data. *Philos. Technol.*, vol. 29(2), pages 137–171, ISSN 2210-5441, doi:10.1007/s13347-015-0202-2, 2016.

PINTO, J. G. and RAIBLE, C. C. Past and recent changes in the North Atlantic oscillation. *WIREs Clim. Change*, vol. 3(1), pages 79–90, ISSN 1757-7780, doi:10.1002/wcc.150, 2012.

PORTIS, D. H., WALSH, J. E., EL HAMLY, M. and LAMB, P. J. Seasonality of the North Atlantic Oscillation. *J. Clim.*, vol. 14(9), pages 2069–2078, doi:10.1175/1520-0442(2001)014<2069:SOTNAO>2.0.CO;2, 2001.

PRAKASH, S., SHARMA, A. and SAHU, S. S. Soil Moisture Prediction Using Machine Learning. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018 Second

International Conference on Inventive Communication and Computational Technologies (ICICCT), pages 1–6. 2018.

RAJAN, K. and SAFFIOTTI, A. Towards a science of integrated AI and Robotics. *Artif. Intell.*, vol. 247, pages 1–9, ISSN 0004-3702, doi:10.1016/j.artint.2017.03.003, 2017.

RAO, A. R. and SRINIVAS, V. V. Regionalization of watersheds by hybrid-cluster analysis. *J. Hydrol.*, vol. 318(1), pages 37 – 56, ISSN 0022-1694, doi:10.1016/j.jhydrol.2005.06.003, 2006.

RASP, S., PRITCHARD, M. S. and GENTINE, P. Deep Learning to represent subgrid processes in climate models. *PNAS*, vol. 115(39), page 9684, doi:10.1073/pnas.1810286115, 2018.

REICHSTEIN, M. Deep learning and process understanding for data-driven Earth system science. *Nature*, vol. 566, doi:10.1038/s41586-019-0912-1, 2019.

RIBES, A., ZWIERS, F. W., AZAÏS, J.-M. and NAVEAU, P. A new statistical approach to climate change detection and attribution. *Clim. Dyn.*, vol. 48(1), pages 367–386, ISSN 1432-0894, doi:10.1007/s00382-016-3079-6, 2017.

RICHMAN, M. B., LESLIE, L. M., RAMSAY, H. A. and KLOTZBACH, P. J. Reducing Tropical Cyclone Prediction Errors Using Machine Learning Approaches. *Procedia Comput. Sci.*, vol. 114, pages 314–323, ISSN 1877-0509, doi:10.1016/j.procs.2017.09.048, 2017.

ROUTSON, C. C. Mid-latitude net precipitation decreased with Arctic warming during the Holocene. *Nature*, vol. 1476, doi:10.1038/s41586-019-1060-3, 2019.

ROWLAND, E. Theory of Games and Economic Behavior. *Nature*, vol. 157(3981), pages 172–173, ISSN 1476-4687, doi:10.1038/157172a0, 1946.

RUBEL, F., BRUGGER, K., HASLINGER, K. and AUER, I. The climate of the European Alps: Shift of very high resolution Köppen-Geiger climate zones 1800-2100. *Meteorol. Z.*, vol. 26(2), pages 115–125, doi:10.1127/metz/2016/0816, 2017.

RUDIN, C. and RADIN, J. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Sci. Rev.*, vol. 1(2), doi:10.1162/99608f92.5a8a3a3d, 2019.

SALCEDO-SANZ, S. Modern meta-heuristics based on nonlinear physics processes: A review of models and design procedures. *Phys. Rep.*, vol. 655, pages 1–70, ISSN 0370-1573, doi:10.1016/j.physrep.2016.08.001, 2016.

SALCEDO-SANZ, S. A review on the coral reefs optimization algorithm: new development lines and current applications. *Prog. Artif. Intell.*, vol. 6, doi:10.1007/s13748-016-0104-2, 2017.

SALCEDO-SANZ, S., GARCÍA-HERRERA, R., CAMACHO-GÓMEZ, C., ALEXANDRE, E., CARRO-CALVO, L. and JAUME-SANTERO, F. Near-optimal selection of representative measuring points for robust temperature field reconstruction with the CRO-SL and analogue methods. *Glob. Planet. Change*, vol. 178, pages 15–34, doi:10.1016/j.gloplacha.2019.04.013, 2019.

SALCEDO-SANZ, S., GARCÍA-HERRERA, R., CAMACHO-GÓMEZ, C., AYBAR-RUÍZ, A. and ALEXANDRE, E. Wind power field reconstruction from a reduced set of representative measuring points. *Appl. Energ.*, vol. 228, doi:10.1016/j.apenergy.2018.07.003, 2018.

SANDERSON, B. M. and O'NEILL, B. C. Assessing the costs of historical inaction on climate change. *Sci. Rep.*, vol. 10(1), page 9173, ISSN 2045-2322, doi:10.1038/s41598-020-66275-4, 2020.

SCHMIDT, A., THORDARSON, T., OMAN, L. D., ROBOCK, A. and SELF, S. Climatic impact of the long-lasting 1783 Laki eruption: Inapplicability of mass-independent sulfur isotopic composition measurements. *J. Geophys. Res. Atmos.*, vol. 117(D23), ISSN 0148-0227, doi:10.1029/2012JD018414, 2012.

SCHMUTZ, C., LUTERBACHER, J., GYALISTRAS, D., XOPLAKI, E. and WANNER, H. Can we trust proxy-based NAO index reconstructions? *Geophys. Res. Lett.*, vol. 27(8), pages 1135–1138, ISSN 0094-8276, doi:10.1029/1999GL011045, 2000.

SCHNASE, J. L., LEE, T. J., MATTMANN, C. A., LYNNES, C. S., CINQUINI, L., RAMIREZ, P. M., HART, A. F., WILLIAMS, D. N., WALISER, D., RINSLAND, P., WEBSTER, W. P., DUFFY, D. Q., MCINERNEY, M. A., TAMKIN, G. S., POTTER, G. L. and CARRIERE, L. Big Data Challenges in Climate Science: Improving the next-generation cyberinfrastructure. *IEEE Geosci. Remote Sens. Mag.*, vol. 4(3), pages 10–22, ISSN 2168-6831, doi:10.1109/MGRS.2015.2514192, 2016.

SEYDOUX, L., BALESTRIERO, R., POLI, P., HOOP, M. D., CAMPILLO, M. and BARANIUK, R. Clustering earthquake signals and background noises in continuous seismic data with unsupervised Deep Learning. *Nat. Commun.*, vol. 11(1), page 3972, ISSN 2041-1723, doi:10.1038/s41467-020-17841-x, 2020.

SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V., LANCTOT, M., DIELEMAN, S., GREWE, D., NHAM, J., KALCHBRENNER, N., SUTSKEVER, I., LILLICRAP, T., LEACH, M., KAVUKCUOGLU, K., GRAEPEL, T. and HASSABIS, D. Mastering the game of Go with deep neural networks and tree search. *Nature*, vol. 529(7587), pages 484–489, ISSN 1476-4687, doi:10.1038/nature16961, 2016.

SILVER, D., HUBERT, T., SCHRITTWIESER, J., ANTONOGLOU, I., LAI, M., GUEZ, A., LANCTOT, M., SIFRE, L., KUMARAN, D., GRAEPEL, T., LILLICRAP, T., SIMONYAN, K. and HASSABIS, D. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, vol. 362(6419), page 1140, doi:10.1126/science.aar6404, 2018.

SLIVINSKI, L. C., COMPO, G. P., WHITAKER, J. S., SARDESHMUKH, P. D., GIESE, B. S., MCCOLL, C., ALLAN, R., YIN, X., VOSE, R.,

TITCHNER, H., KENNEDY, J., SPENCER, L. J., ASHCROFT, L., BRÖNNIMANN, S., BRUNET, M., CAMUFFO, D., CORNES, R., CRAM, T. A., CROUTHAMEL, R., DOMÍNGUEZ-CASTRO, F., FREEMAN, J. E., GERGIS, J., HAWKINS, E., JONES, P. D., JOURDAIN, S., KAPLAN, A., KUBOTA, H., BLANCQ, F. L., LEE, T.-C., LORREY, A., LUTERBACHER, J., MAUGERI, M., MOCK, C. J., MOORE, G. K., PRZYBYLAK, R., PUDMENZKY, C., REASON, C., SLONOSKY, V. C., SMITH, C. A., TINZ, B., TREWIN, B., VALENTE, M. A., WANG, X. L., WILKINSON, C., WOOD, K. and WYSZYNSKI, P. Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Q. J. R. Meteorol. Soc.*, vol. 145(724), pages 2876–2908, doi:10.1002/qj.3598, 2019.

SMERDON, J. E. Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. *WIREs: Clim. Change*, vol. 3, doi:10.1002/wcc.149, 2011.

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. Clim.*, vol. 23, doi:10.1175/2010JCLI3328.1, 2010.

SONG, Y.-C., MENG, H.-D., O'GRADY, M. J. and O'HARE, G. M. P. The application of cluster analysis in geophysical data interpretation. *Comput. Geosci.*, vol. 14(2), pages 263–271, ISSN 1573-1499, doi:10.1007/s10596-009-9150-1, 2010.

SOTO, R., GÓMEZ-PULIDO, J. A., CARO, S. and LANZA-GUTIÉRREZ, J. M. Data Science and AI-Based Optimization in Scientific Programming. *Sci. Program.*, vol. 2019, page 7154765, ISSN 1058-9244, doi:10.1155/2019/7154765, 2019.

SOUSA, P. M., BARRIOPEDRO, D., RAMOS, A. M., GARCÍA-HERRERA, R., ESPÍRITO-SANTO, F. and TRIGO, R. M. Saharan air intrusions as a relevant mechanism for Iberian heatwaves: The record breaking events

of August 2018 and June 2019. *Weather. Clim. Extremes*, vol. 26, page 100224, ISSN 2212-0947, doi:10.1016/j.wace.2019.100224, 2019.

SOUSA, P. M., TRIGO, R. M., BARRIOPEDRO, D., SOARES, P. M. M. and SANTOS, J. A. European temperature responses to blocking and ridge regional patterns. *Clim. Dyn.*, vol. 50(1), pages 457–477, ISSN 1432-0894, doi:10.1007/s00382-017-3620-2, 2018.

STEFFEN, W., BROADGATE, W., DEUTSCH, L., GAFFNEY, O. and LUDWIG, C. The trajectory of the Anthropocene: The Great Acceleration. *Anthr. Rev.*, vol. 2(1), pages 81–98, ISSN 2053-0196, doi:10.1177/2053019614564785, 2015.

STEIG, E. J. and NEFF, P. D. The prescience of paleoclimatology and the future of the Antarctic ice sheet. *Nat. Commun.*, vol. 9(1), page 2730, ISSN 2041-1723, doi:10.1038/s41467-018-05001-1, 2018.

STOTHERS, R. B. The Great Dry Fog of 1783. *Clim. Change*, vol. 32, pages 79–89, doi:10.1007/BF00141279, 1996.

SWARNKAR, A. and SWARNKAR, A. Artificial Intelligence Based Optimization Techniques: A Review. In *(eds. Kalam A., Niazi K., Soni A., Siddiqui S., Mundra A.) Intelligent Computing Techniques for Smart Energy Systems. Lecture Notes in Electrical Engineering*. 2019.

TALENTO, S., SCHNEIDER, L., WERNER, J. and LUTERBACHER, J. Millennium-length precipitation reconstruction over south-eastern Asia: a pseudo-proxy approach. *Earth Sys. Dyn.*, vol. 10(2), pages 347–364, doi:10.5194/esd-10-347-2019, 2019.

TARDIF, R., HAKIM, G. J., PERKINS, W. A., HORLICK, K. A., ERB, M. P., EMILE-GEAY, J., ANDERSON, D. M., STEIG, E. J. and NOONE, D. Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling. *Clim. Past*, vol. 15(4), pages 1251–1273, doi:10.5194/cp-15-1251-2019, 2019.

TCHEMISOVA, T., FREITAS, A., PLAKHOV, A. and WEBER, G. W. Special issue: Optimization in the natural sciences. *Optimization*, vol. 64(6), pages 1363–1365, doi:10.1080/02331934.2015.1027530, 2015.

THORDARSON, T. and SELF, S. Atmospheric and environmental effects of the 1783–1784 Laki eruption: A review and reassessment. *J. Geophys.Res. Atmos.*, vol. 108(D1), pages AAC 7–1–AAC 7–29, ISSN 0148-0227, doi:10.1029/2001JD002042, 2003.

TURING, A. M. I.-Computing Machinery and Intelligence. *Mind*, vol. LIX(236), pages 433–460, ISSN 0026-4423, doi:10.1093/mind/LIX.236.433, 1950.

VACCARO, A., EMILE-GEAY, J., GUILLOT, D., VERNA, R., MORICE, C., KENNEDY, J. and RAJARATNAM, B. Climate field completion via Markov random fields – Application to the HadCRUT4.6 temperature dataset. *J. Clim.*, pages 1–66, doi:10.1175/JCLI-D-19-0814.1, 2021.

VADLAMANI, S. K., XIAO, T. P. and YABLONOVITCH, E. Physics successfully implements Lagrange multiplier optimization. *PNAS*, vol. 117(43), pages 26639–26650, ISSN 0027-8424, doi:10.1073/pnas.2015192117, 2020.

DE VARGAS, R. R. and BEDREGAL, B. R. C. A Way to Obtain the Quality of a Partition by Adjusted Rand Index. In *2013 2nd Workshop-School on Theoretical Computer Science*, pages 67–71. IEEE, 2013.

VRUGT, J. A. and ROBINSON, B. A. Improved evolutionary optimization from genetically adaptive multimethod search. *PNAS*, vol. 104, doi:10.1073/pnas.0610471104, 2007.

WANG, S., LI, G., GONG, Z., DU, L., ZHOU, Q., MENG, X., XIE, S. and ZHOU, L. Spatial distribution, seasonal variation and regionalization of PM2.5 concentrations in China. *Sci. China Chem.*, vol. 58(9), pages 1435–1443, ISSN 1869-1870, doi:10.1007/s11426-015-5468-9, 2015.

WANNER, H., RICKLI, R., SALVISBERG, E., SCHMUTZ, C. and SCHÜEPP, M. Global climate change and variability and its influence on Alpine

climate — concepts and observations. *Theor. Appl. Climatol.*, vol. 58(3), pages 221–243, ISSN 1434-4483, doi:10.1007/BF00865022, 1997.

WILKS, D. S. *Chapter 13 - Canonical Correlation Analysis (CCA)*, vol. 100 of *Statistical Methods in the Atmospheric Sciences*, pages 563–582. Academic Press, 2011.

YANG, X.-S., CHIEN, S. F. and TING, T. O. Computational Intelligence and Metaheuristic Algorithms with Applications. *Sci. World J.*, vol. 2014, page 425853, ISSN 2356-6140, doi:10.1155/2014/425853, 2014.

YOO, D. G. and KIM, J. H. Meta-heuristic algorithms as tools for hydrological science. *Geosci. Lett.*, vol. 1(1), page 4, ISSN 2196-4092, doi:10.1186/2196-4092-1-4, 2014.

YUKIMOTO, S., KAWAI, H., KOSHIRO, T., OSHIMA, N., YOSHIDA, K., URAKAWA, S., TSUJINO, H., DEUSHI, M., TANAKA, T., HOSAKA, M., YABU, S., YOSHIMURA, H., SHINDO, E., MIZUTA, R., OBATA, A., ADACHI, Y. and ISHII, M. The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component. *J. Meteorol. Soc. Jpn. Ser. II*, vol. 97(5), pages 931–965, doi:10.2151/jmsj.2019-051, 2019.

ZAMBRI, B., ROBOCK, A., MILLS, M. J. and SCHMIDT, A. Modeling the 1783-1784 Laki Eruption in Iceland: 2. Climate Impacts. *J. Geophys. Res. Atmos.*, vol. 124(13), pages 6770–6790, doi:10.1029/2018JD029554, 2019.

ZHANG, Y., MOGES, S. and BLOCK, P. Optimal Cluster Analysis for Objective Regionalization of Seasonal Precipitation in Regions of High Spatial-Temporal Variability: Application to Western Ethiopia. *J. Clim.*, vol. 29(10), pages 3697–3717, doi:10.1175/JCLI-D-15-0582.1, 2016.

ZHOU, S., ZHOU, K., WANG, J., YANG, G. and WANG, S. Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. *Front. Earth Sci.*, vol. 12(3), pages 491–505, ISSN 2095-0209, doi:10.1007/s11707-017-0682-8, 2018.

Zishka, K. M. and Smith, P. J. The Climatology of Cyclones and Anticyclones over North America and Surrounding Ocean Environs for January and July, 1950–77. *Mon. Weather Rev.*, vol. 108(4), pages 387–401, doi:10.1175/1520-0493(1980)108<0387:TCOCAA>2.0.CO;2, 1980.