CLINT

CLIMATE INTELLIGENCE

REVIEW OF ML ALGORITHMS FOR CLIMATE SCIENCE

November 2022

This project is part of the H2020 Programme supported by the European Union, having received funding from it under Grant Agreement No 101003876

Programme Call:	Building a low-carbon, climate resilient future: climate action in support of the Paris Agreement (H2020-LC-CLA-2018-2019-2020)
Grant agreement ID:	101003876
Project Title:	CLINT
Partners:	POLIMI (Project Coordinator), CMCC, HEREON, CSIC, SMHI, HKV, E3M, TCDF, DKRZ, IHE, ECMWF, UAH, JLU, OGC, UCM
Work-Package:	WP2: Climate Intelligence
Deliverable #:	D2.1
Deliverable Type:	Report
Contractual Date of Delivery:	30 November 2022
Actual Date of Delivery:	29 November 2022
Title of Document:	Review of ML algorithms for Climate Science
Responsible partner:	POLIMI
Author(s):	Paolo Bonetti, Alberto Maria Metelli, Marcello Restelli, Guido Ascenso, Matteo Giuliani, Andrea Castelletti, Sancho Salcedo-Sanz, Claudia Bertini.
Content of this report:	Review of Machine Learning methods and applications on Climate Science. The review contains a theoretical overview of the main ML subfields of interest and their algorithm, together with applications for the detection, causation and attribution of these methods on the four classes of Extreme Events addressed in the project.
Availability:	This report is public.

Document revisions		
Author	*Revision content*	*Date*
Eduardo Zorita	D21_v0_EZ: internal review	16.11.2022
Étienne Plésiat	D21_v0: internal review	17.11.2022
Christopher Kadow	D21_v0: internal review	17.11.2022
Paolo Bonetti	D21_v1: revision after internal review	23.11.2022
Andrea Castelletti	D21_v1: revision	24.11.2022
Paolo Bonetti	D21_v2: revision after Coordinator review	25.11.2022
Elena Matta	D2.1_v3: final check for submission	29.11.2022
Paolo Bonetti	D2.1_F: final version	29.11.2022

About CLINT

The main objective of CLINT is the development of an Artificial Intelligence framework composed of Machine Learning techniques and algorithms to process big climate datasets for improving Climate Science in the detection, causation, and attribution of Extreme Events (EEs), namely tropical cyclones, heatwaves and warm nights, droughts, and floods. The CLINT AI framework will also cover the quantification of the EE impacts on a variety of socio-economic sectors under historical, forecasted, and projected climate conditions and across different spatial scales (from European to local), ultimately developing innovative and sectorial AI-enhanced Climate Services. Finally, these services will be operationalized into Web Processing Services, according to the most advanced open data and software standards by Climate Services Information Systems, and into a commercial Demonstrator to facilitate the uptake of project results by public and private entities for research and Climate Services development.

More information: https://climateintelligence.eu/

Disclaimer

Use of any knowledge, information or data contained in this document shall be at the user's sole risk. Neither the CLINT consortium nor any of its members, their officers, employees or agents shall be liable or responsible, in negligence or otherwise, for any loss, damage or expense whatever sustained by any person as a result of the use, in any manner or form, of any knowledge, information or data contained in this document, or due to any inaccuracy, omission or error therein contained.
The European Commission shall not in any way be liable or responsible for the use of any such knowledge, information or data, or of the consequences thereof.
This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of it.

Copyright notice

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DL	Deep Learning
DR	Dimensionality reduction
EE	Extreme Event
FFNN	Feedforward Neural Networks
FS	Feature Selection
ML	Machine Learning
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Support Vector Machine

EXECUTIVE SUMMARY

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that leverages the information contained in data to inductively address tasks. In particular, ML algorithms consider a set of data, the training set, to learn patterns that can be exploited to add information or make predictions given new unseen test data. Therefore, ML algorithms learn from observations and optimize a set of parameters to improve their performance as measured by predefined metrics.
ML techniques are more and more frequently applied in a variety of fields (e.g., medicine, biology, Earth science, social media), where they exploit the large amount of data available to inductively extract or interpret complex patterns in the data that are subsequently employed to generalize to new unseen data. Moreover, a variety of ML algorithms exists to address different problems (e.g., clustering to identify subgroups of data, supervised learning to predict a target value) and to consider different types of data (e.g., images, text and sequences, time series).

The CLINT project aims to develop an ML-enhanced framework to address detection, causation, and attribution of climatic Extreme Events (EE). The focus is on tropical cyclones, droughts, heatwaves and warm nights, compound events and concurrent extremes, which are ensembles of critical climatic events. Specifically, WP2 aims to identify and develop ML techniques to handle the large amount of available spatio-temporal climate data. The final purpose is to provide suitable algorithms that perform well on the different applications addressed in CLINT. The methodological workflow followed in WP2 is to analyse the state-of-the-art techniques available in the literature and eventually develop new methods to address specific needs. For this reason, this first deliverable is focused on state-of-the-art analysis, both focusing on methods and applications.

After a brief introduction to Machine Learning and on the EE addressed in CLINT, the document introduces from a methodological perspective the most relevant subfield of ML identified to address detection, causation, and attribution of extremes. Then, the focus is on the analysis of the state-of-the-art ML applications to address each of the three problems. In particular, a different chapter is dedicated to each EE, which has its own peculiarities and is addressed in a different way in the literature.
Finally, methodological challenges and improvement opportunities are discussed to conclude the review.

KEYWORDS

Machine Learning; Feature Selection; Supervised Learning; Deep Learning; Causal Inference; Detection; Causation; Attribution; Extreme events

1 INTRODUCTION

1.1 Machine Learning Overview

Machine Learning (ML) is a branch of computer science and a subfield of Artificial Intelligence (AI), which aims to extract information and relevant patterns from data that can be generalized to gather information on new unseen samples and exploit them to make decisions. ML is composed of three main subfields: supervised learning, unsupervised learning, and reinforcement learning (a detailed analysis of supervised, unsupervised, and reinforcement learning can be found in (Bishop & Nasrabadi, 2006), (Zaki, Meira Jr, & Meira, 2014), (Sutton & Barto, 2018)).

Given a set of input features x_i and related target outputs y_i, the goal of **supervised learning** is to estimate the unknown model that is able to reconstruct the output from the inputs in a way that it is also able to estimate the output of a set of new unseen inputs, called the test set. Supervised learning can be divided into three main subfields:

- **Classification**, where the target is one of *K* discrete classes, and the goal is to assign each input to a class.
- **Regression,** where the target is a continuous variable, and the goal is to learn a mapping from the input that produces a prediction as close as possible to the target.
- **Probability Estimation,** where the goal is to associate to each input a probability distribution over a set of possible events.

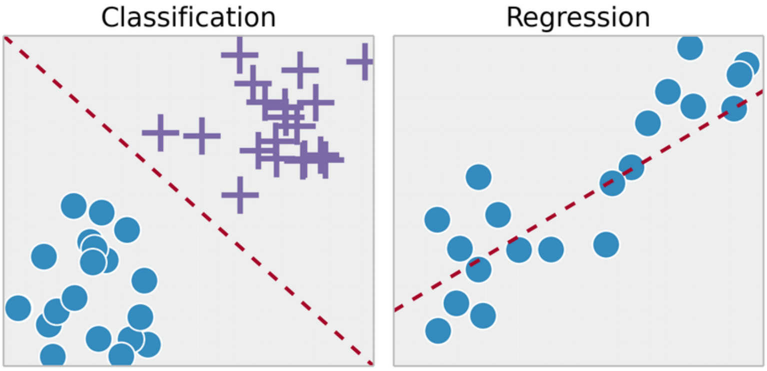

Figure 1-1: an example of classification and regression, the first tries to predict the group a feature belongs to, while the second a real-valued output. (Credits for image (Matanga, 2017))

In the **unsupervised learning** scenario, the only available information is a set of unlabelled input features x_i and the goal is to learn an efficient (e.g., compact) representation of the input. This is usually done in two ways, depending on the information of interest that should be extracted from the data. The first option is represented by **clustering** techniques, whose goal is to group data in order to maximize the similarity of data within the same group and minimize the similarity between different groups. The other standard unsupervised approach is **dimensionality reduction**, which focuses on projecting data into a lower dimensional space, preserving the structure among data while reducing their dimension, which is useful for computational complexity, memory usage, and algorithm performances.

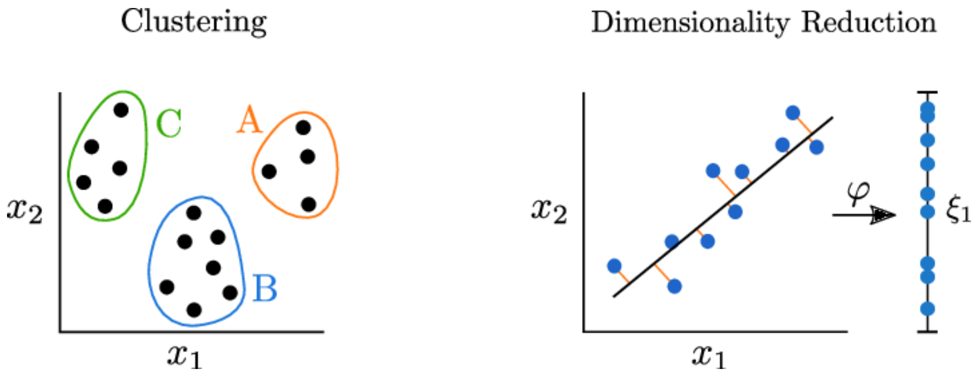

Figure 1-2: an example of unsupervised clustering and dimensionality reduction, which respectively allow to group input data into clusters of similar data or to project them into lower dimensional spaces. (Credits for image (Beck & Kurz, 2020))

Finally, **reinforcement learning** is the ML approach to solve sequential decision-making problems, which aims to find the optimal distribution of actions to take in each possible state in order to maximize the cumulative reward (usually modelling the problem as a Markov Decision Process).

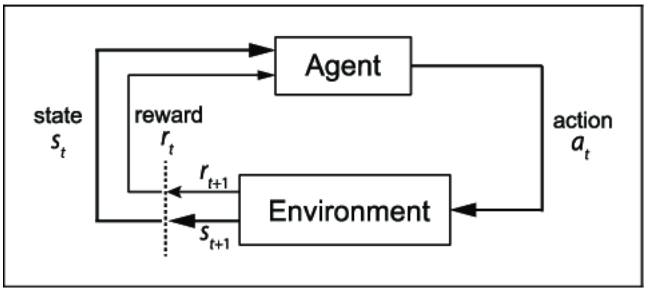

Figure 1-3: scheme of the interaction between agent and environment in a reinforcement learning setting. (Credits for image (Galatzer-Levy, Ruggles, & Chen, 2018))

1.2 Motivation and Outline

The CLINT project aims to address the problems of detection, causation, and attribution of EE, enhancing traditional methods with ML. For this reason, WP2 is focused on providing state-of-the-art techniques, together with the design of novel algorithms, to address these problems. In particular, the objectives of WP2 are to develop:

- advanced identification of relevant features (feature extraction) algorithms for EE detection;
- advanced data-driven causal inference algorithms for EE causation analysis;
- advanced neural computation algorithms for EE attribution;
- advanced data-driven models for EE forecasting;
- advanced spatial predictive models for reconstructing past EE.

As a first step, this document aims to revise the state-of-the-art techniques applied in the Literature for the detection, causation, and attribution of the EEs addressed in CLINT both from a methodological and an applicative perspective. In particular, the ML subfields identified as relevant to the tasks of interest will be presented. Then, the focus will be on revising some recent existing ML applications for each EE under analysis.

The following table offers an overview of the number of applications that are presented in this report for the detection, causation, and attribution of the four categories of EE addressed in CLINT. From the table it is immediately possible to identify that there is a robust literature addressing the detection of droughts and tropical cyclones with ML, with a few works also related to their attribution and causation. Moreover, a few works in the literature address heatwaves (and warm nights) with ML, and only a few recent works try to address compound events and concurrent extremes with ML.

Table 1-1: number of applications analysed in this report for the detection, causation, and attribution with ML of each EE addressed in CLINT.

Number of references				
	Droughts	**Tropical Cyclones**	**Heatwaves**	**Compound Events**
Detection	23	41	6	6
Causation	6	6	3	0
Attribution	6	6	3	0

1.3 Outline

The deliverable is structured into seven chapters:

- **Chapter 1** (current) provides an overview of machine learning methods and introduces the structure of the document.
- **Chapter 2** analyses the main subfields of ML identified as relevant for EE detection, causation and attribution.
- **Chapter 3** presents relevant works and methods representing the state-of-the-art of detection, causation and attribution of droughts with ML.
- **Chapter 4** presents relevant works and methods representing the state-of-the-art of detection, causation and attribution of Tropical Cyclones with ML.
- **Chapter 5** presents relevant works and methods representing the state-of-the-art of detection, causation and attribution of Heatwaves and Warm Nights with ML.
- **Chapter 6** presents relevant works and methods representing the state-of-the-art of detection, causation and attribution of Compound Events and Concurrent Extremes with ML.
- **Chapter 7** concludes the document with final considerations.

2 CHALLENGES AND METHODS: AN ML PERSPECTIVE

2.1 Chapter Overview

In order to enhance standard approaches for the detection, causation and attribution of EE with ML techniques, the first prerequisite is to identify the subfields of ML most related to these three problems and the commonalities and peculiarities of each EE under analysis.

The main peculiarity of the available datasets is their spatio-temporal structure with fine granularity, which means that they are historical series of highly correlated variables, whose correlation diminishes with separation. For this reason, dimensionality reduction techniques can provide an efficient data-driven way to aggregate variables that are close in space and share a large amount of information, reducing the dimension of the dataset so that the performance of the ML models and the space and time complexity required are improved.

Detection aims to identify the drivers of an EE. Therefore, given a set of candidate features, the most straightforward way to identify the most impactful drivers with a data-driven approach is to make use of feature selection techniques. Then, to better investigate if the candidate drivers are meaningful to predict the EE, supervised learning techniques are the general solution that is needed.

Attribution, which aims to identify the impact of anthropogenic climate change on EE, can benefit from feature selection and supervised learning. The former can be adapted to identify whether the most relevant variables are related to human activities. Supervised learning is then applied to train and compare models with data produced by climate models with and without anthropogenic climate trends to conclude whether an observed EE pattern can be attributed to climate change.

Finally, the identification of causal relationships within a set of variables can be addressed with causal inference, in particular with the subfield of causal discovery. Since the data available are usually time series, specific algorithms related to sequences can be adopted.

In the following table, the above-mentioned techniques are summarised, with their impact on detection, causation and attribution of EE. In the following sections of this chapter, each of the mentioned subfields of ML will be introduced, resulting in a methodological overview of the state-of-the-art methods for each subfield identified as significant for the three tasks.

Table 2-1: overview of the identified ML subfields and their importance for EE detection, causation, and attribution.

Detection	
Dimensionality reduction ● Aggregate fine-grained spatial features in a dynamical data-driven way ● No need to manually extract aggregations of variables	**Feature selection** ● Identify relevant variables ● Identify non-redundant variables

Detection	
● Faster computations and less risk of overfitting of ML models	● Faster computations and less risk of overfitting if many non-relevant input features are removed
Supervised learning ● Detect an EE through the most relevant drivers ● Quantify the performance of the prediction to confirm that the identified variables are drivers of the considered EE ● Automatically extract relevant features (with some techniques such as convolutional neural networks)	**Causal Inference** ● See causation
Attribution	
Dimensionality reduction ● Same advantages of detection ● More complex to apply since it is usually necessary to keep separated the variables related to human influence	**Feature selection** ● Identify if anthropogenic driver variables are relevant and non-redundant candidate drivers of the EE under analysis
Supervised learning ● Evaluate the improvement of prediction performance of the models with anthropogenic features to find statistical evidence of their impact on the EE	**Causal Inference** ● See causation
Causation	
Dimensionality reduction ● Still applicable to reduce the dimensionality of features	**Feature selection** ● Applicable as pre-processing step aimed to reduce the set of candidate causal features
Supervised learning ● Applicable with causal discovery techniques	**Causal Inference** ● Classical causal analysis with interventions is not applicable, need to rely on observational causal inference ● Causal discovery identified as the core component, designed to identify the causal link among variables

2.2 Dimensionality Reduction

Dimensionality reduction (DR) is a subfield of ML usually considered as a preprocessing technique, applied to the features before feeding a supervised learning approach with them. DR is particularly useful when high-dimensional data are available, which should be reduced to obtain more manageable features, i.e., less prone to overfitting and curse of dimensionality, two of the most common issues in ML.

Specifically, DR consists in projecting the data into a lower-dimensional space, trying to preserve as much as possible their high-dimensional structure. More rigorously, DR can be seen as a function $\phi: \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times d}$, mapping the original dataset with D features into a reduced dataset with d<<D features, obtained by combining the original features through the transformation ϕ.
The goal of this projection is therefore to reduce the (huge) dimensionality of the original dataset while keeping as much information as possible in the reduced dataset, which is usually done by preserving a distance (e.g., Euclidean, geodesic) or the probability of a point to have the same neighbours after the projection ((Zaki, Meira Jr, & Meira, 2014) contribute a broader introduction with more rigorous computations of classical approaches).

DR is particularly important for Earth sciences ML approaches in general, since climatic spatio-temporal data usually consists of thousands of values of a variable for different locations on Earth (e.g., gridded Sea Surface Temperature data with 0.25 degrees granularity), where each one can be considered as a different feature. In this context, the dimension must be reduced since, otherwise, the number of features can be much larger than the number of samples available (e.g., one sample per day for the last century). DR methods allow to reduce the dimension keeping the information of each original feature through a projection, at the cost of impairing the interpretability of the reduced features. In the next section, Feature Selection will be presented, which has the opposite tradeoff: it preserves interpretability, but it discards many of the original features.

DR approaches can be divided into two macro categories: unsupervised and supervised. The first ones focus on projecting the data, reducing the dimension without considering the subsequent supervised task, but only the preservation of the structure among data. The supervised techniques, on the other hand, aim to obtain a projection that tries to keep the structure among the data and to obtain a reduced set of features that can get the best performance on the prediction of the target. Moreover, it is possible to distinguish between linear and non-linear projections for each of the two types of methods: the first ones are simpler to interpret, while the others can learn a more complex manifold. In the next two subsections, the main approaches of these families are revised.

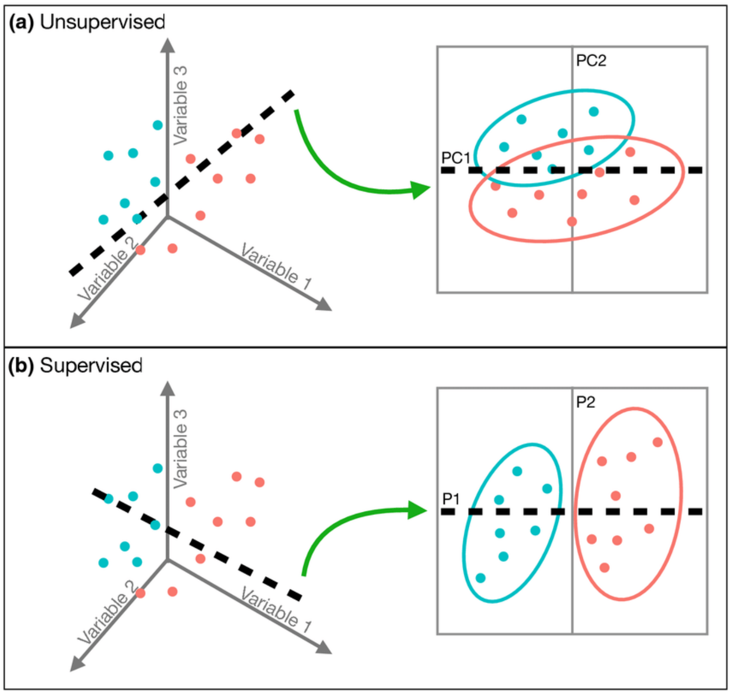

Figure 2-1: unsupervised and supervised dimensionality reduction. The first projects data to preserve their structure as much as possible while the second maximizes the separation between samples of different classes. (Credits for image (Scott & Crone, 2021))

2.2.1 Unsupervised Dimensionality Reduction

Classical dimensionality reduction methods can be considered unsupervised learning techniques that generally do not consider the target, but focus on projecting the dataset through the minimization of a given cost function.

The most popular unsupervised linear dimensionality reduction technique is Principal Components Analysis (PCA) (Pearson, 1901) (Hotelling, 1933), a linear method that embeds the data into a linear subspace of dimension *d*, describing as much as possible the variance in the original dataset. Specifically, PCA is based on projecting the data in orthogonal directions minimizing the mean-squared error, whose solution can be proved to be equivalent to projecting onto the eigenvectors of the covariance matrix of the dataset. The dimensionality reduction is, therefore, performed by selecting the first *d* projections (principal components), which are the projections through the *d* eigenvectors associated with the *d* largest eigenvalues. Moreover, thanks to the orthogonality of these vectors, the amount of variance preserved is equal to the sum of variances of each reduced feature (which is the sum of the corresponding eigenvalues).

One of the main difficulties of applying PCA in real problems is that it performs linear combinations of possibly all the *D* features, usually with different coefficients, hindering the interpretability of each principal component and suffering from the curse of dimensionality. To overcome this issue, some variants have been introduced, like sparse variable PCA (svPCA) (Ulfarsson & Solo, 2011), which regularize the loss through a term that forces most of the weights of the projection to be zero. This mitigates the interpretability issue, but it possibly leads to neglecting many features for the projection, whose informativity would be lost.

A number of variants exist to overcome the different issues of PCA (e.g., out-of-sample generalization, linearity, sensitivity to outliers), to extend its applicability or to approach the problem from a different perspective. Some relevant approaches based on PCA are introduced in the following, while an extensive overview can be found in (Sorzano, Vargas, & Montano, 2014).
Incremental PCA (Artac, Jogan, & Leonardis, 2002) can update the projection weights online and is particularly useful for streaming data. Singular value decomposition (SVD) (Golub & Reinsch, 1971) leads to the same result as PCA from an algebraic perspective through matrix decomposition. It decomposes the data matrix as the product of three matrices and from them, it is possible to compute the principal components faster. It is also possible to perform a non-linear PCA projection (Girolami & Fyfe, 1997), considering a general function rather than a linear projection and still minimizing the least squares error. This procedure allows projecting onto a more complex manifold, with a cost in terms of computational complexity and interpretability.
Another technique related to PCA is Factor Analysis (FA) (Thurstone, 1931), which is a generative approach. It assumes that the features are generated from a smaller set of latent variables, called factors, and tries to identify them by looking at the covariance matrix. In particular, standard FA assumes that each feature is a (linear) combination of the factors with Gaussian noise. Both PCA and Factor Analysis can reduce through rotations the number of features that are combined for each reduced component to improve the interpretability, but their coefficients can still be different and hard to interpret.
Finally, Independent Component Analysis (Hyvarinen, 1999) is an information theory-based approach that looks for independent components (not only uncorrelated as PCA) that are not constrained to be orthogonal. It is also a generative model, whose objective is to find a matrix such that the components are as independent as possible, and each feature is a (linear) combination of these components. This method is more focused on splitting different signals mixed between features rather than reducing their dimensionality, which can be done as a subsequent step with feature selection. This step would be simplified since the new features are independent.
A broader overview of linear dimensionality reduction techniques can be found in (Cunningham & Ghahramani, 2015).

In contrast to the linear nature of PCA, many non-linear approaches exist, following the idea that data can be projected onto non-linear manifolds. It is possible to distinguish between convex techniques, where the solution space is convex and is guaranteed to reach a global optimum, and non-convex techniques. An extensive overview of these two families of methods can be found in (Van Der Maaten, Postma, Van den Herik, & others, 2009) and in the following some of the most relevant approaches are discussed.
The approaches that optimize a convex objective function usually identify the projection coefficients through a generalized eigenproblem. Some of them try to preserve the global similarity of data. Among these, Isomap (Tenenbaum, Silva, & Langford, 2000) preserves pairwise geodesic distance among samples, rather than the Euclidean distance preserved by PCA. Kernel PCA (Shawe-Taylor, Cristianini, & others, 2004) reformulates PCA by performing standard PCA on the kernel matrix, providing a nonlinear mapping of the original dataset. MVU (Weinberger, Sha, & Saul, 2004) is a variation of Kernel PCA that also learns the kernel matrix through the definition of a neighbourhood

graph on the data, similar to Isomap. Then, it maximizes the Euclidean distance as PCA under the constraint that the distances in the neighbourhood graph do not change, preserving the local geometry between samples. Also, Kernel Entropy Component Analysis (Jenssen, 2009) generalizes classical PCA. It is an information-theoretic extension of PCA that, rather than maximizing the variance explained by the reduced features, attempts to maximize their entropy, estimating the distribution probability of the original features with a kernel function. Finally, Diffusion Maps (Lafon & Lee, 2006) considers again graphs that have edges representing the distance between each pair of the samples, rather than only having edges with neighbour samples as done by Isomap. The objective here is to better preserve the distance between the data after the projection.

Other approaches optimize a convex objective function focusing on local similarity of data. LLE (Roweis & Saul, 2000), similarly to Isomap, constructs a graph representation of the samples. However, it only tries to preserve local properties by writing each sample as a combination of its nearest neighbours, assuming locally linear manifolds. Also, Laplacian Eigenmaps (LE) (Belkin & Niyogi, 2001) try to preserve local properties among data minimizing the weighted distance between each sample and its *k* nearest neighbours by exploiting the Laplacian of the graph. Finally, LTSA (Zhang & Zha, 2004) considers the local tangent space similarly to LLE, but it looks for the projection that allows to reconstruct the same local tangent space to a sample that can be found in the original huge dimensional feature space.

Finally, other non-linear methods optimize a non-convex objective function with different purposes. Sammon Mapping (Sammon, 1969) focuses on rescaling the Euclidean distance-based cost function of PCA by assigning weights inversely proportional to the distance, with the purpose of giving equal importance to the preservation of distance between samples that are close or not. Also, more complex structures like neural networks can be adopted for projecting the features: Multilayer Autoencoders (Hinton & Salakhutdinov, 2006), at the end of the encoder phase, provide a lower dimensional feature vector.

Finally, non-convex approaches can be adopted to align mixtures of models, as done in LLC (Teh & Roweis, 2002). This method consists in computing a mixture of linear models on the data and then aggregating them to obtain the final lower dimensional projection.

2.2.2 Supervised Dimensionality Reduction

Supervised dimensionality reduction is a less-known but powerful approach. It is a less general approach, since it directly encodes the optimization of the learning of a target variable. This makes it suitable when the main goal is to perform classification or regression rather than learning an unsupervised data projection into a lower dimensional space. The methods of this subfield are usually based on classical unsupervised dimensionality reduction, adding the regression or classification loss in the optimization phase. In this way, the reduced dataset is the specific projection that balances the maximization of the performance of the considered supervised problem and the preservation of the structure among data. This is usually done in classification settings, minimizing the distance within the same class and maximizing the distance between different classes in the same fashion as Linear Discriminant Analysis (Fisher, 1936). The other possible approach is to directly integrate the loss function for classification or regression. Following the taxonomy presented in (Chao, Luo, & Ding, 2019), where it is possible to find a broader overview of supervised DR methods, these approaches can be divided into PCA-based, NMF-based (mostly linear), and manifold-based (mostly non-linear).

A well-known PCA-based algorithm is Supervised PCA. The most straightforward approach of this kind has been proposed by (Bair, Hastie, Paul, & Tibshirani, 2006). It is a heuristic approach that applies classical PCA only to the subset of features having large correlation with the target. A more advanced approach can be found in (Barshan, Ghodsi, Azimifar, & Jahromi, 2011), where the original dataset is orthogonally projected onto a space where the features are uncorrelated, simultaneously maximizing the dependency between the reduced dataset and the target by exploiting Hilbert–Schmidt independence criterion.

Many variants of Supervised PCA exist, e.g., to make it a non-linear projection or to make it able to handle missing values (Yu, Yu, Tresp, Kriegel, & Wu, 2006).

NMF-based algorithms have better interpretability than PCA-based, but they focus on the non-negativity property of features, which is not a general property of problems and restricts their applicability to a subset of settings. These methods factorize the dataset into two non-negative matrices, with the idea of approximating the data matrix with the product of two non-negative matrices through the maximization of the Frobenius loss function or the KL divergence. Two groups of NMF-based methods exist: Supervised NMF and discriminative NMF. Supervised NMF consists in directly including the loss function in the learning of the two matrices, as in (Jing, Zhang, & Ng, 2012) through the optimization of three terms: the first is related to the best approximation of the original matrix through the Frobenius norm, the second one focuses on minimizing the prediction loss and the third one maximizes the difference of the projected data belonging to different classes. Discriminative NMF, on the other hand, is applied by (Lu, et al., 2016). The main idea is, similarly to LDA, to improve the predictability maximizing the distance between classes while minimizing the distance within the same class, together with minimizing the KL divergence between the original matrix and its approximated decomposition.

Finally, manifold-based methods perform non-linear projections, usually applying a supervised variation of an unsupervised non-linear approach, with higher accuracy on the final supervised predictions with higher computational costs. Starting from unsupervised Isomap, (Ribeiro, Vieira, & Carvalho das Neves, 2008) include the maximization of the dissimilarity between samples of different classes and the minimization of dissimilarity within the same class. Also (Zhang, et al., 2018) propose a supervised version of Isomap, by projecting the data of the same class on the same manifold and trying to maximize the distance between manifolds. In (Zhang S.-q. , 2009) a supervised variation of LLE is considered, focusing on preserving local structures, rather than global ones as Isomap. It is based again on the LDA idea of encoding the maximization of dissimilarity between classes together with the preservation of the capability of LLE to express a sample as a linear combination of its neighbours. Finally, (Raducanu & Dornaika, 2012) proposes a supervised version of LE where again the LDA idea of minimizing the margin between data of different classes and maximizing it for samples of the same class.

2.3 Feature Selection

Feature Selection (FS) is a subfield of ML usually considered a preprocessing technique, that can be applied after or in place of DR, before performing a supervised learning task. FS is complementary to DR, and it is fundamental in the presence of high-dimensional data that should be reduced to prevent overfitting and curse of dimensionality. Exactly as DR, FS consists in reducing the dimension

of the data from a high dimension *D* to a lower dimension *d<<D*, which is done with FS by discarding a set of variables and retaining without transforming a limited number of *d* features, choosing the most relevant and non-redundant ones. In this way, the advantage of FS is that the interpretability of the reduced features is preserved, since they are a subset of the original ones, while the main disadvantage is that a set of features is completely discarded, which can be a loss of information.
More specifically, the first objective of FS is to identify the relevant features, which are the ones that determine the conditional distribution of the target, given the others. Moreover, FS aims to select only the set of non-redundant features. This means that this set is minimal, as no subset would contain the same information on the target. An extensive introduction of the concept of FS and an overview of classical FS approaches can be found in (Guyon & Elisseeff, 2003).

As already discussed for DR, FS is particularly important for applying ML on Earth sciences, which usually deal with spatio-temporal datasets with thousands of features representing values of a variable at different locations on the Earth. Moreover, with many FS approaches it is possible to estimate the amount of information that a certain meteorological variable measured at a certain location has about a target EE. This property can be particularly useful if the final task is the detection of the main drivers, or the attribution of observed pattern to human influence.

FS approaches can be divided into three categories - wrapper, embedded, and filter - depending on the procedure they follow to select features (Chandrashekar2014). In particular, wrapper FS techniques are focused on the prediction performance of a learning model, and they aim to select the subset of features that has the best validation performance. For this reason, these approaches lead to a good performance of the subsequent supervised approach if it is performed with the same learning algorithm, but they have poor generalization capability and they should be run again at each change supervised technique. Moreover, it is impossible to perform an evaluation of each subset in practice since it would require evaluating a combinatorial number of subsets. For this reason, a greedy search is usually applied forward or backwards. This search sequentially adds or removes features, eventually improved by considering more advanced heuristic criteria. Wrapper methods will not be discussed in more detail due to their computational complexity and lack of generalizability, which make them impractical for Earth sciences applications. The interested reader can find a more detailed description of wrapper methods in (Kohavi & John, 1997).

Filter methods, on the other hand, usually select features in a model-free fashion, i.e., without considering a specific supervised learning approach but ranking the features depending on some criterion of relevance and non-redundancy. Therefore, these methods are applied before the learning phase, so, compared to wrapper methods, they are more general approaches that must only be run once. However, they are less optimized for the specific application, which may lead to poorer performance when applied with a specific model. Due to the reduced computational complexity of their application and the clear importance score that they provide, filter methods are often considered the best choice with huge dimensional datasets, and they will be extensively described in the next section.

Finally, embedded FS approaches directly embed the FS phase inside the supervised training phase of a learning algorithm, trying to combine the computational efficiency of filter methods with the specificity of wrappers. They can be performed by adding a regularization term to the training cost function that forces the coefficients related to many features to be zero, so that the learning is only performed focusing on the most relevant features. However, these methods are very specific to some learning algorithms, and they do not clearly identify the general relevant drivers. For this reason, they will not be treated in this deliverable, where the main purpose is to identify techniques which are able to detect the candidate drivers of an EE, and this selection should not depend on the choice of the learning algorithm.

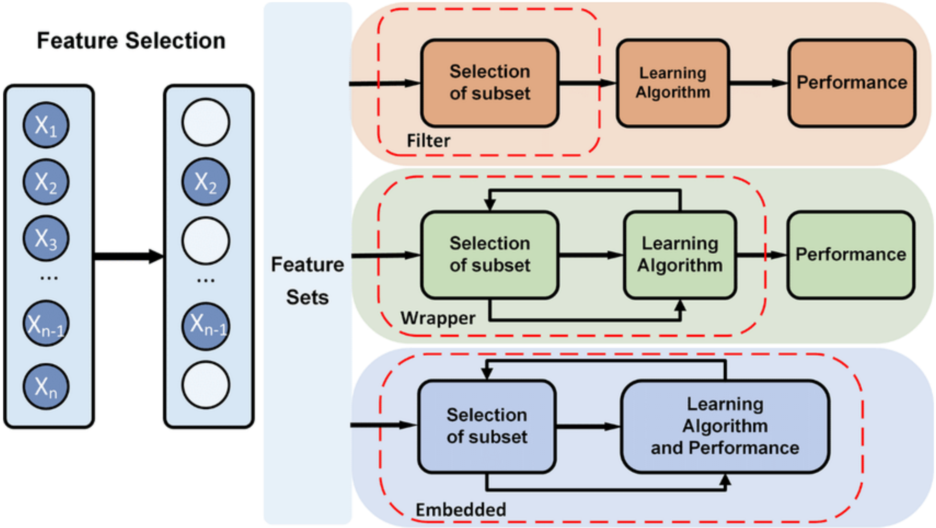

Figure 2-2: filter, wrapper and embedded FS methods. The first ones are performed once before and independently from the learning algorithm, the second ones are performed together with the learning method, while the third ones are part of the learning algorithm itself. (Credits for image (Xie, Li, Zhou, He, & Zhu, 2020))

2.3.1 Filter Feature Selection

In the following, some of the main state-of-the-art filter FS methods will be presented, following the taxonomy by (Li, et al., 2017) and assuming the samples to be independent and identically distributed. In this setting, it is possible to identify four main categories of filter methods: similarity-based, information theoretical based, sparse learning based and statistical based.

Similarity based methods. These methods assess the importance of a feature based on its ability to preserve the overall data similarity. In general, this can be done in an unsupervised fashion, considering a similarity between each pair of samples, or in a supervised way, deriving the similarity from the labels.

Laplacian score (He, Cai, & Niyogi, 2005) is a traditional unsupervised approach that focuses on local neighbours, considering the following similarity function $S(i,j) = e^{-\frac{\|x_i - x_j\|_2^2}{t}}$ if x_i is one of the nearest neighbours of x_j, 0 otherwise. Then, the Laplacian score is defined for each feature as a measure of the feature's locality-preserving power, which consists of maximizing the sum of differences between each pair of features multiplied by their similarity score. In this algorithm, the *d* best features are selected by evaluating all the scores individually, without considering the minimization of the relevance, which may lead to a selection of variables that share a lot of information. SPEC algorithm (Zhao & Liu, 2007) can be considered a variant of the Laplacian score method. It is still based on the idea of giving importance to a feature depending on its ability to preserve local data similarity, but this is done both for supervised and unsupervised settings. It associates a similarity measure between samples which is positive only if they belong to the same class and null otherwise. Also, the score proposed by SPEC is a generalization of the Laplacian score. It introduces hyperparameters in the computations to leverage the relative importance between features.

On the other hand, the Fisher score (Duda, Hart, & others, 2006) is a supervised method that selects features that maximise the similarity among features within the same class and minimise the similarity between samples of different classes. Also, the Fisher score can be seen as a supervised variation of the Laplacian score, which considers all the samples of a class as neighbours. Another supervised approach, ReliefF (Robnik-Šikonja & Kononenko, 2003), focuses on the opposite task of Laplacian-type scores, since it aims to identify the features that can separate samples from different classes, rather than focusing on the ones that maximize the similarity within the same class.

In general, for all the introduced approaches, it is possible to conclude that they are useful to identify the most promising features, they are fast to compute, and they do not rely on any specific learning algorithm. On the other hand, they do not focus on feature redundancy and on the interaction between features, therefore they should be considered together with one of the approaches discussed in the rest of the section.

Information-theoretical based methods. This family of FS methods considers different measures from the field of information theory to evaluate the relevance and non-redundancy of the selected features. Since these algorithms are able to exploit the information of the target, quantify the shared amount of information and evaluate a subset of features altogether, they are among the most promising filter methods, although the estimators of information-theoretical measures are usually difficult to compute and need a large number of samples to be reliable.

The classical idea behind these approaches is to maximize the amount of information shared between a feature and the target, or the amount of additional information on the knowledge of the target that one feature adds with respect to the already selected features, meanwhile minimizing the information shared between the feature under analysis and the already selected ones.

One of the first methods of this category is Mutual Information Maximization (Lewis, 1992), which only focuses on selecting the features that share the maximum amount of information with the target, ignoring the minimization of redundancy. The redundancy is addressed by mutual information feature selection (Peng, Long, & Ding, 2005). It selects the features with large mutual information between feature and target and small mutual information between features. More advanced techniques, such as conditional infomax feature extraction (Lin & Tang, 2006), joint mutual information (Brown, Pocock, Zhao, & Luján, 2012), and conditional mutual information maximization (Fleuret, 2004) introduce conditional mutual information to take into account, together with the maximization of information shared between feature and target, the minimization of the difference between the mutual information between the feature under analysis and the already selected ones and their conditional mutual information given the target. With this term, the information that two selected features share which is not relevant for the target is minimized, minimizing, therefore, the redundancy. A more recent approach (Beraha, Metelli, Papini, Tirinzoni, & Restelli, 2019) proposes forward and backward methods that maximize the conditional mutual information between the feature under analysis and the target, given the already selected ones. In this way, the paper produces theoretical guarantees of the amount of information that is lost in

performing the reduction, proving the intuition that there is a trade-off between the amount of information lost and the desired reduction of the number of features.
In conclusion, there are a lot of different formulations of (conditional) mutual information FS approaches, which are theoretically able to balance high relevance and low redundancy, and whose main disadvantages are the difficulty of the estimation of the considered quantities and the necessity to include the target in the analysis, and therefore without the possibility to perform the methods in an unsupervised fashion.

Sparse learning-based methods. This family of filter methods is similar to embedding FS. Indeed, these methods aim to regularize the learning of a supervised algorithm by imposing the coefficients associated with many variables to be null. In particular, they usually add a regularization term on the loss of a supervised learning technique, which is usually the l_p norm or the $l_{p,q}$ norm of the coefficients. The filter approach of these techniques is the addition of this term to any supervised or clustering algorithm, without the need that the method has been originally designed with an embedded feature selection. Different methods are available in the literature (Hara & Maehara, 2017), (Peng & Fan, 2017), (Nie, Huang, Cai, & Ding, 2010) but they all suffer from a lack of generalization and computational costs, therefore they will not be discussed further in this review.

Statistical based methods. This category of FS approaches can usually be seen as an initial filtering technique that is applied before other more advanced techniques among the ones discussed before. In particular, they compute some statistical quantities to identify the variables that should not be considered. For example, the most straightforward statistical approach is to compute the variance of each feature and remove the ones that have small variability, assuming that these would not be able to discriminate between samples of different classes. Also, independence tests (e.g., Chi-Square score (Liu & Setiono, 1995)) can be exploited to evaluate if a feature is independent of the target and subsequently remove it from the candidate's relevant features.

2.4 Supervised Learning

Supervised learning is the final step to address the detection, attribution and forecast of an EE, where there is a set of variables, possibly aggregated through DR or filtered with FS, and an ML model is designed to produce an output which predicts one or more characteristics of the EE.
The most straightforward supervised learning approach is linear regression (Montgomery, Peck, & Vining, 2021). It linearly combines the input variables and identifies the set of coefficients that minimize the mean squared error between the predicted values and the observed ones. Then, the performance of the model is evaluated through its ability to predict the target given new samples as input. The main limitation of this approach is the strong assumption of linearity, which can be relaxed considering the non-linear transformation of the input variables as features. This, however, should be done manually and is still limited. For this reason, many non-linear supervised learning approaches exist and they enlarge the hypothesis space by considering more complex possible models. Among these, Decision Trees (Quinlan, 1986) use a tree-like model to identify groups of samples with similar input features and similar output, Support Vector Machines (Cortes & Vapnik, 1995) try to identify boundaries in the feature space that maximally separate different samples, K-nearest Neighbors (Fix & Hodges, 1989) focus on the target value of the most similar samples to predict the value of the label of the sample under analysis.

More advanced techniques are based on an ensemble of models (Sagi & Rokach, 2018), which can be useful to reduce the bias (boosting) or the variance (bagging) that a single model may suffer from. To reduce the variance of a supervised learning method, bootstrap technique (Breiman, Bagging predictors, 1996) can be exploited to produce different datasets from a unique one through sampling with replacement. Then, a model is trained on each of them, and the final output is a balance of the outputs (e.g., the mean or the median of the predictions). Random Forest algorithm (Breiman, Random forests, 2001) is one of the most famous boosting methods. Random Forest considers an ensemble of decision trees, eventually training each with a different dataset obtained through bootstrap. To further reduce the correlation between different trees, which results in a reduction of variance, each decision node of the tree is based only on a subset of features, and this subset randomly varies across different trees. Extremely Randomized Trees (Geurts, Ernst, & Wehenkel, 2006) is a variation of Random Forests that additionally can perform the split on a node randomly, further reducing the variance. Boosting techniques, on the other hand, are focused on reducing the possible bias of a single model. The most famous approach of this kind is AdaBoost (Freund & Schapire, 1997), which aims to reduce the bias through the assignment of different weights to the samples, giving more importance to the samples that have been misclassified previously. Gradient Boosting (Breiman, Arcing the edge, 1997), (Friedman, 2002), on the other hand, is based on the idea of learning a cascade of predictors, where each of them is learning the residual of the previous one. An improvement of Gradient Boosting is eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016), a software library that encodes a version of Gradient Boosting with many technical modifications that help to regularize its behaviour and improve the predictive performances.
For a broader overview of classical supervised learning techniques see (Bishop & Nasrabadi, 2006).

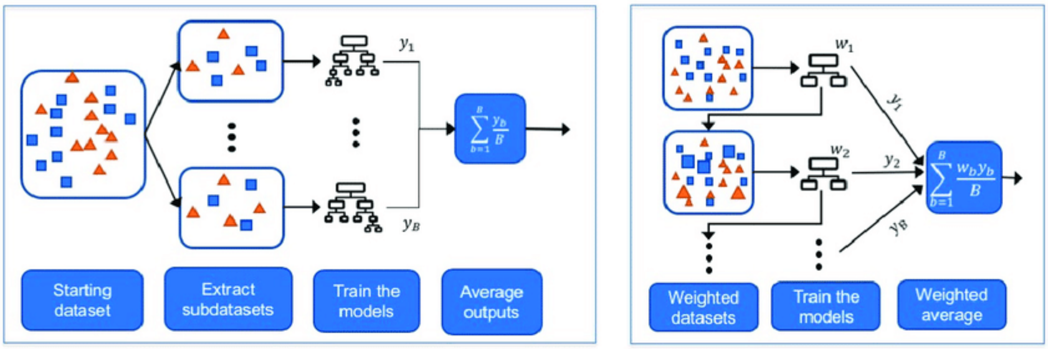

Figure 2-3: supervised learning with model ensembles. Bagging considers different weak learners focusing on the reduction of variance, boosting sequentially focuses on misclassified samples to reduce bias. (Credits for image (Seccia, et al., 2021))

2.4.1 Deep Learning

Meteorological quantities that are usually considered candidate drivers for the occurrence of an EE are spatially and temporally distributed. For this reason, the need for more complex models able to take into account the sequential nature of data and their proximity arises.

Artificial Neural Networks (ANN) (Pouyanfar, et al., 2018) are state-of-the-art techniques applied in many fields to address structured problems, such as text mining or image classification, and they are able to extract features in a fully data-driven way at the cost of performing complex combinations of features that make it difficult to understand the importance of the input features on the prediction and therefore to interpret the results and the extracted features. The main idea of ANNs is to sequentially apply basic non-linear transformations that altogether form a complex network able to represent complex non-linear functions from the input features to the target prediction.

Feedforward Neural Networks (FFNN) (see (Schmidhuber, 2015) for a detailed overview) are the simplest form of ANNs, which are designed for unstructured data. The input data are fed to the first layer nodes and the information is propagated forwardly until the nodes of the output layer are reached. Then, the difference in the prediction with respect to the observed value of the target is computed, and the weights of each node are updated following the classical backpropagation phase. One of the most important parameters to tune, together with the number of nodes and layers, is the activation function, which is the one that makes non-linear the transformations at each node of each layer.

Another category of ANNs is Recurrent Neural Networks (Rumelhart, Hinton, & Williams, 1986), which are specifically designed for sequential data and time series. The main idea behind RNNs is that the information associated with some nodes is propagated to the next iteration of the network, enhancing the network with a memory signal from past samples to predict the present value of the target. A more advanced RNN-based technique is Long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997), which is particularly focused on propagating the memory signal in order to keep important information from the past for many subsequent iterations. Another advanced approach is gated recurrent units (GRU) (Cho, Van Merriënboer, Bahdanau, & Bengio, 2014). It is a variation of LSTM with fewer hyperparameters and some changes in the structure of the network, that make it faster and reduce the memory cost.

Convolutional Neural Networks (CNN) (LeCun, Bottou, Bengio, & Haffner, 1998), on the other hand, are ANNs specifically designed to take into account the spatial dependencies of data. In particular, they are usually applied to images and have also been extensively applied to spatially distributed climate data, as will be discussed in the next chapters. The main idea of CNNs is to perform local aggregations of variables that are closed in space and to propagate this information through many layers, in order to finally extract a set of relevant features that can be exploited to perform the final prediction.

More recently, generative deep learning approaches have also been explored. One of the most applied categories of these methods are generative adversarial networks (GANs) (Goodfellow, et al., 2020). The basic concept of GANs is that, given a training set, the model is able to generate new data with the same distribution as the original set. The model obtains good performance when another model, called *discriminator*, is not able to distinguish between a real sample and a sample generated by the model.

In a standard application of ANNs, the output layer consists of a single output node that encodes the prediction of a single scalar target. More advanced techniques allow for multiple outputs, and they will be discussed in the next subsection.

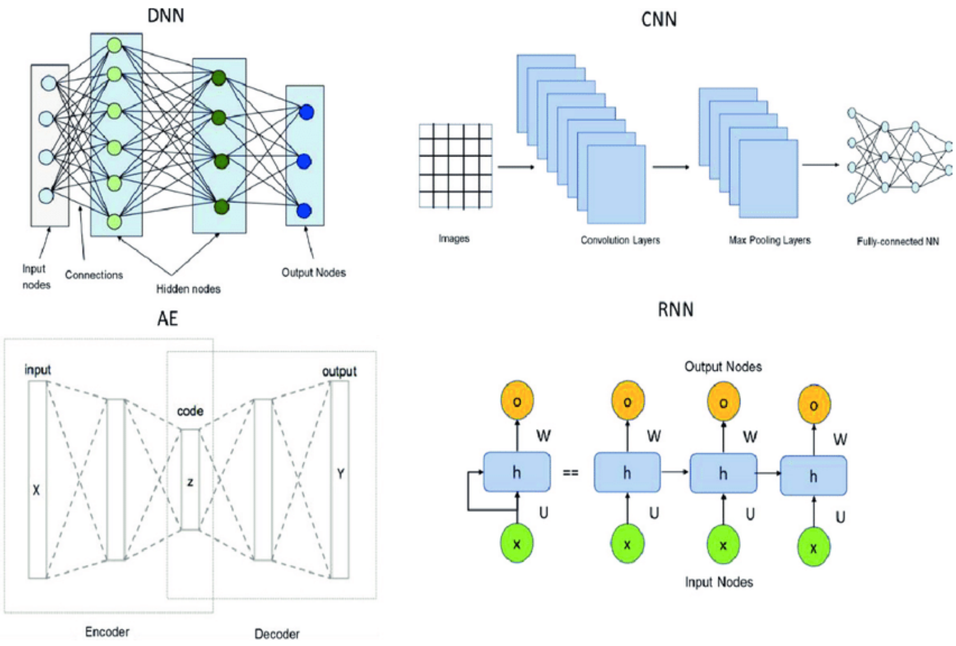

Figure 2-4: four common neural network structures. Feedforward deep NN connect layers and nodes in a forward way, convolutional NN extract features considering spatial adjacencies, autoencoders reconstruct the input after encoding it into a lower dimensional space and can be used for dimensionality reduction, recurrent NN take into account the sequential nature of data with the propagation of a hidden state. (Credits for image (Lo, Gui, Honda, & Davis, 2019))

2.4.2 Multi-Task Learning

Traditional ML methods are composed of a model which is optimized through a training dataset to predict a single scalar value. In some situations, it would be advantageous to consider multiple prediction tasks within the same model, so that the learning algorithm does not specialize on a single task. This reduces the risk of overfitting and makes the model more robust when new unseen data are considered. This intuition is reminiscent of the idea that a human's knowledge of different related tasks is simultaneously exploited to take decisions and predict future values for each of them. This ML approach is called Multi-Task Learning (MTL) (Caruana, 1997). The main concept of MTL is the sharing of information between tasks that can usually be achieved in two ways. One is *feature sharing,* which consists in identifying the set of features simultaneously relevant to each of the considered tasks. The second is *parameter sharing,* that is the sharing of (some) model parameters between different tasks. In the following, the main concepts related to these two approaches are discussed, the interested reader may refer to (Zhang & Yang, 2021) for an extensive review of MTL with more technical details.

In feature-based MTL, the main objective is to identify a powerful feature representation that can boost the overall performance of each individual task. The simplest method of this kind is to identify a subset of relevant shared features in a pre-processing step, through feature selection approaches that consider as target the vector of all the considered tasks.
Another way of performing feature based MTL can be through feature extraction. An intuitive approach of this kind is an ANN that takes as input a set of shared selected features and performs some transformations in the first layers, for example convolutional transformations to images,

obtaining a set of extracted features. Then, some task-specific layers produce the output of each task, which may be boosted by the shared extracted features.

On the other hand, parameter-based MTL is based on the idea of sharing model parameters during the training phase to tune parameters capable of generalizing among different tasks, reducing overfitting. Task clustering approach (Thrun & O'Sullivan, 1996) is one of the most common parameter-based MTL approaches. It relies on the idea of identifying clusters of very similar tasks and then tuning a model, for example, an FFNN, with common nodes in the first layers and nodes in the final layers specific to each cluster.

In conclusion, MTL can be seen as a generalization of classical single task supervised learning, which is particularly useful for applications that share a lot of information, such as the evaluation of an EE for adjacent regions, or when the target are different EEs in the same region. Therefore, it is a particularly relevant option for applying supervised learning methods in this project.

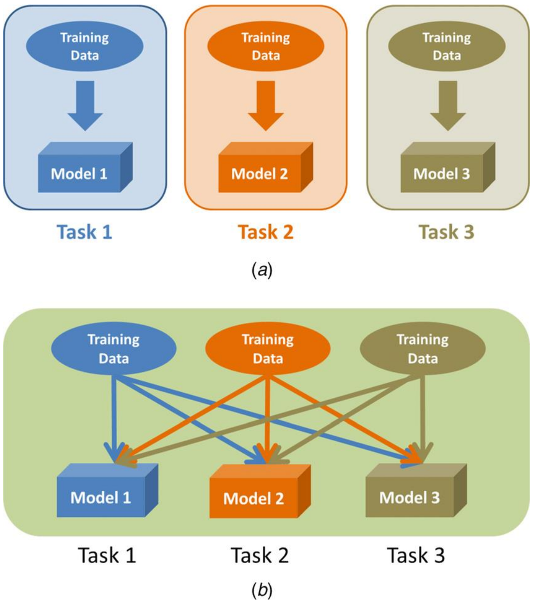

Figure 2-5: difference between single task and multi-task supervised learning. The first independently considers data to learn specific models, the latter shares features and models to leverage the information sharing. (Credits for image (Shao, Ren, Wang, Jin, & Hu, 2016))

2.5 Causality

Classical ML methods focus on the amount of information shared between the input features and the target, minimizing a certain expected prediction loss. This setting is designed to identify the function of the features, among the ones in the hypothesis space of the model, which leads to the most accurate prediction on new unseen data. In general, these methods do not inspect the causal relationship between the feature and the target, with the risk of considering some features that simply have spurious correlations with the target as relevant. Inspecting the cause-effect relationship between variables is particularly important for Earth applications, since there are complex interconnected physical processes that lead to the occurrence of an Extreme Event. The identification of some driver as the real cause of an event would be an important added value to the simple correlation. Moreover, the prediction of an ML model based only on features that are indeed the physical cause of the event may lead to a more robust prediction. For these reasons, causal inference is an important tool to inspect the relationship between features and target and it will be introduced in the next subsection. Moreover, since the data usually considered for Earth science applications are observations and it is not possible to perform controlled experiments, an entire subsection will be devoted to causal discovery for observational data, which is the subfield of causal inference that tries to discover, with just the data analysis, the causal relationship between variables.

2.5.1 Causal Inference

Causal Inference is based on the idea that *correlation does not imply causation*. The observed association between two variables (e.g., correlation) usually results from the combination of two components: a causal association and a confounding association (due, for example, to a common causal variable, which has not been observed, called latent confounder).

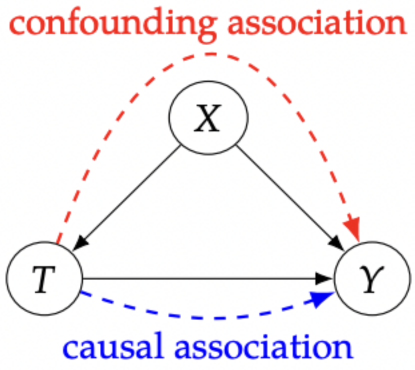

Figure 2-6: total association can be both due to a causal association, to a confounding association or to a mixture between causal and confounding relationships.

In causal studies, the most common causal quantity is the *individual treatment effect (ITE)*, defined as $Y_i|do(T=1) - Y_i|do(T=0) = Y_i(1) - Y_i(0)$. The *ITE* represents the difference between the values that a variable of interest Y_i, associated with the individual *i,* assumes whether a binary treatment *T* is performed on the specific individual ($do(T=1)$) or not ($do(T=0)$). As an example, *T* may represent the occurrence of a Tropical Cyclone and *Y* the amount of precipitation in a region. In this case it may be interesting to detect if, under the same conditions and on the same region *i*, having a TC or not changes the value Y_i of the amount of precipitation.

More in general, given a population of individuals, the quantity of interest becomes the *average treatment effect (ATE)*, defined as $E[Y(1)] - E[Y(0)]$. The ATE represents the expected value of the effect of *T* over the entire population under analysis. Recalling the example above, this may represent the expected value of the effect over different regions on precipitations due to the presence of a TC. This quantity is, however, different from the difference of the expected value of the variable *Y* given the observation of *T*, $E[Y|T=1] - E[Y|T=0]$, since the two observed populations may not be equally distributed. Indeed, in the example introduced above, the subpopulation with *T=0* may contain different heterogeneous regions than the one with *T=1*. Moreover, the correct causal quantity of the ATE can not be directly computed from samples, since for each individual *i,* only *T=0* or *T=1* can be observed. This is the *fundamental problem of causal inference.*

In controlled experiments interventions are possible and the value of *T* can be prescribed. Therefore, the ATE can be estimated with Randomized Control Trial (RCT) technique. It consists in randomly splitting a population into two groups: in the first one it is prescribed *T=1* and in the second one *T=0*. In this way, if the population is sufficiently large, the distribution of the two groups is the same with high probability. On the other hand, when no experiments are possible and only observational data are available, confounders should be controlled to infer the causal relationships. This is usually the case with Earth science applications, since it is not possible to intervene to make a climatic event happen. In this setting, the classical workflow consists in identifying the causal quantity of interest (e.g., the ATE), then identifying a statistical quantity that is equal to the causal quantity of interest, and finally estimating this statistical quantity from data. The easiest way to identify the ATE with a statistical quantity is to evaluate the conditional expected value with respect to all the possible values of the remaining features *W* and assuming that there are no latent confounders. In this way it holds that: $E[Y(1) - Y(0)] = E_w\big[E[Y|T = 1, w] - E[Y|T = 0, w]\big]$ and it is finally possible to estimate the right-hand side of the equivalence with data.

Many subfields of causal inference exist that explore different facets of the identification and identification steps, such as potential outcomes, causal models, unobserved confounders, instrumental variables, causal discovery from observations or interventions, transfer learning and counterfactuals. In (Pearl & others, 2000), (Peters, Janzing, & Schölkopf, 2017) it is possible to find an extensive overview of causal inference. An important aspect of these approaches is that the causal unidirectional link between feature and target is assumed to exist, and the problem is only to quantify the strength of this link. This assumption is not realistic in EE applications, where the knowledge of the causal relationship and its direction is usually the objective of the causal analysis. For this reason, in the following subsection the focus will be redirected to causal discovery for observational data, which aims to identify the causal structure between observational variables.

2.5.2 Causal Discovery

Classical causal discovery (Eberhardt, 2017), (Glymour, Zhang, & Spirtes, 2019) aims to identify a causal graph among a set of variables, considering four main assumptions: Markov assumption, faithfulness, causal sufficiency and acyclicity. According to the first two assumptions, two variables are independent on a causal graph, conditioned on a set of other variables, if and only if their distributions are conditional independent. Causal sufficiency assumes that there are no unobserved confounders and acyclicity does not allow the presence of cycles in the graph.

The PC (Peter and Clark) algorithm (Spirtes P. , et al., 2000) is the first and most famous algorithm designed to identify the causal graph from observations with the four mentioned assumptions. Starting from a complete undirected graph, it exploits (conditional) independence tests to remove connections (edges) between variables that are not causally related, and it considers some orientation rules to identify as much causal directions as possible. In particular, the outcome of the algorithm is a partially directed graph (PDG), where only some edges are oriented. The set of all the directed acyclic graph that are in accordance with the identified PDG form the Markov equivalence class. This equivalence class is the outcome of the PC algorithm, i.e., the output is a set of possible causal graphs that contains the correct causal graph with high probability.

Many variations of the PC algorithm exist. These non-parametric algorithms aim to add more orientation rules to reduce the number of undirected edges or speed up the computations. Other non-parametric techniques consider the relaxation of the four assumptions of PC algorithm, such as FCI algorithm (Spirtes P. , Glymour, Scheines, & Heckerman, 2000), which allows the presence of unobserved latent confounders, or the CCD (Richardson, 2013) algorithm, that removes the acyclicity assumption. Moreover, some model-based techniques exist that improve the algorithms' orientation rules by assuming structural causal models (SCM). In particular, a well-known result states that if each variable can be expressed as a linear combination of its causes with an additive Gaussian noise, then the PC algorithm is complete, and it is not possible to identify the direction of the undirected edges. On the contrary, if the SCM is linear with additive non-Gaussian noise or it is non-linear, then the acyclic directed causal graph can be identified under some technical assumptions.

In order to apply causal discovery to Earth Sciences, a lot of challenges arise with respect to classical approaches. The most important differences are related to the spatio-temporal nature of data, which involves autocorrelation, time lags, and spatial scales. Moreover, in a complex physical system, the assumption of causal sufficiency is a rather strict requirement and approaches that allow the presence of latent confounders are more realistic. A complete overview of the challenges of causality for Earth science applications can be found in (Runge, et al., 2019). Most of the approaches adopted in this setting are causal discovery methods for time series after the extraction and aggregation of the variables of interest so that the spatial structure is no longer considered during the causal discovery phase.

The classical causal discovery approach for time series is Granger Causality, which defines causality as the impact that the historical series of a variable has on the prediction of another one. Although this approach is intuitive and can be easily applied in different contexts, it may suffer from the curse of dimensionality, and it may not correctly evaluate the impact of a variable when many variables that may also be autocorrelated, are considered. For this reason, the PCMCI algorithm (Runge, Nowack, Kretschmer, Flaxman, & Sejdinovic, 2019) evaluates the impact of a variable on another considering only a subset of candidate causes. This leads to a better control of effects due to the high dimensionality and autocorrelation. An improved version to identify causal effects for time series with Granger causality is given by the GResPC algorithm (Moneta, Chlaß, Entner, & Hoyer, 2011). This algorithm also considers contemporary causal effects between variables that may arise when the time lag of the effect is smaller than the considered time steps. This approach applies classical Granger causality to past data to identify their Granger-causal impact on the future and then it applies the PC algorithm to the prediction residuals of the present, to identify contemporary causal relationships. As already discussed for classical Granger causality and PCMCI, the PCMCI+ algorithm (Gerhardus & Runge, 2020) is an extension of PCMCI algorithm that allows for contemporary causal relations that is more able than GResPC algorithm to handle high dimensional data and autocorrelations. Finally, the SVAR-FCI algorithm is a time-series variation of the FCI algorithm that allows both for contemporaneous causal relations and unobserved latent confounders, which further extends the GRresPC algorithm to latent confounders. Similarly, the LPCMCI algorithm (Runge, 2020) performs the same extension to the PCMCI+ algorithm.

Other definitions of causality, or measures to define the causal relationships between variables, have been introduced in the literature. From information theory, transfer entropy (Schreiber, 2000) and directed information (Massey & others, 1990) are two asymmetric quantities designed to quantify the amount of causal information from one series to another. In particular, transfer entropy quantifies the amount of information that the last *l* value of a candidate causal variable shares with a target variable, conditioning on the last *l* values of the target itself. Similarly, the directed information quantifies the amount of information shared between all the historical values of a candidate causal variable and the actual value of the target, conditioning on all the historical values of the target itself. Moreover, the final directed information at time *N* is the sum of the amount of this conditional mutual information at each timestep. These two quantities are clearly related, as shown in (Liu & Aviyente, 2012) and they can be considered as causal counterparts of the classical quantities of conditional entropy and conditional mutual information.

Another way to address causality is through invariance. Indeed, considering, for example, a law from physics, invariance should hold in different heterogeneous environments, while spurious correlations vary with context. Therefore, the causal idea behind invariance is that if a variable causes the target, it is invariant when the distribution changes. In (Arjovsky, Bottou, Gulrajani, & Lopez-Paz, 2019) this idea is exploited by performing supervised learning over different training heterogeneous environments, guaranteeing optimality also for unseen environments, if the SCM between the invariant features and the target does not change. The proposed algorithm consists of a projection onto an invariant space and, subsequently, a unique supervised learning optimization that is valid in each environment. Another way to address causality as invariance is proposed by (Bühlmann, 2020). Here, all the invariant subsets of features are first considered. Their intersection is then a conservative set of invariant causal features, since it is proved that the set of causal variables must be invariant, while the opposite case may not hold.

Finally, recent approaches exist to better characterize problems with latent confounders (e.g., (Chen, Cai, Zhang, & Hao, 2021)) or to integrate causality in advanced machine learning techniques (e.g., (Varando, Fernández-Torres, & Camps-Valls, 2021)).

3 MACHINE LEARNING FOR DROUGHTS

3.1 Overview

Droughts are extreme events consisting of prolonged periods of water supply deficit (Pedro-Monzonís, Solera, Ferrer, Estrela, & Paredes-Arquiola, 2015). The identification of the main drivers of these extreme events may be particularly difficult in highly regulated water systems where natural and anthropogenic dynamics and interventions coexist (Zaniolo, Giuliani, & Castelletti, 2019).

Droughts are usually identified through traditional meteorological, agricultural and hydrological indices (Spinoni, et al., 2019) (e.g., SPI, Standardized Precipitation Index (McKee, Doesken, Kleist, & others, 1993); SPEI, Standardized Precipitation and Evapotranspiration Index (Vicente-Serrano, Beguería, & López-Moreno, 2010)). However, they may fail in the detection of the impact of a drought event, especially with highly regulated water systems, where the effects of meteorological droughts may be mitigated for months by efficient management of water systems without incurring in water shortage (Zaniolo M. , Giuliani, Castelletti, & Pulido-Velazquez, 2018). For this reason, in recent years, several studies have suggested that satellite data can be a valid alternative to classical indices identifying droughts (e.g., NDVI index (Tucker, 1979)).

Machine Learning approaches are a valid alternative to traditional drought indices for the design of data-driven drought indices. Indeed, feature selection and supervised learning approaches may lead to the identification of the main meteorological drivers (e.g., temperature, precipitation, lake water levels) and they may provide a model that, combining their values, is able to detect the presence and/or the intensity of a drought event. Moreover, causal inference methods can provide a causal interpretation of the identified drivers and attribution approaches can give information about the anthropological influence.

3.2 Detection and Forecast

Several approaches based on Random Forests (RF) have been performed in the literature for drought detection and forecast. The main advantages of these model ensemble techniques are the reduction of variance and the embedded feature importance that they perform during the training phase.

In particular, in (Rhee & Im, 2017) RF and Extremely Randomized Trees (ERT), a variation of RF with more randomness aimed to further reduce variance, are applied using SPI and SPEI indices of Korean regions as target and using precipitation, land surface temperature, NDVI and air temperature at different temporal aggregation and scale as input variables.

In (Park, Im, Jang, & Rhee, 2016), sixteen variables from MODIS and TRMM satellite sensors are used to detect drought conditions in different regions of the USA, considering SPI index and crop yields data as targets for meteorological and agricultural droughts, respectively. In this study, RF and two variants (boosted regression trees and Cubist tool) are applied, and RF obtains the best validation performance. The focus of this work is also on the detection of the main predictors identified by the RF approach, which can be the main drivers of the drought event. In particular, land surface-related variables (e.g., Land Surface Temperature and Evapotranspiration) are identified as the main drivers for short-term meteorological drought, while vegetation-related variables (e.g., Normalized Difference Vegetation Index (NDVI)) are the most important variables identified by RF for long-term meteorological droughts.

A study on agricultural droughts hazard in the south-east region of Queensland, Australia, can be found in (Rahmati, et al., 2020). In this work, RF are applied together with other tree-based supervised learning approaches: classification and regression trees and boosted regression trees. Moreover, the results are shown in comparison with other supervised learning methods: multivariate adaptive regression splines and support vector machines (SVM). The RF model yields the best performance. Specifically, the relative departure of soil moisture (RDSM) is considered as target and eight environmental factors are considered as independent variables: elevation, slope, topographic wetness index (TWI), soil depth, clay content, sand content, plant available water holding capacity (PAWC), and mean annual precipitation. Considering the feature importance provided by the RF model, PAWC, mean annual precipitation, and clay fraction are the three main drivers of droughts identified by this approach.

Another application of RF in South-Eastern Australia is (Feng, Wang, Li Liu, & Yu, 2019), which tries to determine if satellite data related to drought factors can be applied to detect agricultural droughts. In this application, the SPEI index is again considered as a target, while thirty remotely sensed drought factors are considered as features (e.g., precipitation and the already mentioned NDVI index). They are obtained from satellite data from MODIS and TRMM. RF, SVM and Artificial Neural Networks (ANN) are considered as regression models, but RF again achieves the best test performance. Moreover, the VSURF method has been applied as a feature selection approach to identify the candidate drivers before the application of the supervised models. From the feature importance calculated by the RF method, the three-month average precipitation is detected as the most important feature.

In (Hobeichi, Abramowitz, Evans, & Ukkola, 2022), RF are used to construct a new drought indicator connecting historical droughts' impacts with a number of drought-related climate features. The new index is shown to have better detection capabilities than commonly used drought indicators. Besides, the RF-based index is fully automated, provides information at the grid scale, and can be used for forecasting droughts with up to 3-month lead time.

In close connection with droughts, (Sutanto, van der Weert, Wanders, Blauhut, & Van Lanen, 2019) applies RF to predict not only the meteorological and agricultural drought condition, but the impact of droughts, exploiting SPI and SPEI indices as features and some specific impact indices as a target.

Another category of model ensemble techniques applied in the literature are boosting methods (e.g., XGBoost), which focus on the reduction of bias, iteratively learning models by giving different weights to each sample, inversely proportional to the prediction error that they produce.

In (Zhang R. , Chen, Xu, & Ou, 2019), a classification task is performed in Shaanxi province, China. Three levels of drought conditions (moderate, severe, and extreme) are classified depending on the SPEI index. ANN and XGBoost are considered as learning methods. Meteorological data (pressure, temperature, humidity, wind speed, precipitation, sunshine duration) and climate indices related to ocean oscillations are considered as features. A feature selection based on the VIF score is first applied, and then XGBoost is shown to achieve the best performance for this task.

In (Mokhtar, et al., 2021) the estimation of the SPEI index for the Tibetan Plateau (China) is addressed. Climate variables are used for the prediction: precipitation amount, temperature, solar radiation, sunshine hours, wind speed and relative humidity. Two model ensembles (RF, XGBoost) and two methods based on ANN (Convolutional Neural Networks and Long-term short memory) have been applied and XGBoost is shown to obtain the best performance for the prediction considering all input variables, although all four models show interesting results depending on the choice of subsets of input variables.

In (Zhang, Abu Salem, Hayes, & Tadesse, 2020), XGBoost is used to predict multi-category drought impacts based on Standardized Precipitation Indexes with different time aggregations. The analysis includes the application of the Synthetic Minority Oversampling Technique (SMOTE) and Random Undersampling on the training datasets in order to balance the class distribution.

Approaches based on Support Vector Machines (SVM) are also often adopted in the literature, especially in classification settings. These methods are particularly effective in high-dimensional spaces, since they can also be trained when the number of features is smaller than the number of samples.

SVM, Artificial Neural Networks and K-Nearest Neighbors classification methods are used by (Khan, et al., 2020) to detect three drought conditions in different cropping seasons for Pakistan: moderate, severe, and extreme. These classes have been identified from SPEI index and the candidate features used are reanalysis data. Specifically, after the application of Recursive Feature Elimination Feature Selection, three drivers have been detected: relative humidity, temperature and wind speed. Validation results identify SVM as the best-performing method in this application.

In (Roodposhti, Safarrad, & Shahabi, 2017), data from both synoptic stations and satellite data (MODIS) are combined to derive a drought sensitivity map (DSM) for vegetation of the Iranian province of Kermanshah. In particular, monthly precipitations from 13 different Iranian Meteorological Organization stations are considered to extract the SPEI index, together with the EVI index, which is used as a satellite-derived drought index. These two indicators of the soil conditions are considered as the input variables of two One-Class SVM, which produce a drought sensitivity map based on the classification of pixels into 5 categories.

In recent years, Artificial Neural Networks (ANN) and Deep Learning in general have also been successfully applied for the detection and forecasting of droughts. These methods, with respect to classical approaches, have the advantage of having a larger hypothesis space, therefore they can learn more complex patterns between data, and they usually do not require preprocessing steps like feature extraction.

In (Belayneh & Adamowski, 2013) the average SPI index evaluated over one and three months is used to predict its next values through ANN and SVM for regression (SVR). The region considered for this study is the Awash River Basin, Ethiopia, and the ANN obtains the best regression performance both in terms of MSE and R-squared. An extension of these results can be found in (Belayneh, Adamowski, & Khalil, 2016), where ANN and SVR approaches are combined with a preprocessing approach based on wavelet analysis. The test results indicate that the coupled model with wavelet analysis and neural network model is the best performing. Furthermore, (Belayneh A. , Adamowski, Khalil, & Ozga-Zielinski, 2014) address the same autoregression problem in the long-term (six and twelve months) and again the ANN approach enhanced by the preprocessing approach based on wavelet analysis is the best performing method, outperforming also classical ARIMA models for time-series prediction. Finally, (Belayneh A. , Adamowski, Khalil, & Quilty, 2016) consider both short- and long-term predictions of the SPI index with ANN and SVR, enhanced by wavelet analysis-based preprocessing as before. Moreover, Bootstrap and Boosting are applied both on ANN and on SVR in order to reduce the variance or the bias of the models, respectively. The best-performing method for this problem is the wavelet boosting ANN.

In (Felsche & Ludwig, 2021), different ANN models are applied to predict drought occurrence in Lisbon and Munich with a one-month lead-time based on SPI. The most influential predictors are extracted among a set of several local atmospheric and land-related variables, together with large-scale climate indices, by using linear correlation analysis. The results showed that the most effective model is an ANN with many hidden layers, with the addition of dropout layers to avoid overfitting. The SHapley Additive ExPlanations (SHAP) algorithm has been used to investigate the contribution of each input variable to the overall predictive power of the ANN, revealing the influence of large-scale climate variables in predicting droughts.

A different method based on neural networks has been applied by (Deo & Şahin, 2015). Specifically, in this work, Extreme Learning Machines (ELM) are applied for the prediction of the Effective Drought Index (EDI) in eastern Australia. ELMs are ANN with randomly chosen weights. They are extremely fast since they do not perform backpropagation, and they do not update the weights of each node, but they only compute a generalized inverse to update the coefficients of the output layer. In this study, the considered features are five parameters describing the spatio-temporal characteristics of the data (year, month, latitude, longitude and elevation) and eight meteorological variables (precipitation, air temperature, maximum and minimum air temperature, and four large-scale indices). In this setting, the monthly EDI is detected with ELM and classical ANN. The first outperforms the latter in terms of MSE and R-squared. A drawback of the application of this kind of networks is the lack of interpretability, so there is no clear identification of the main drivers.

Also (Li, et al., 2021) considers ELM, together with RF and SVM for regression to predict the SPEI index for the Colorado, Danube, Orange, and Pearl River basins with the antecedent SST fluctuation pattern (ASFP) as feature. The result shows that the ELM is the best-performing method among the three models applied.

(Dikshit, Pradhan, & Alamri, 2020) address the problem of detecting the SPEI index at different timescales for the region of New South Wales (NSW), Australia. Thirteen features are used as input, both related to meteorological data (precipitation, temperature, evapotranspiration, cloud cover, sea surface temperature) and climatic indices. ANN and SVM for regression are performed as supervised learning methods. ANN outperforms SVM in terms of MSE and R-squared. Moreover, the importance of features obtained from the ANN identifies that the sea surface temperature and the indices related to ocean oscillations are not relevant drivers for the drought event considered.

A different drought index, the Palmer Drought Severity Index, is predicted in (Tufaner & Özbeyaz, 2020) for the Adiyaman province, within the Middle Euphrates Section of the Southeastern Anatolia Region. Considered features are the monthly average of temperature, pressure, wind speed, relative humidity, rainfall, and some meteorological variables such as potential evapotranspiration, available water capacity and runoff. Four different regression algorithms are applied in this work (Linear regression, ANN, SVM, and Decision Trees), and ANN results to be the best-performing method.

A more advanced deep learning approach can be found in (Wu, Yin, He, & Li, 2022). In the paper, LSTMs are considered to detect and predict the three-month SPI index with atmospheric variables, combining dynamical models with ML and showing a significant improvement in the prediction, especially when the lead time exceeds one month.

Finally, (Zaniolo M. , Giuliani, Castelletti, & Pulido-Velazquez, 2018) introduce the *FRamework for Index-based Drought Analysis* (FRIDA) that uses an ELM-based wrapper for performing a feature selection in order to automatically construct a drought index representing a surrogate of the drought conditions of a river basin. The FRIDA index is computed by combining all the relevant available information from a set of candidate hydrometeorological observations and it is applied to the case study of the Jucar river basin (Spain), where it is shown to outperform a traditional index in representing the drought condition.

Other methods and drought indices can be found in the literature to address problems related to droughts. In (Aghelpour, Mohammadi, Biazar, Kisi, & Sourmirinezhad, 2020), different supervised learning approaches are applied. They focus on multivariate drought indices, which try to summarize different types of droughts that may happen simultaneously. The considered targets are the joint deficit index (JDI) and the multivariate standardized precipitation index (MSPI). Precipitation, temperature and the previous values of the considered indices are the features under analysis. The study is conducted in the widest climatic zone of Iran and the ML methods applied are group method of data handling (GMDH), generalized regression neural network (GRNN), least squared support vector machine (LSSVM), adaptive neuro-fuzzy inference system (ANFIS) and ANFIS optimized with three heuristic optimization algorithms. A feature selection technique based on entropy is also applied to filter the non-relevant features, which allows to conclude that precipitation and temperature are important drivers for this problem. Among the considered models, GMDH application obtains the best performance and in general the models obtain a better prediction of the MSPI index than the JDI index.

3.3 Causation

Causal inference and causal discovery methods can add value to the detection of droughts described in the previous section. Indeed, once the main candidate drivers of a drought event are identified, it would be relevant to understand if the features identified by the ML models can be labelled as *causes* of the extreme event or they are simply correlated with it. Many causal discovery approaches have been designed in recent years, although applications on complex, realistic problems like droughts are only a few.

One of the most famous causal discovery approaches for time series is Granger Causality, which is applied by (Gupta & Jain, 2021) to provide a causal interpretation of the connection between droughts and ENSO in India. SPI and SPEI indices were considered to represent drought conditions and four climatic indices were chosen to represent climatic conditions (southern oscillation index, northern oscillation index, NINO3 and NINO3.4 sea surface temperature indices) at different time scales (3,6,9 and 12 months). Granger causality was applied both with a linear and a nonlinear approach. Firstly, a linear regression and an ANN were trained considering all lagged values of the two drought indices to predict the value of current month. Then, the same models were trained considering also the lagged values of the climate indices as input features and the improvement of the prediction results quantifies the causal relationship between ENSO and drought conditions. In particular, most of the climatic indices showed a similar connection with the two drought indices, with the ANN-based approach identifying a larger area with significant causal effect with respect to the linear approach. Moreover, the impact of climatic indices was found to be larger with SPEI than the only precipitation-based index SPI.

In (Varando, Fernández-Torres, & Camps-Valls, 2021), Granger Causality is used as an additional criterion for training an autoencoder neural network to learn latent feature representations that are Granger-cause of the target index. The approach is demonstrated by analyzing the relationship between ENSO and vegetation greenness represented via the normalized difference vegetation index (NDVI), an index that is often used as a proxy to quantify the impacts of drought and plant stress.

Also (Noorbakhsh, Connaughton, & Rodrigues, 2020) focuses on inspecting the causal relationship between ENSO (identified by the Sea Surface Temperature (SST)) and drought conditions, finding that the cold SSTs in the previous year influence the occurrence of droughts in Ethiopia, as expected. In particular, droughts are represented as a univariate time series with monthly data considering

SPI index. Rotated Principal Component Analysis, on the other hand, is applied on the global SST dataset, obtaining 76 aggregated time series as features. As causal discovery method, PCMCI has been selected, which is a variation of Granger causality and the PC algorithm, specifically designed for applications on the Earth Sciences. The two strongest links identified by the method are an autocorrelation causal relationship of the SPI index with its one-month lagged value and a causal link between the 12-months lagged second principal component of the SST, whose strongest contribution comes from the ENSO region of the Pacific Ocean, and the SPI.

In a recent work (Shi, Zhao, Liu, Cai, & Zhou, 2022), Convergent Cross Mapping (CCM) method, an alternative to Granger causality based on dynamical systems, is applied for causal discovery on droughts for the Pearl River Basin in China. In particular, the interest of this study is to identify if there is a causal relationship between the meteorological and the hydrological drought condition. The SPI index is again considered as a meteorological drought index, while SRI is chosen to represent hydrological droughts. This work shows that there is a causal relationship between lagged meteorological droughts and hydrological droughts, as expected (scarcity of precipitation in the past causes scarcity of water in the future), having an almost constant impact from 1 to 6-month lag.

In (Rajsekhar, Singh, & Mishra, 2015) causal discovery is applied to droughts with an approach based on information theory quantities. In particular, Transfer Entropy is considered to evaluate the (possibly non-linear) causal effect in a way similar to Granger causality (i.e., measuring the additional information provided by the possible cause with respect to the autoregression) with a measure from information theory similar to (conditional) mutual information, but asymmetric. The study area considered is Texas, in the United States, where the Multivariate Drought Index (MDI) is computed as representative of droughts. Precipitation, runoff, soil moisture and evapotranspiration are considered as possible causes of the extreme event. From the results, it is possible to conclude that in the west of Texas, for the years 2015-2099, it is expected that precipitation will mostly cause droughts, together with runoff, evapotranspiration will also contribute with less strength and soil moisture will not be much relevant.

A climatic event related to droughts in India are summer monsoons, since a significant variation of their values from the average can lead to floods or droughts. In (Saha, Soni, Finley, & Monteleoni) causal discovery is applied to identify the regions of the Pacific Ocean influencing the Indian summer monsoons. Specifically, PCMCI algorithm is applied to identify the causal link between Sea Surface Temperature (SST) and Indian monsoons. The Pacific Ocean was divided into 24 areas and the average SST was computed for each of them, while average precipitation was computed for the entire India and for its four regions. These 29 variables were considered as the nodes of a causal graph and different regions of the ocean were identified as cause of monsoons for the different regions of India, allowing to conclude that precipitation is affected by many areas of the Pacific Ocean and not only by the areas that define the Nino index. Then, an RF model was applied considering the 3, 5, 8 and 10 regions of the Ocean identified as most significant for each region of India, with a test error below 10% for each one of them. In contrast, considering the NINO 3.4 index as it is traditionally done, results in a worse test performance.

3.4 Attribution

Attribution of human impact, climate change and global warming for droughts with machine learning has been addressed in a few papers, therefore, there are a lot of possibilities to extend the scientific knowledge on attribution both from a methodological and an applicative perspective.

In particular, feature selection for the identification of the main drivers allows to identify the candidate drivers of a drought event, and domain knowledge can attribute the obtained results to climate change. In (Richman & Leslie, 2018) several variables are studied to detect the drivers of the Cape Town droughts, focusing on a particularly impactful event in 2015-2017. The considered features include the Southern Annular Mode, Atlantic Meridional Mode, Indian Ocean Dipole, an Integrated Southern Hemisphere temperature index and several El Niño indices. Different feature selection techniques are applied to identify the impact of each candidate driver on the precipitation of the considered region, based on linear correlation in a forward and backward greedy search, genetic search algorithms, or wrapper techniques, which consider one of the mentioned techniques in a wrapper way, together with SVM. From the cross-validation results, it is possible to identify some attributes that emerge among most methods, other attributes that may be drivers since they are detected by some techniques, and attributes that are not selected by any method and that can be discarded. A peculiar result is that El Niño SST-based index and the El Niño atmospheric index (Southern Oscillation Index) are identified as ineffective in the prediction of precipitation, although they are traditionally assumed to be relevant variables. This is addressed as alarming, since the decreased predictive capacity of ENSO phases, which is usually the basis for seasonal forecasts, is a clear signal of climate change. (Richman & Leslie, 2020) analyze temperature and precipitation data from Perth, Australia, considering the Cape-Town drought analysis as a preliminary study. The candidate drivers considered are global temperature, sea surface temperature and a set of climate indices (Dipole Mode Index, Southern Oscillation Index, Niño 3, Niño 4, Pacific Decadal Oscillation, Antarctic Oscillation, Atlantic Meridional Oscillation, North Pacific Index, Interdecadal Pacific Oscillation) and their cross products. Preliminary statistical analysis (e.g., permutation test) was performed to show the clearly increasing pattern of temperature and the decreasing pattern of precipitation, as already found for Cape Town. Then, after the application of different feature selection techniques based on correlation or wrapper methods, linear regression, ANN, RF and SVM were applied, with the linear model and SVM found to perform better. ANN often had large prediction errors, which suggests relationships among the selected attributes and climate change, since ANN are known to have poor performance when there is a trend in the data.

Also (Hartigan, MacNamara, & Leslie, 2020) analyse precipitation and temperature trends for Canberra, Australia, finding a mean temperature increase and a stable annual precipitation due to a summer precipitation increase balanced by an autumn precipitation decrease. Wavelet analysis was then performed, showing ENSO influence on precipitation and temperature in Canberra, with less impact since 2000. Moreover, linear regression and SVM for regression were applied both as wrapper feature selection methods for attribution and for prediction, using as targets the maximum temperature and the mean precipitation. The identified relevant attributes for precipitation are ENSO, the southern annular mode, Indian Ocean Dipole and Tasman Sea SST anomalies. For the maximum temperature, Indian Ocean Dipole and global warming attributes are selected, showing a relevant trend due to climate change. In continuation with this study, (Hartigan, MacNamara, Leslie, & Speer, 2020) found, with permutation testing and other statistical analysis, a decreasing trend for annual precipitation across the Sidney catchment area, with significant reduction in summer and autumn. Then, wavelet analysis confirmed that the ENSO influence on precipitation has greatly weakened from the 2000s, suggesting the increased impact of climate change. Finally, considering as features the Atlantic Meridional Oscillation, the Dipole Mode Index (DMI), the global sea surface temperature anomalies, the global temperature anomalies, Niño3.4, the Tripole Index for the Interdecadal Pacific Oscillation, the Southern Annular Mode, the Southern Oscillation Index and the

Tasman Sea surface temperature anomalies, together with their two-way interactions, linear regression, SVM for regression and RF were applied for wrapper feature selection and prediction. SVM obtained the best score, with Niño3.4 and the interaction between DMI and TSST as selected attributes, representing respectively the influence of ENSO and global warming.

In (Shi, et al., 2020) the human impact was evaluated in a region of China where policies of ecological restoration have been applied since the 1980s. The dynamic characteristics of vegetation were addressed in terms of satellite data with NDVI as target, applying RF to identify feature importance. Nine meteorological factors were considered, with the average temperature, minimum temperature, maximum temperature and average relative humidity that have been identified as the most relevant. The combined effect of these variables contributes less to NDVI prediction than human activities. Human activity was identified as the most important attribute to the growth of NDVI index, with an increase of more than 40% from 1990 to 2015, allowing to conclude that the applied policies have been relevant.

A study related to attribution is (Prodhan, et al., 2022), where droughts were studied in South Asia. This work aimed to evaluate the future drought projections, assuming that they will be affected by climate change, global warming and increasing food demand. Five climate models were considered, taking into account a high greenhouse gas emission scenario, and three crop models were used to evaluate the future yield loss risk for rice, wheat and maize. Starting from these models, the SPEI index was considered as an input feature for identifying droughts and the yield loss risk index (YLRI) was adopted as target for identifying the yield loss risk. An ensemble ML technique was chosen: RF and Gradient Boosting were applied together, with an ANN that considered their output to produce the final prediction. The good fit of the climate model considering high emissions and climate change on the historical data underlines the reliability of a model that considers progressive increasing temperature for South Asia, with a projected level of SPEI index suggesting more duration and less intensification of droughts. The ML model showed good validation performance on historical data and predicts an increase of YLRI for the next years, with the largest risk associated with rice crops.

4 MACHINE LEARNING FOR TROPICAL CYCLONES

4.1 Overview

Tropical cyclones (TC) are some of the most devasting extreme events, causing extensive damage along their path either through flooding (caused by the torrential rainfall of the TCs themselves or by the storm surge they cause in coastal areas) or through the sheer force of their winds, which can reach upwards of 90 m/s (Chen, Giese, & Chen, 2020).

TC genesis, intensity, and trajectory are usually predicted with low accuracy by traditional dynamical models due to the complexity of the underlying physical processes and the coarse resolution of most models, which are unable to resolve the fine-grained wind fields that spin up tropical cyclones (Emanuel, 2018). For this reason, ML data-driven methods can be an added value to the knowledge and prediction of these events.

Following the taxonomy introduced in a recent literature review on the application of ML to tropical cyclones analysis (Chen, Zhang, & Wang, 2020), it is possible to identify two ways in which ML is applied to study TCs. First, fully data-driven ML models can be applied for genesis, track, intensity, and disastrous impact forecast or detection. Second, ML can be exploited to improve existing numerical models by automatically pre-processing data, tuning model parameters, or post-

processing the results. As highlighted in the review, the literature on ML methods applied to TC analysis is still young, and there remain several open challenges and opportunities.
(Chen, Zhang, & Wang, 2020) focus on applications involving classical ML methods (e.g., decision trees, logistic regression, support vector machines), ensemble methods (e.g., random forests, Boosting), and deep learning approaches (feedforward and convolutional neural networks). An overview of these applications is reported here, together with more recent results.

4.2 Detection and Forecast

This section is divided into five subsections, each describing the state of the art of applications of ML on a different aspect of tropical cyclones. Depending on the problem and the type of data (e.g., reanalysis or satellite, categorical or continuous), different ML models are most suitable for the detection and forecast of a specific subproblem.

4.2.1 TC Genesis

Forecasting the genesis of TCs typically consists in identifying a set of precursor variables that allow a skilful estimate of the probability that a TC may occur at a certain point in time and space, which can be seen as a classification problem.
Using non-ML methods, domain experts have identified key precursors~~,~~ which have proved to be good predictors of TC genesis in both reanalysis and simulated data: the Coriolis parameter, low-level relative vorticity, vertical wind shear, mid-troposphere relative humidity, ocean thermal energy, and the difference between the equivalent potential temperatures at the surface and at 500 hPa (Gray, 1998). However, the predictor variables may change depending on whether the lead time for the forecast is short (1-3 days) or long (months), leading to the distinction between short- and long-term TC genesis forecasting.

For short-term TC genesis forecasting, (Wijnands, Qian, & Kuleshov, 2016) identify, through feature selection based on mutual information and the Peter-Clark algorithm for directed acyclic graphs, potential vorticity (600 hPa), relative vorticity (925 hPa), and vertical wind shear (200–700 hPa) as the main drivers of TC genesis with 12-72 hours lead time, which are confirmed to be relevant by the satisfactory performance of a linear model. Similarly, (Zhang, Fu, Peng, & Li, 2015) used a decision tree and identified that maximum 800-hPa relative vorticity, sea surface temperature, precipitation rate, divergence averaged between 1000- and 500-hPa levels, and 300-hPa air temperature anomaly were the five main drivers of TC in the western North Pacific. In the same region, (Matsuoka, Nakano, Sugiyama, & Uchida, 2018) applied deep CNNs to simulated outgoing longwave radiation data to detect TCs. Differently from previous approaches, with CNNs there is no need to identify the drivers of the event, since the network automatically performs the feature extraction, usually with an improvement of performances at the cost of reduced interpretability.

More recently, (Pillay & Fitchett, 2021) used random forests to determine which large-scale atmospheric fields explain why some storms evolve into TCs and some do not, focusing on the Southern Hemisphere. Their analysis found that SST, air temperature, geopotential height, vertical wind shear, and relative humidity account for a large amount of the variability in the formation of TCs, and that SST was the single best predictor of whether a pre-existing storm will evolve into a TC.

Similarly, (Zhang, et al., 2019) used decision trees, k-nearest neighbours, ANNs, SVMs, AdaBoost and RF to determine whether pre-existing mesoscale convective systems (MCSs) would evolve into TC, concluding that AdaBoost is the best-performing method for short-term TC genesis forecast over the tropics and identifying low-level vorticity and genesis potential index as the most relevant input features.

Satellite data are also considered in some studies. (Park M.-S. , Kim, Lee, Im, & Park, 2016) extracted eight dynamic and hydrologic indices from wind and rainfall satellite images, applying a decision tree for detecting the TC, identifying circulation symmetry and intensity as the main drivers. (Kim, Park, Im, Park, & Lee, 2019) extended this study also considering RF and SVM, showing that SVM is the most performing algorithm for this problem.

As for long-term TC genesis forecasting, (Richman & Leslie, 2012) considered 108 predictors and used a wrapper feature selection method and an SVM for regression, reaching better results than a baseline linear model. This method is then refined in a subsequent work (Richman, Leslie, Ramsay, & Klotzbach, 2017). Beyond SVM-based methods, (Nath, Kotal, & Kundu, 2016) proposed an ANN to address the problem of determining the number of seasonal TC in the North Indian Ocean during the post-monsoon season, identifying geopotential height at 500 hPa, relative humidity at 500 hPa, sea-level pressure, and zonal wind at 700 hPa and 200 hPa for the preceding month (September) as the main drivers.

More recently, (Sun, Xie, Shah, & Shen, 2021) addressed the problem of predicting the number of TC in a season over the Atlantic basin, considering regional and global monthly features such as Pacific SST-related climate indices, El Nino Southern Oscillation (ENSO) related indices, and atmospheric and teleconnection indices, for a total of 34 features. To identify the most relevant ones, a clustering method was applied on the available features and the variable that is the most correlated with the target was selected in each cluster as a candidate driver. Then, Lasso regression was applied, and the results were shown to be comparable to a classical forecast and to be better when a subregion of the entire basin is considered.

Finally, (Asthana, Krim, Sun, Roheda, & Xie, 2021) addressed long-term TC prediction with ML, specifically applying CNNs to extract relevant features in order to predict the number of TC generations over the North Atlantic basin, with relevant performance results, at the cost of losing interpretability due to the complexity of the chosen model.

4.2.2 TC Track

TC tracking consists of predicting the future movement of a TC by considering the characteristics of the TC, of the general atmospheric circulation, and of the land/ocean (e.g., presence of mountains or cold ocean water) over which it passes. In the literature, feedforward and recurrent neural networks have shown promising results, together with advanced deep learning techniques like ConvLSTM, combining convolutional and recurrent neural networks or generative models like GANs. (Chen, Zhang, & Wang, 2020) review many applications of ML for TC tracking, dividing them into path prediction, predictors mining, and similarity search. These applications are mainly related to forecasting, since they typically do not try to identify the drivers that generate the TC, but rather forecast the trajectory of the TC depending on its characteristics and the conditions of the environment. Therefore, as these tasks are beyond the scope of CLINT, only a brief overview of the use of ML for this problem will be presented here.

In particular, *path prediction* has been addressed by (Ali, Kishtawal, & Jain, 2007) and (Wang, Zhang, & Fu, 2011) with feedforward neural networks, with good performance for up to 24 hours. (Moradi Kordmahalleh, Gorji Sefidmazgi, & Homaifar, 2016) and (Alemany, Beltran, Perez, & Ganzfried, 2019) applied recurrent neural networks to the problem to better consider the time component, obtaining better performances on longer lead times (up to 120 hours of forecast). To consider both space and time components, (Kim, et al., 2019) consider ConvLSTM neural networks. A more advanced approach, (Rüttgers, Lee, & You, 2018) and (Rüttgers, Lee, Jeon, & You, 2019), applies Generative Adversarial Networks to generate an accurate image of the future location of a TC given a satellite image of its past location.

A more recent work (Tan, Chen, & Wang, 2021) addresses TC tracking with an ensemble method, Gradient Boosting Decision Tree, which is shown to produce satisfactory results that outperform a standard numerical model (CLIPER) in the TC tracks in the Western North Pacific.

Predictors mining is more related to detection, since it aims to use ML to identify predictors that can be used by the trajectory forecasting models. This topic is not much addressed in the literature, since most of the TC tracking methods are based on path prediction. In (Zhang, Leung, & Chan, 2013) an approach with decision trees can be found, but there is room to apply some more advanced feature selection approaches, similar to what is done for TC genesis.

Finally, another approach to TC tracking is *similarity search,* which consists of developing search algorithms to identify similar historical cases to forecast the evolution of the TC under analysis. (Wang, Han, Lin, Shen, & Zhang, 2018) uses a neural network to perform the search, whereas usually clustering methods are applied to form clusters of TCs with similar characteristics. As non-exhaustive examples, (Kim & Seo, 2016) apply self-organising maps, while (Camargo, Robertson, Barnston, & Ghil, 2008) and (Ramsay, Camargo, & Kim, 2012) consider the path together with its shape and location to generate the clusters.

4.2.3 TC Intensity

In TC genesis, the main interest is identifying the probability that a TC will occur in a certain region, and it is usually framed as a classification problem. TC intensity estimation, on the other hand, is mainly tackled with regression methods. In this case, the goal is to predict the intensity of the TC, usually quantified as the maximum wind speed at a certain instant of time. As for TC tracking, the main application of ML focuses on predicting the next states of the TC based on the current and previous conditions. Since the principal type of data for this task are satellite images, the most widely adopted approaches are based on CNNs, which take as input the satellite image of a TC and try to predict a subsequent state of the TC, in terms of intensity and, potentially, also position. As it is a sequential problem, approaches also involving RNNs and LSTMs have been applied. As these data-driven approaches are mostly focused on satellite images and neural networks, the focus of the approaches is not usually on detection but on forecasting.

In (Chen, Zhang, & Wang, 2020) different intensity forecasting methods are reported. First, intensity estimation, which is strictly related to detection, is addressed: it focuses on identifying the main features of the current TC satellite image which can explain its intensity. This is different from detecting the main meteorological drivers that generate a TC and it can be seen more as a feature extraction data-driven approach directly applied to images of TCs. For this reason, approaches based on CNNs, which can automatically extract important features from satellite images, are the most applied. In particular, (Pradhan, Aygun, Maskey, Ramachandran, & Cecil, 2017) use a CNN to classify images of TCs into intensity classes. Their focus is on the inner representation of the CNN, seen as a feature extraction method to identify the characteristics of the TC. (Wimmers, Velden, & Cossuth, 2019) performed also framed TC intensity estimation as a classification task, using CNNs on satellite TC images, focusing on identifying which features are the most relevant and from which satellite they came from. More recent applications combine CNNs to extract the relevant characteristics of the TC from images with LSTMs and RNNs to keep track of the sequential nature of the evolution of the TC in time, as in (Lee, Im, Cha, Park, & Sim, 2019).

More recently, (Kar & Banerjee, 2021) address the problem of forecasting the intensity of TCs over the Bay of Bengal using satellite images. In particular, they propose a novel feature extraction method and the chosen features are used as inputs to five different intensity classifiers: naïve bayes, an SVM, a logistic model tree, a random tree and a random forest, with the random forest being the best performing approach. Also (Kim, Moon, Won, Kang, & Kang, 2021) address the problem of TC intensity forecast as a classification task, identifying the probability that a TC reaches a maximum intensity greater than 70 knots during its lifetime. Using a decision tree, they identify that the main drivers for TC intensification are the ocean thermal structures, the TC’s past trajectory, and the latitude of the TC’s current position.

Differently from intensity estimation, intensity prediction focuses on predicting the evolution of the intensity of the TC, potentially relying on the characteristics extracted in the estimation phase. Due to the sequential nature of this problem in time, recent approaches are based on RNNs (Pan, Xu, & Shi, 2019) or RNNs combined with CNNs (Chen, et al., 2019).

A problem related to TC intensity is intensity change prediction, which usually consists in a classification problem estimating the probability that the TC will get more or less intense. Examples of applications of ML on this task can be found again in (Chen, Zhang, & Wang, 2020) and they will not be addressed here since they are outside the scope of this report.

4.2.4 TC Weather and Disastrous Impact

Following the taxonomy of (Chen, Zhang, & Wang, 2020), the last purely data-driven family of ML approaches available in the literature for TCs is the forecast of severe rainfalls, storm surges, and wind fields, which are largely influenced by the occurrence of a TC. Since in this report we focus on TCs exclusively, these lines of research will only be briefly mentioned. In particular, ANNs, RFs and SVMs for regression are the methods usually applied to forecast the precipitation or storm surge height, while for wind fields recent approaches focus on CNN.

4.2.5 TC Numerical Models Improvement

To conclude this section based on the review from (Chen, Zhang, & Wang, 2020) and enriched by more recent applications of ML for TCs, this paragraph introduces a group of applications that apply ML to enhance numerical models of meteorology rather than constructing fully data-driven approaches for detection and forecast.

This can be done as a pre-processing step, applying ML to identify meaningful initial conditions for the numerical model. Although there exist some preliminary results based on giving a numerical model information on whether an area is identified as a TC region (Lee, Hall, Stewart, & Govett, 2018), there are no relevant applications in the literature for this task, since it is challenging to identify a proper supervised problem which can allow an ML method to learn good initial conditions.

Another way of applying ML to improve numerical models is to use ML algorithms to tune the hyperparameters of numerical models. This approach is much more studied, with some approaches that focus on learning the distribution and variability of the variables involved (Loridan, Crompton, & Dubossarsky, 2017) or parametrising the variables (typically as neural networks) (Jiang, Xu, & Wei, 2018).

A more recent work, (Baki, Chinta, Balaji, & Srinivasan, 2021), focus on hyperparameter tuning of the WRF model, which consists of more than 100 variables and simultaneously models the physics and the evolution of multiple meteorological quantities. Sensitivity analysis and the MARS method are used, which can be seen as a non-linear generalisation of multivariate linear regression. Ten TCs from the Bay of Bengal that happened in the previous ten years have been simulated and different meteorological variables have been analysed (wind speed, temperature, surface pressure, total precipitation, planetary boundary layer height, outgoing longwave radiation flux, downward shortwave radiation flux, and downward longwave radiation flux). Simulations have been then compared to observational data and the simulations based on the hyperparameters tuned with ML have been shown to outperform the simulations based on classical values of the parameters.

Finally, ML can be applied in the post-processing phase of a numerical model, elaborating the outputs to produce a more accurate result. Typical applications are studies based on the numerical model outputs for genesis, track, and intensity forecast and detection. Using the numerical models as simulators, ML methods take as input their outputs to perform the prediction tasks introduced in previous sections. Many successful results exist for this approach (Matsuoka, Nakano, Sugiyama, & Uchida, 2018), (Racah, et al., 2017), (Kim, et al., 2019). The main drawback of these applications is the bias introduced in considering numerical simulations rather than observational data, which should be inspected more in-depth from a theoretical perspective.

4.3 Causation

As presented in the previous paragraph, different meteorological drivers have been detected in many studies and a variety of input features from reanalysis data and satellite images have been considered to model the genesis, track and intensity of TCs. Causal inference and causal discovery methods can add value to the detection of candidate drivers, since they are able to identify if a candidate driver may cause the target, or if their relationship is just a correlation due to the fact that they have a similar response to external factors or they are caused by the same confounder.

Causality has been studied extensively in the last years from a theoretical perspective, but few methods were applied to TCs.

Granger causality is the most well-known causal approach for observational data, and it has been applied to TCs with statistical tests already in (Elsner, 2007), where a model with Atlantic SST as common cause of global mean near-surface air temperature (GT) and TC activity is compared with a model with GT causing Atlantic SST, which in turn causes TC activity. The results identify the possibility of a causal relationship between Atlantic SST and GT, with a causal direction from GT to SST, suggesting the validity of the second model.

(Dong, Lian, & Zhang, 2019) also consider Granger causality together with recurrent neural networks (GRUs in particular) to predict TC tracks. GRUs have been applied considering the causal meteorological variables identified as features (the maximum sustained wind speed is the one with the maximum causal relationship with the location of the TC) and results show good performance with respect to traditional deep learning methods in predicting the location of the TC.

In (Bai, Zhang, Bao, San Liang, & Guo, 2018), Information Flow, a concept based on information theory similar to transfer entropy and directed information, is used to derive that ENSO and Pacific decadal oscillation are the two most relevant causes of TC genesis position in the Atlantic basin. Moreover, for the Western area also the western Pacific subtropical high has a relevant causal effect, while in the Eastern region the monsoon trough has an impactful causal effect.

Also (Zhang, Ma, Li, Chen, & Bai, 2022) consider information flow for identifying causal relationships. The focus of their work is on TC genesis frequency in the Australian region, and the causal analysis allows them to conclude that the Atlantic meridional mode, Atlantic multidecadal oscillation, and north tropical Atlantic Sea surface temperature anomalies are all candidate causes of TC genesis frequency in this region, providing also a significant improvement of performance of predictors that consider them.

Similarly, (Pfleiderer, Schleussner, Geiger, & Kretschmer, 2020) work on the forecast of TC activity in late spring for the summer season, quantified as accumulated cyclone energy (ACE). The PCMCI algorithm is applied, and it is found that warm SST in the Atlantic and La Niña conditions in May are candidate causes of an active hurricane season, which is predicted with a satisfactory performance considering the selected candidate causes.

Finally, in (Bertrand, Pfleiderer, Kretschmer, Geiger, & Schleussner, 2019) causal discovery methods are applied. They conclude that Atlantic SST and mean sea level pressure over the Pacific have a causal relationship with the number of TCs expected in a season over the Atlantic basin.

4.4 Attribution

In this section we address two different aspects of attribution: the impact of anthropogenic climate change on TC and the impact of TC on human society. Attribution of TCs to human impact and climate change can make people aware of the possibility that changing their behaviour can modify the disastrous impact of TCs and, on the other hand, estimating the impact that TCs have on the economy or health provides an insight into the relevance that it has in real life.

Attribution of TCs genesis or intensity to human activity and climate change is a topic addressed in the literature with classical numerical models (Wehner, Zarzycki, & Patricola, 2019), (Knutson, et al., 2019), (Reed, Stansfield, Wehner, & Zarzycki, 2020). However, to date there are no publications addressing the problem with ML, leaving space to design ML-based applications and methods to address this aspect.

On the other hand, some studies exist that attribute the impact of TC to some specific societal problems.

(Nethery, et al., 2020) use ML methods to identify health impacts of TC in the United States. In particular, mortalities in the Medicare population, respiratory disease hospitalisations, chronic obstructive pulmonary disease hospitalisations, and cardiovascular disease hospitalisations in the Medicare fee-for-service population are considered. Moreover, temporally detailed track and feature data for each Atlantic-basin TC have been considered to characterise the TCs. Bayesian methods are selected as ML approaches. First, for each health outcome, causal inference submodels are used to estimate the excess health events attributable to historic TCs. Then, a predictive model for each health outcome is designed in order to attribute the relationship between the county-specific TC health effects and the features of the TC and county, with the purpose of characterising how these features modify the impact of the TC on health in a specific county. As a main result, respiratory hospitalisation is identified as the most TC-attributable health problem, with maximum windspeed being a strong predictor of the TC impact on this kind of health risk.

Another important aspect addressed by ML in (Wendler-Bosco & Nicholson, 2022) is the destructive impact of TCs in the United States, with a particular focus on the economic consequences. Eight ML approaches (such as RFs and SVMs) were considered in this study, with features characterising as target the TCs and storm damage ratio, which is the ratio between immediate storm damages (in terms of dollars) and annual gross domestic product. Tree-based methods perform best in this setting, and some main features are identified, such as the storm size and the speed of the TC movement being important for the prediction, while maximum wind speed is found not to be relevant, despite the fact that it is usually considered a key identifying feature of a TC.

Finally, an application of ML to evaluate the impact of events related to TCs on ecosystems can be found in (Zhang X. , et al., 2021), which identifies through an RF the typhoon Lekima impact on Chinese forests, computing with small uncertainty the proportion of damage.

5 MACHINE LEARNING FOR HEATWAVES AND WARM NIGHTS

5.1 Overview

Heatwaves are extreme events, usually defined as prolonged periods of maximum or average daily temperature significantly higher than the average (Russo, Sillmann, & Fischer, 2015). These events broadly impact many social, economic and environmental systems, including elderly health, animals, ecosystems, etc. (Stillman, 2019). For example, in agriculture, prolonged periods of very hot temperature have a harmful effect on plants and water availability conditions that may compromise yields, even in the presence of irrigation plans (Lobell, Cahill, & Field, 2007). Warm nights, on the contrary, are episodes of elevated night temperatures that are more problematic from a perspective of diffusion of diseases and pests. Warm nights are not necessarily related to heatwaves, and different drivers and processes may be associated with each of them. As already discussed, these extreme events may have a severe impact on society, and, on the other hand, they are becoming more and more frequent due to global warming (Chapman, Watkins, & Stainforth, 2019), (Bador, et al., 2017). Therefore, attribution to anthropogenic climate change is also a relevant problem, with the purpose of identifying possible changes in human habits to mitigate the occurrence of these events. For these reasons, Machine Learning (ML) approaches may be good methods to inspect and predict in a data-driven fashion the most relevant drivers and the possible occurrence of such events. Subsequently, causal inference can lead to new insights regarding the two phenomena, their drivers and their interconnection. Finally, attribution with ML methods can better identify the impact of humans on the occurrence of these events. A more advanced step may be to investigate the relationship between the occurrence of the two events, in order to identify common drivers and the possibility of their simultaneous occurrence (further details for these concurrent extremes problems can be found in the next chapter). In the literature, some approaches are available for the detection, forecast, causation and attribution of extreme temperatures using ML, applied only on night-time data to detect warm nights. However, only a few applications address heatwaves and even fewer applications are available that directly address warm nights events. Due to the restricted amount of relevant recent papers on these topics, they will be discussed altogether in the following section.

5.2 Detection, Forecast, Causation and Attribution

5.2.1 Heatwaves

Only a few papers address the problems of detection and forecast of heatwaves with Machine Learning techniques in the literature. (Chattopadhyay, Nabizadeh, & Hassanzadeh, 2020) combine analogue forecasting (which can be seen as a K-nearest-neighbor approach) with deep learning to predict heatwaves (and cold spells). Surface air temperature and geopotential height at 500 mb have been adopted as daily data and the ANN chosen is a CapsNet. The network is trained in a supervised manner by auto labeling the data, assigning a label from 0 to 4 each day depending on the value of temperature over North America several days later. From the results, it is possible to conclude that the network is able to predict the occurrence of a heatwave (or cold wave) from 1 to 5 days ahead and using only the values of geopotential height with satisfactory results, outperforming standard ANN and logistic regression. From this result, it is also possible to conclude that geopotential height is a candidate driver for heatwaves. Also (Asadollah, et al., 2022) address the forecast and detection of heatwaves, focusing on the eight climate regions of Iran. Daily maximum temperature has been adopted to identify heatwave days, while six reanalysis meteorological features at four different pressure levels have been employed as candidate drivers (air temperature, geopotential height, relative humidity, specific humidity, U wind, V wind). The ML models selected are decision trees, RF and AdaBoost with decision tree predictors. After applying PCA to reduce collinearity among features, AdaBoost is found to be the best performing method for this application, using humidity and wind component as features, which are therefore identified as the main drivers.

Attribution problems, i.e., the causal relationships between anthropogenic climate change and heatwaves have been also addressed in some works. In (Pasini, Racca, Amendola, Cartocci, & Cassardo, 2017), attribution is addressed with neural networks. In particular, the paper underlines that Global Climate Model ensembles show human impact as the main driver of temperature increase. The aim of the work is to prove the robustness of these results with a completely data-driven approach. The yearly average global temperature is considered as target of two neural network ensemble models. The first one takes as input the complete set of natural and anthropogenic factors RFANTH (anthropogenic forcing), RFSOLAR (solar activity), RFVOL (volcanoes) from 1850 to 2010. The second ANN takes as input the same quantities, but the value of anthropogenic forcing is fixed at its preindustrial value of 1850. The validation results show that in the first case the increasing trend of global temperature is clearly identified by the network, with good performance scores, while in the second one the recent increase in temperatures is not identified, with a constant trend from 1960. Moreover, a more detailed analysis has been performed on residuals, identifying the main drivers for the years 1960-2010 two anthropogenic factors: greenhouse gases as the main driver and black carbon as a less but still significant one. On the other hand, in the period 1910-1975 solar irradiation is identified as the main driver. (Park & Kim, 2018) address heatwave attribution focusing on the relationship between temperature and people hospitalised with heat illness. Nineteen meteorological variables, together with daily data of heat illness hospital patients from 2011 to 2016 in Seoul metropolitan area, Korea, are considered. The purpose of this study was to determine the threshold temperature that defines a heatwave, using a data-driven approach that depends on the meteorological quantities together with the impact of the high temperature on human health. In this work, a spline regression model (MARS) was applied to identify the relevant changing points of the relationship between variables. From the results it is possible to conclude that low temperatures have little impact on the number of people hospitalised for heat-related illnesses, while this only becomes relevant when the temperature is above 32.58 C for two consecutive days.

Finally, causality for heatwaves has been addressed in some articles, including recent works (Vijverberg & Coumou, 2022). In particular, in that paper the causal relationship between Sea Surface Temperature (SST) and Rossby waves (RW) is discussed, since the latter have been identified to be related to extreme heatwaves. Since the two considered variables are highly correlated, PCMCI approach has been applied as a causal discovery method to infer the relationship SST-RW for eastern and western US. With this method it is confirmed that both in the west and in the east SST is a plausible cause for RW at a daily scale. However, only for the eastern US there exists a long-lead time causal link from SST to RW, suggesting that it is possible to perform a long-lead predictability from SST to RW and therefore to temperature. Two previous papers already addressed causality in heatwaves by applying the Granger method. In (Ratnam, Behera, Ratna, Rajeevan, & Yamagata, 2016) causality is inspected for Indian heatwaves. In particular, two different analyses have been performed for India's north-central area and east coast. Granger causality is exploited in the north-central area to identify 200 hPa geopotential height anomalies over North Atlantic as causes of the daily maximum temperatures with a two-day lag. On the Indian east coast, on the other hand, Granger causality is applied between the daily 850hPa eddy stream function anomalies in West Pacific and daily maximum temperature, identifying that the first causes the latter with a lag of one day. Finally, (Li, Tam, Tai, & Lau, 2021) exploit Granger causality to identify the relationship between summer heatwaves and vegetation cover (LAI index). In this study, a strong correlation between LAI index and heatwaves is identified. In particular, in Central Europe and the southern and southeastern parts of North America, heatwaves are more frequent with lower LAI, while for the northwestern and northeastern parts of North America, the opposite holds. These results are supported by further analysis and Granger causality application through statistical tests.

5.2.2 Warm Nights

To the authors' knowledge, there are no specific works in the literature devoted to ML applications in the detection, causation and attributions of warm nights. For this reason, in this section, a few recent works on ML applications for extreme temperatures are briefly discussed, considering that similar methodologies can be applied directly to warm nights. (Paniagua-Tineo, et al., 2011) addressed the problem of predicting the daily maximum temperature with SVM for regression in different measuring stations in Europe. Meteorological variables (e.g., temperature, precipitation, relative humidity, air pressure) are considered as features together with the daily synoptic situation of the day and the monthly cycle. From those results, it is possible to conclude that the model can accurately predict the maximum temperature for the subsequent 24 hours. The SVM for regression is also shown to outperform ANN-based methods in this task and this good performance suggests that the selected variables are relevant drivers of the maximum temperature value. More recently, (Peng, Zhi, Ji, Ji, & Tian, 2020) applied ANN and Natural Gradient boosting to predict maximum temperature with a lead time of 1-35 days over East Asia. The method is shown to outperform a classical model output postprocessing (EMOS), with better performances in more than the 90% of the overall area. Relating forecast and causality, (Oettli, et al., 2022) proposed a hybrid approach to predict the air temperature through the sea surface temperature (SST), combining dynamical forecast and nine different ML methods that converge into an ensemble model. The proposed methodology is applied to Japan's central region, and the experimental results show a satisfactory performance for two months lead time. Moreover, the causality between SST and air temperature is also addressed. In particular, the information flow, a causal quantity related to transfer entropy, is evaluated between the SST at each point of the available global grid and the air temperature of the region of interest, identifying the regions of the Earth where the SST is a candidate to be causally related to the air temperature of central Japan. (Attanasio, 2012) also applies causality analysis to temperature anomalies, looking for evidence of the impact of natural and anthropogenic warming on the temperature extremes, which makes this work related both to causality and attribution. In this paper, Granger causality is adopted to show that there is little connection between natural forcings and global temperature anomalies, while the greatest influence on them is from CO_2, with also a relevant but smaller influence of methane. Although this result seems to indicate that there is no causal relationship between meteorological drivers and temperature anomalies, it must be underlined that this experiment is conducted assuming linear settings and it only considers total solar irradiance, cosmic ray intensity and stratospheric aerosol optical thickness as candidate meteorological causes. Some specific applications of ML for attribution of the maximum temperature value to climate change also exist. For example, in (Chithra, et al., 2015) ANNs are applied to evaluate the climate change impact on the monthly maximum temperature of the Chaliyar river basin, India. Considering reanalysis data and dynamical model predictors in a climate change context, different ANNs have been trained and validated for different seasons. Then, considering different dynamical models, the input variables have been simulated for the future and in all scenarios maximum temperature is predicted to increase by the next hundred years between one and three degrees.

6 MACHINE LEARNING FOR COMPOUND EVENTS AND CONCURRENT EXTREMES

Compound events are combinations of events, that are not necessarily extremes, and can lead to significant impacts for humans and ecosystems. The individual events may occur simultaneously or within a specific time period, and they may have additive or even multiplicative effects.
Concurrent extremes are defined as the occurrence of two extreme weather events, simultaneously or with a certain time lag. The events may be of a different type (e.g., a heatwave and a drought) and occur in the same location with a temporal lag or of the same type in two different locations within a specific time period. (Toreti, Cronie, & Zampieri, 2019) provides an overview of the problem of the identification of concurrent extremes. The paper proposes a methodology based on the marked inhomogeneous J-Function (Cronie & van Lieshout, 2016), using an application focused on wheat-producing regions.

The detection and study of compound events and concurrent extremes is vital for a coherent impact assessment (Zampieri, Ceglar, Dentener, & Toreti, 2017) (Zscheischler, et al., 2018). However, the scarcity of data available where these multiple events have been labelled, the large variety of combinations of events that may be addressed as concurrent extremes and the broad definition of compound events hinder the direct application of ML techniques to provide a data-driven forecast of their occurrence, or to discriminate between drivers and causes. For this reason, very little literature addresses these problems with ML, making this literature stream largely unexplored, with much room for improvement.

A recent technical report (Feng, et al., 2021) introduces a possible ML pipeline that can be followed to address different problems related to compound events and concurrent extremes with ML. The first step is the identification of these events, which is still an open problem, since there are few datasets to train the models on. Unsupervised ML techniques can also work with unlabeled observations. Then the focus is on detecting the main drivers to identify some variables that lead to the occurrence of these events, which can be precursors for a future event. Moreover, explainable AI methods are suggested to evaluate the drivers' importance in ML model prediction. Finally, model ensembles and probabilistic models are suggested as possible relevant supervised learning techniques to perform a final forecast of these events. A collection of articles (Zhang, Murakami, Khouakhi, & Luo, 2021) also presents concurrent extremes as a relevant open topic. This collection is composed of articles related to dynamical models, statistics or machine learning, all aiming to advance this topic. Two articles are particularly relevant from this collection. (Huang, et al., 2021) focused on solar radiation prediction, performing twelve different ML methods, identifying meteorological variables as crucial for the performance of the models and selecting XGboost and a stacking ensemble model as the best-performing methods. The models confirm that the maximum of the mean ground temperature increases with solar radiation, leading to the conclusion that solar radiation is one of the most impactful variables for concurrent extremes. Moreover, (Wang, Zhao, Gao, Zhang, & Feng, 2021) combine ML approaches (Isolation Forests) to identify the set of outliers, which can be a valid alternative to labelled datasets. In the paper, a statistical model is applied to identify the critical points. The experiment is conducted in China, studying the connection of extreme events between the Pearl River Delta and the Yangtze River Delta regions, considering TC, precipitation and temperature data. A significant correlation is found between the heatwaves in the Pearl River Delta and the extreme precipitation in the Yangtze River Delta identified by the method, suggesting that their occurrence can be considered as concurrent extremes. Some preliminary work has also been done for compound events and machine learning. For example, (Sweet & Zscheischler, 2022) present a study where crop yield failure is explained through ML, with a data-driven approach based on RF that identifies the critical meteorological conditions that lead to a severe impact on agricultural yields.

Since detection and forecast have very preliminary ML applications for these events, to the authors' knowledge, causality and attribution have not been addressed in the literature so far. Indeed, they usually require a robust methodology for detection to add significance to the results.

7 CONCLUSIONS

Detection, causation, and attribution of EE are difficult tasks, with many physical processes involved that are difficult to be represented by classical dynamical models. ML can enhance the capability to perform these tasks by exploiting the information brought by observational data, with the possibility to also consider simulated data and estimating the bias increased by the simulations.
In particular, ML can provide efficient methods to process big climate datasets, with the purpose of extracting information and patterns from spatially and temporally distributed data of climatological variables associated with EE.

Different challenges arise when dealing with spatio-temporal climatic datasets:

- Highly correlated variables, with correlations decreasing with distance, which need to be discarded or aggregated to avoid high collinearity between features.
- Time delays, that make more difficult the feature extraction process, since the value of a variable in a certain position may impact the target in a different position many timesteps later.
- Highly non-linear dependencies, that lead to the necessity to consider a complex model and quantities that are able to evaluate non-linear relationships, rather than classical correlations.
- High dimensional datasets, usually with more features than samples, that underline the necessity to reduce the dimensionality before estimating a ML model.

This document provides an extensive overview of ML methods suitable to tackle the different challenges that arise in this context, with a methodological overview of dimensionality reduction, feature selection, supervised learning and causal inference. The first two subfields are proposed to reduce the huge number of variables and to select the relevant candidate drivers, to feed supervised learning techniques. Finally, causal inference may be able to give a causal insight of the identified dependencies. Then, the focus of this work is on a literature review of ML applications for the detection, causation and attribution of the EE addressed in CLINT.
A major pillar of this document is indeed the state-of-the-art analysis to identify how the challenges that arise in this context are addressed in the literature for the EE under analysis. In particular, the methodologies addressed in Chapter 2 are applied in most of the works reviewed in the subsequent chapters, together with other methods such as oversampling and undersampling to address imbalanced classification or explainable AI algorithms to identify the relevant variables complex models (e.g., ANNs).

From the literature review performed, it is possible to draw some conclusions that are common to all the EE analysed. On one hand, detection with ML is the most addressed task in the literature, with many applications that apply different feature selection techniques and a final prediction with ensemble or ANN-based algorithms. On the other hand, causation and attribution with ML are less inspected in the literature: the first is a relatively new field, where methods able to scale to huge dimensions have been introduced only recently and usually relies on strong assumptions, while attribution is a variation of detection focused on human impact and can be addressed with standard ML techniques.

There are also peculiarities that are specific of each extreme events. Detection of droughts is largely addressed in the literature, mostly focused on the detection of drought indices such as the SPI and the SPEI. Also detection of TCs is a well-studied topic and many aspects can be analysed: genesis, tracking, intensity, impact, together with the improvement of indices. A few works have been found on the detection of heatwaves and warm nights with ML, that mostly focus on the mean or the maximum temperatures. Fewer works have been identified for compound events and concurrent extremes, that have not been clearly formulated as ML problems yet. Finally, as already discussed, causation and attribution are in general less studied, with some relevant results found for droughts and TCs.

In conclusion, this document is a review of the subfields of ML identified as relevant for the detection, causation and attribution of EE and of the existing applications available in the literature. This is a first step that allows to identify which methods can be used in the framework of CLINT and which methods need to be improved to provide successful results in the project, with the design of new algorithms.

REFERENCES

Aghelpour, P., Mohammadi, B., Biazar, S. M., Kisi, O., & Sourmirinezhad, Z. (2020). A theoretical approach for forecasting different types of drought simultaneously, using entropy theory and machine-learning methods. *ISPRS International Journal of Geo-Information, 9*, 701.

Alemany, S., Beltran, J., Perez, A., & Ganzfried, S. (2019). Predicting hurricane trajectories using a recurrent neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, pp. 468–475.

Ali, M. M., Kishtawal, C. M., & Jain, S. (2007). Predicting cyclone tracks in the north Indian Ocean: An artificial neural network approach. *Geophysical research letters, 34*.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Artac, M., Jogan, M., & Leonardis, A. (2002). Incremental PCA for on-line visual learning and recognition. *2002 International Conference on Pattern Recognition*, *3*, pp. 781–784.

Asadollah, S. B., Khan, N., Sharafati, A., Shahid, S., Chung, E.-S., & Wang, X.-J. (2022). Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models. *Stochastic environmental research and risk assessment, 36*, 1959–1974.

Asthana, T., Krim, H., Sun, X., Roheda, S., & Xie, L. (2021). Atlantic hurricane activity prediction: A machine learning approach. *Atmosphere, 12*, 455.

Attanasio, A. (2012). Testing for linear Granger causality from natural/anthropogenic forcings to global temperature anomalies. *Theoretical and applied climatology, 110*, 281–289.

Bühlmann, P. (2020). Invariance, causality and robustness. *Statistical Science, 35*, 404–426.

Bador, M., Terray, L., Boe, J., Somot, S., Alias, A., Gibelin, A.-L., & Dubuisson, B. (2017). Future summer mega-heatwave and record-breaking temperatures in a warmer France climate. *Environmental Research Letters, 12*, 074025.

Bai, C., Zhang, R., Bao, S., San Liang, X., & Guo, W. (2018). Forecasting the tropical cyclone genesis over the Northwest Pacific through identifying the causal factors in cyclone–climate interactions. *Journal of Atmospheric and Oceanic Technology, 35*, 247–259.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association, 101*, 119–137.

Baki, H., Chinta, S., Balaji, C., & Srinivasan, B. (2021). Determining the sensitive parameters of WRF model for the prediction of Tropical cyclones in Bay of Bengal using Global sensitivity analysis and Machine learning. *arXiv preprint arXiv:2107.04824*.

Barshan, E., Ghodsi, A., Azimifar, Z., & Jahromi, M. Z. (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition, 44*, 1357–1371.

Beck, A., & Kurz, M. (2020, October). A Perspective on Machine Learning Methods in Turbulence Modelling. *A Perspective on Machine Learning Methods in Turbulence Modelling*. doi:10.13140/RG.2.2.17469.69608

Belayneh, A., & Adamowski, J. (2013). Drought forecasting using new machine learning methods. *Journal of Water and Land Development*.

Belayneh, A., Adamowski, J., & Khalil, B. (2016). Short-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet transforms and machine learning methods. *Sustainable Water Resources Management, 2*, 87–101.

Belayneh, A., Adamowski, J., Khalil, B., & Ozga-Zielinski, B. (2014). Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. *Journal of Hydrology, 508*, 418–429.

Belayneh, A., Adamowski, J., Khalil, B., & Quilty, J. (2016). Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmospheric research, 172*, 37–47.

Belkin, M., & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems, 14*.

Beraha, M., Metelli, A. M., Papini, M., Tirinzoni, A., & Restelli, M. (2019). Feature selection via mutual information: New theoretical insights. *2019 International Joint Conference on Neural Networks (IJCNN)*, (pp. 1–9).

Bertrand, M., Pfleiderer, P., Kretschmer, M., Geiger, T., & Schleussner, C.-F. (2019). Using discovery algorithms to forecast seasonal tropical cyclone genesis in the Atlantic. *Geophysical Research Abstracts*, *21.*

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123–140.

Breiman, L. (1997). *Arcing the edge.* Tech. rep., Technical Report 486, Statistics Department, University of California.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5–32.

Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research, 13*, 27–66.

Camargo, S. J., Robertson, A. W., Barnston, A. G., & Ghil, M. (2008). Clustering of eastern North Pacific tropical cyclone tracks: ENSO and MJO effects. *Geochemistry, Geophysics, Geosystems, 9*.

Caruana, R. (1997). Multitask learning. *Machine learning, 28*, 41–75.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*, 16–28.

Chao, G., Luo, Y., & Ding, W. (2019). Recent advances in supervised dimension reduction: A survey. *Machine learning and knowledge extraction, 1*, 341–358.

Chapman, S. C., Watkins, N. W., & Stainforth, D. A. (2019). Warming trends in summer heatwaves. *Geophysical Research Letters, 46*, 1634–1640.

Chattopadhyay, A., Nabizadeh, E., & Hassanzadeh, P. (2020). Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems, 12*, e2019MS001958.

Chen, A., Giese, M., & Chen, D. (2020). Flood impact on Mainland Southeast Asia between 1985 and 2018—The role of tropical cyclones. *Journal of flood risk management, 13*, e12598.

Chen, R., Wang, X., Zhang, W., Zhu, X., Li, A., & Yang, C. (2019). A hybrid CNN-LSTM model for typhoon formation forecasting. *GeoInformatica, 23*, 375–396.

Chen, R., Zhang, W., & Wang, X. (2020). Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere, 11*, 676.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, (pp. 785–794).

Chen, W., Cai, R., Zhang, K., & Hao, Z. (2021). Causal discovery in linear non-gaussian acyclic model with multiple latent confounders. *IEEE Transactions on Neural Networks and Learning Systems*.

Chithra, N. R., Thampi, S. G., Surapaneni, S., Nannapaneni, R., Reddy, A., & Kumar, J. D. (2015). Prediction of the likely impact of climate change on monthly mean maximum and minimum temperature in the Chaliyar river basin, India, using ANN-based models. *Theoretical and Applied Climatology, 121*, 581–590.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*, 273–297.

Cronie, O., & van Lieshout, M. N. (2016). Summary statistics for inhomogeneous marked point processes. *Annals of the Institute of Statistical Mathematics, 68*, 905–928.

Cunningham, J. P., & Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research, 16*, 2859–2900.

Deo, R. C., & Şahin, M. (2015). Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. *Atmospheric Research, 153*, 512–525.

Dikshit, A., Pradhan, B., & Alamri, A. M. (2020). Temporal hydrological drought index forecasting for New South Wales, Australia using machine learning approaches. *Atmosphere, 11*, 585.

Dong, P., Lian, J., & Zhang, Y. (2019). A novel data-driven approach for tropical cyclone tracks prediction based on Granger causality and GRU. *2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, (pp. 70–75).

Duda, R. O., Hart, P. E., & others. (2006). *Pattern classification.* John Wiley & Sons.

Eberhardt, F. (2017). Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics, 3*, 81–91.

Elsner, J. B. (2007). Granger causality and Atlantic hurricanes. *Tellus A, 59*, 476–485.

Emanuel, K. (2018). 100 years of progress in tropical cyclone research. *Meteorological Monographs, 59*, 15–1.

Felsche, E., & Ludwig, R. (2021). Applying machine learning for drought prediction using data from a large ensemble of climate simulations. *Nat. Hazards Earth Syst. Sci. Discuss*, 1–20.

Feng, P., Wang, B., Li Liu, D., & Yu, Q. (2019). Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia. *Agricultural Systems, 173*, 303–316.

Feng, Y., Maulik, R., Wang, J., Balaprakash, P., Huang, W., Rao, V., Sullivan, R. (2021). *Characterization of Extremes and Compound Impacts: Applications of Machine Learning and Interpretable Neural Networks.* Tech. rep., Artificial Intelligence for Earth System Predictability.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics, 7*, 179–188.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique, 57*, 238–247.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine learning research, 5*.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*, 119–139.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis, 38*, 367–378.

Galatzer-Levy, I., Ruggles, K., & Chen, Z. (2018, January). Data Science in the Research Domain Criteria Era: Relevance of Machine Learning to the Study of Stress Pathology, Recovery, and Resilience. *Chronic Stress, 2*, 247054701774755. doi:10.1177/2470547017747553

Gerhardus, A., & Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems, 33*, 12615–12625.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning, 63*, 3–42.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 3-42.

Girolami, M., & Fyfe, C. (1997). Stochastic ICA contrast maximisation using Oja's nonlinear PCA algorithm. *International journal of neural systems, 8*, 661–678.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics, 10*, 524.

Golub, G. H., & Reinsch, C. (1971). Singular value decomposition and least squares solutions. In *Linear algebra* (pp. 134–151). Springer.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*, 139–144.

Gray, W. M. (1998). The formation of tropical cyclones. *Meteorology and atmospheric physics, 67*, 37–69.

Gupta, V., & Jain, M. K. (2021). Unravelling the teleconnections between ENSO and dry/wet conditions over India using nonlinear Granger causality. *Atmospheric Research, 247*, 105168.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*, 1157–1182.

Hara, S., & Maehara, T. (2017). Enumerate lasso solutions for feature selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31.*

Hartigan, J., MacNamara, S., & Leslie, L. M. (2020). Application of machine learning to attribution and prediction of seasonal precipitation and temperature trends in Canberra, Australia. *Climate, 8*, 76.

Hartigan, J., MacNamara, S., Leslie, L. M., & Speer, M. (2020). Attribution and prediction of precipitation and temperature trends within the Sydney catchment using machine learning. *Climate, 8*, 120.

He, X., Cai, D., & Niyogi, P. (2005). Laplacian score for feature selection. *Advances in neural information processing systems, 18*.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science, 313*, 504–507.

Hobeichi, S., Abramowitz, G., Evans, J. P., & Ukkola, A. (2022). Toward a Robust, Impact-Based, Predictive Drought Metric. *Water Resources Research, 58*, e2021WR031829.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*, 1735–1780.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology, 24*, 417.

Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C., & Zeng, Z. (2021). Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Frontiers in Earth Science, 9*, 596860.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks, 10*, 626–634.

Jenssen, R. (2009). Kernel entropy component analysis. *IEEE transactions on pattern analysis and machine intelligence, 32*, 847–860.

Jiang, G.-Q., Xu, J., & Wei, J. (2018). A deep learning algorithm of neural network for the parameterization of typhoon-ocean feedback in typhoon forecast models. *Geophysical Research Letters, 45*, 3706–3716.

Jing, L., Zhang, C., & Ng, M. K. (2012). SNMFCA: Supervised NMF-based image classification and annotation. *IEEE transactions on image processing, 21*, 4508–4521.

Kar, C., & Banerjee, S. (2021). Tropical cyclone intensity classification from infrared images of clouds over Bay of Bengal and Arabian Sea using machine learning classifiers. *Arabian Journal of Geosciences, 14*, 1–17.

Khan, N., Sachindra, D. A., Shahid, S., Ahmed, K., Shiru, M. S., & Nawaz, N. (2020). Prediction of droughts over Pakistan using machine learning algorithms. *Advances in Water Resources, 139*, 103562.

Kim, H.-K., & Seo, K.-H. (2016). Cluster analysis of tropical cyclone tracks over the western North Pacific using a self-organizing map. *Journal of Climate, 29*, 3731–3751.

Kim, M., Park, M.-S., Im, J., Park, S., & Lee, M.-I. (2019). Machine learning approaches for detecting tropical cyclone formation using satellite data. *Remote Sensing, 11*, 1195.

Kim, S., Kim, H., Lee, J., Yoon, S., Kahou, S. E., Kashinath, K., & Prabhat, M. (2019). Deep-hurricane-tracker: Tracking and forecasting extreme climate events. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (pp. 1761–1769).

Kim, S.-H., Moon, I.-J., Won, S.-H., Kang, H.-W., & Kang, S. K. (2021). Decision-tree-based classification of lifetime maximum intensity of tropical cyclones in the tropical western north pacific. *Atmosphere, 12*, 802.

Knutson, T., Camargo, S. J., Chan, J. C., Emanuel, K., Ho, C.-H., Kossin, J. (2019). Tropical cyclones and climate change assessment: Part I: Detection and attribution. *Bulletin of the American Meteorological Society, 100*, 1987–2007.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence, 97*, 273–324.

Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE transactions on pattern analysis and machine intelligence, 28*, 1393–1403.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*, 2278–2324.

Lee, J., Im, J., Cha, D.-H., Park, H., & Sim, S. (2019). Tropical cyclone intensity estimation using multi-dimensional convolutional neural networks from geostationary satellite data. *Remote Sensing, 12*, 108.

Lee, Y.-J., Hall, D., Stewart, J., & Govett, M. (2018). Machine learning for targeted assimilation of satellite data. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (pp. 53–68).

Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.*

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR), 50*, 1–45.

Li, J., Tam, C.-Y., Tai, A. P., & Lau, N.-C. (2021). Vegetation-heatwave correlations and contrasting energy exchange responses of different vegetation types to summer heatwaves in the Northern Hemisphere during the 1982–2011 period. *Agricultural and Forest Meteorology, 296*, 108208.

Li, J., Wang, Z., Wu, X., Xu, C.-Y., Guo, S., Chen, X., & Zhang, Z. (2021). Robust meteorological drought prediction using antecedent SST fluctuations and machine learning. *Water Resources Research, 57*, e2020WR029413.

Lin, D., & Tang, X. (2006). Conditional infomax learning: An integrated framework for feature extraction and fusion. *European conference on computer vision*, (pp. 68–82).

Liu, H., & Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. *Proceedings of 7th IEEE international conference on tools with artificial intelligence*, (pp. 388–391).

Liu, Y., & Aviyente, S. (2012). The relationship between transfer entropy and directed information. *2012 IEEE Statistical Signal Processing Workshop (SSP)*, (pp. 73–76).

Lo, B., Gui, R., Honda, H., & Davis, K. (2019, September). Artificial Intelligence-Based Drug Design and Discovery. doi:10.5772/intechopen.89012

Lobell, D. B., Cahill, K. N., & Field, C. B. (2007). Historical effects of temperature and precipitation on California crop yields. *Climatic change, 81*, 187–203.

Loridan, T., Crompton, R. P., & Dubossarsky, E. (2017). A machine learning approach to modeling tropical cyclone wind field uncertainty. *Monthly Weather Review, 145*, 3203–3221.

Lu, Y., Lai, Z., Xu, Y., Li, X., Zhang, D., & Yuan, C. (2016). Nonnegative discriminant matrix factorization. *IEEE Transactions on Circuits and Systems for Video Technology, 27*, 1392–1405.

Malinsky, D., & Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*, (pp. 23–47).

Massey, J., & others. (1990). Causality, feedback and directed information. *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, (pp. 303–305).

Matanga, Y. (2017, September). *Analysis of Control Attainment in Endogenous Electroencephalogram Based Brain Computer Interfaces.* Ph.D. dissertation. doi:10.13140/RG.2.2.10493.05608

Matsuoka, D., Nakano, M., Sugiyama, D., & Uchida, S. (2018). Deep learning approach for detecting tropical cyclones and their precursors in the simulation by a cloud-resolving global nonhydrostatic atmospheric model. *Progress in Earth and Planetary Science, 5*, 1–16.

McKee, T. B., Doesken, N. J., Kleist, J., & others. (1993). The relationship of drought frequency and duration to time scales. *Proceedings of the 8th Conference on Applied Climatology*, *17*, pp. 179–183.

Mokhtar, A., Jalali, M., He, H., Al-Ansari, N., Elbeltagi, A., Alsafadi, K., Rodrigo-Comino, J. (2021). Estimation of SPEI meteorological drought using machine learning algorithms. *IEEE Access, 9*, 65503–65523.

Moneta, A., Chlaß, N., Entner, D., & Hoyer, P. (2011). Causal search in structural vector autoregressive models. *NIPS Mini-Symposium on Causality in Time Series*, (pp. 95–114).

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Moradi Kordmahalleh, M., Gorji Sefidmazgi, M., & Homaifar, A. (2016). A sparse recurrent neural network for trajectory prediction of atlantic hurricanes. *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, (pp. 957–964).

Nath, S., Kotal, S. D., & Kundu, P. K. (2016). Seasonal prediction of tropical cyclone activity over the north Indian Ocean using three artificial neural networks. *Meteorology and Atmospheric Physics, 128*, 751–762.

Nethery, R. C., Katz-Christy, N., Kioumourtzoglou, M.-A., Parks, R. M., Schumacher, A., & Anderson, G. B. (2020). Integrated causal-predictive machine learning models for tropical cyclone epidemiology. *arXiv preprint arXiv:2010.11330*.

Nie, F., Huang, H., Cai, X., & Ding, C. (2010). Efficient and robust feature selection via joint ℓ2, 1-norms minimization. *Advances in neural information processing systems, 23*.

Noorbakhsh, M., Connaughton, C., & Rodrigues, F. A. (2020). Discovering causal factors of drought in Ethiopia. *Proceedings of the 10th International Conference on Climate Informatics*, (pp. 72–78).

Oettli, P., Nonaka, M., Richter, I., Koshiba, H., Tokiya, Y., Hoshino, I., & Behera, S. (2022). Combining dynamical and statistical modeling to improve the prediction of surface air temperatures 2 months in advance: A hybrid approach. *Frontiers in Climate, 4*.

Pan, B., Xu, X., & Shi, Z. (2019). Tropical cyclone intensity prediction based on recurrent neural networks. *Electronics Letters, 55*, 413–415.

Paniagua-Tineo, A., Salcedo-Sanz, S., Casanova-Mateo, C., Ortiz-García, E. G., Cony, M. A., & Hernández-Martín, E. (2011). Prediction of daily maximum temperature using a support vector regression algorithm. *Renewable Energy, 36*, 3054–3060.

Park, J., & Kim, J. (2018). Defining heatwave thresholds using an inductive machine learning approach. *Plos one, 13*, e0206872.

Park, M.-S., Kim, M., Lee, M.-I., Im, J., & Park, S. (2016). Detection of tropical cyclone genesis via quantitative satellite ocean surface wind pattern and intensity analyses using decision trees. *Remote sensing of environment, 183*, 205–214.

Park, S., Im, J., Jang, E., & Rhee, J. (2016). Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. *Agricultural and forest meteorology, 216*, 157–169.

Pasini, A., Racca, P., Amendola, S., Cartocci, G., & Cassardo, C. (2017). Attribution of recent temperature behaviour reassessed by a neural-network method. *Scientific reports, 7*, 1–10.

Pearl, J., & others. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress, 19*.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science, 2*, 559–572.

Pedro-Monzonís, M., Solera, A., Ferrer, J., Estrela, T., & Paredes-Arquiola, J. (2015). A review of water scarcity and drought indexes in water resources planning and management. *Journal of Hydrology, 527*, 482–493.

Peng, H., & Fan, Y. (2017). A general framework for sparsity regularized feature selection via iteratively reweighted least square minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31.*

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence, 27*, 1226–1238.

Peng, T., Zhi, X., Ji, Y., Ji, L., & Tian, Y. (2020). Prediction skill of extended range 2-m maximum air temperature probabilistic forecasts using machine learning post-processing methods. *Atmosphere, 11*, 823.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms.* The MIT Press.

Pfleiderer, P., Schleussner, C.-F., Geiger, T., & Kretschmer, M. (2020). Robust predictors for seasonal Atlantic hurricane activity identified with causal effect networks. *Weather and Climate Dynamics, 1*, 313–324.

Pillay, M. T., & Fitchett, J. M. (2021). On the conditions of formation of Southern Hemisphere tropical cyclones. *Weather and Climate Extremes, 34*, 100376.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR), 51*, 1–36.

Pradhan, R., Aygun, R. S., Maskey, M., Ramachandran, R., & Cecil, D. J. (2017). Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Transactions on Image Processing, 27*, 692–702.

Prodhan, F. A., Zhang, J., Sharma, T. P., Nanzad, L., Zhang, D., Seka, A. M., Mohana, H. P. (2022). Projection of future drought and its impact on simulated crop yield over South Asia using ensemble machine learning approach. *Science of The Total Environment, 807*, 151029.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*, 81–106.

Rüttgers, M., Lee, S., & You, D. (2018). Prediction of typhoon tracks using a generative adversarial network with observational and meteorological data. *arXiv preprint arXiv:1812.01943*.

Rüttgers, M., Lee, S., Jeon, S., & You, D. (2019). Prediction of a typhoon track using a generative adversarial network and satellite images. *Scientific reports, 9*, 1–15.

Racah, E., Beckham, C., Maharaj, T., Ebrahimi Kahou, S., Prabhat, M., & Pal, C. (2017). Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Advances in neural information processing systems, 30.*

Raducanu, B., & Dornaika, F. (2012). A supervised non-linear dimensionality reduction approach for manifold learning. *Pattern Recognition, 45*, 2432–2444.

Rahmati, O., Falah, F., Dayal, K. S., Deo, R. C., Mohammadi, F., Biggs, T., Bui, D. T. (2020). Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia. *Science of the Total Environment, 699*, 134230.

Rajsekhar, D., Singh, V. P., & Mishra, A. K. (2015). Integrated drought causality, hazard, and vulnerability assessment for future socioeconomic scenarios: An information theory perspective. *Journal of Geophysical Research: Atmospheres, 120*, 6346–6378.

Ramsay, H. A., Camargo, S. J., & Kim, D. (2012). Cluster analysis of tropical cyclone tracks in the Southern Hemisphere. *Climate dynamics, 39*, 897–917.

Ratnam, J. V., Behera, S. K., Ratna, S. B., Rajeevan, M., & Yamagata, T. (2016). Anatomy of Indian heatwaves. *Scientific reports, 6*, 1–11.

Reed, K. A., Stansfield, A. M., Wehner, M. F., & Zarzycki, C. M. (2020). Forecasted attribution of the human influence on Hurricane Florence. *Science advances, 6*, eaaw9253.

Rhee, J., & Im, J. (2017). Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data. *Agricultural and Forest Meteorology, 237*, 105–122.

Ribeiro, B., Vieira, A., & Carvalho das Neves, J. (2008). Supervised Isomap with dissimilarity measures in embedding learning. *Iberoamerican Congress on Pattern Recognition*, (pp. 389–396).

Richardson, T. S. (2013). A discovery algorithm for directed cyclic graphs. *arXiv preprint arXiv:1302.3599*.

Richman, M. B., & Leslie, L. M. (2012). Adaptive machine learning approaches to seasonal prediction of tropical cyclones. *Procedia Computer Science, 12*, 276–281.

Richman, M. B., & Leslie, L. M. (2018). The 2015-2017 Cape Town drought: Attribution and prediction using machine learning. *Procedia Computer Science, 140*, 248–257.

Richman, M. B., & Leslie, L. M. (2020). Machine Learning for Attribution of Heat and Drought in Southwestern Australia. *Procedia Computer Science, 168*, 3–10.

Richman, M. B., Leslie, L. M., Ramsay, H. A., & Klotzbach, P. J. (2017). Reducing tropical cyclone prediction errors using machine learning approaches. *Procedia computer science, 114*, 314–323.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning, 53*, 23–69.

Roodposhti, M. S., Safarrad, T., & Shahabi, H. (2017). Drought sensitivity mapping using two one-class support vector machine algorithms. *Atmospheric research, 193*, 73–82.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science, 290*, 2323–2326.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature, 323*, 533–536.

Runge, J. (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. *Conference on Uncertainty in Artificial Intelligence*, (pp. 1388–1397).

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., others. (2019). Inferring causation from time series in Earth system sciences. *Nature communications, 10*, 1–13.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances, 5*, eaau4996.

Russo, S., Sillmann, J., & Fischer, E. M. (2015). Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters, 10*, 124003.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*, e1249.

Saha, M., Soni, D., Finley, B., & Monteleoni, C. (n.d.). CAUSAL LINK DETECTION AND THE PREDICTION OF THE INDIAN SUMMER MONSOON.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers, 100*, 401–409.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85–117.

Schreiber, T. (2000). Measuring information transfer. *Physical review letters, 85*, 461.

Scott, E., & Crone, E. (2021, May). Using the right tool for the job: the difference between unsupervised and supervised analyses of multivariate ecological data. *Oecologia, 196*. doi:10.1007/s00442-020-04848-w

Seccia, R., Romano, S., Salvetti, M., Crisanti, A., Palagi, L., & Grassi, F. (2021, February). Machine Learning Use for Prognostic Purposes in Multiple Sclerosis. *Life, 11*, 122. doi:10.3390/life11020122

Shao, C., Ren, J., Wang, H., Jin, J., & Hu, S. (2016, September). Improving Machined Surface Variation Prediction by Integrating Multi-Task Gaussian Process Learning with Cutting Force Induced Surface Variation Modeling. *Journal of Manufacturing Science and Engineering, 139*. doi:10.1115/1.4034592

Shawe-Taylor, J., Cristianini, N., & others. (2004). *Kernel methods for pattern analysis.* Cambridge university press.

Shi, H., Zhao, Y., Liu, S., Cai, H., & Zhou, Z. (2022). A new perspective on drought propagation: Causality. *Geophysical Research Letters, 49*, e2021GL096758.

Shi, Y., Jin, N., Ma, X., Wu, B., He, Q., Yue, C., & Yu, Q. (2020). Attribution of climate and human activities to vegetation change in China using machine learning techniques. *Agricultural and Forest Meteorology, 294*, 108146.

Sorzano, C. O., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

Spinoni, J., Barbosa, P., De Jager, A., McCormick, N., Naumann, G., Vogt, J. V., Mazzeschi, M. (2019). A new global database of meteorological drought events from 1951 to 2016. *Journal of Hydrology: Regional Studies, 22*, 100593.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search.* MIT press.

Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V., & Wimberly, F. (2000). Constructing Bayesian network models of gene expression networks from microarray data.

Stillman, J. H. (2019). Heat waves, the new normal: summertime temperature extremes will impact animals, ecosystems, and human communities. *Physiology, 34*, 86–100.

Sun, X., Xie, L., Shah, S. U., & Shen, X. (2021). A Machine Learning Based Ensemble Forecasting Optimization Algorithm for Preseason Prediction of Atlantic Hurricane Activity. *Atmosphere, 12*, 522.

Sutanto, S. J., van der Weert, M., Wanders, N., Blauhut, V., & Van Lanen, H. A. (2019). Moving from drought hazard to impact forecasts. *Nature Communications, 10*, 1–7.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Sweet, L.-b., & Zscheischler, J. (2022). *Using interpretable machine learning to identify compound meteorological drivers of crop yield failure.* Tech. rep., Copernicus Meetings.

Tan, J., Chen, S., & Wang, J. (2021). Western North Pacific tropical cyclone track forecasts by a machine learning model. *Stochastic Environmental Research and Risk Assessment, 35*, 1113–1126.

Teh, Y., & Roweis, S. (2002). Automatic alignment of local representations. *Advances in neural information processing systems, 15*.

Tenenbaum, J. B., Silva, V. d., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science, 290*, 2319–2323.

Thrun, S., & O'Sullivan, J. (1996). Discovering structure in multiple learning tasks: The TC algorithm. *ICML*, *96*, pp. 489–497.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological review, 38*, 406.

Toreti, A., Cronie, O., & Zampieri, M. (2019). Concurrent climate extremes in the key wheat producing regions of the world. *Scientific reports, 9*, 1–8.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment, 8*, 127–150.

Tufaner, F., & Özbeyaz, A. (2020). Estimation and easy calculation of the Palmer Drought Severity Index from the meteorological data by using the advanced machine learning algorithms. *Environmental Monitoring and Assessment, 192*, 1–14.

Ulfarsson, M. O., & Solo, V. (2011). Vector l_0 Sparse Variable PCA. *IEEE Transactions on Signal Processing, 59*, 1949–1958.

Van Der Maaten, L., Postma, E., Van den Herik, J., & others. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res, 10*, 13.

Varando, G., Fernández-Torres, M.-A., & Camps-Valls, G. (2021). Learning Granger Causal Feature Representations. *AGU Fall Meeting 2021.*

Vicente-Serrano, S. M., Beguería, S., & López-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate, 23*, 1696–1718.

Vijverberg, S., & Coumou, D. (2022). The role of the Pacific Decadal Oscillation and ocean-atmosphere interactions in driving US temperature predictability. *npj Climate and Atmospheric Science, 5*, 1–11.

Wang, L., Zhao, Q., Gao, S., Zhang, W., & Feng, L. (2021). A new extreme detection method for remote compound extremes in southeast china. *Frontiers in Earth Science, 9*, 630192.

Wang, Y., Han, L., Lin, Y.-J., Shen, Y., & Zhang, W. (2018). A tropical cyclone similarity search algorithm based on deep learning method. *Atmospheric Research, 214*, 386–398.

Wang, Y., Zhang, W., & Fu, W. (2011). Back Propogation (BP)-neural network for tropical cyclone track forecast. *2011 19th International Conference on Geoinformatics*, (pp. 1–4).

Wehner, M. F., Zarzycki, C., & Patricola, C. (2019). Estimating the human influence on tropical cyclone intensity as the climate changes. In *Hurricane risk* (pp. 235–260). Springer.

Weinberger, K. Q., Sha, F., & Saul, L. K. (2004). Learning a kernel matrix for nonlinear dimensionality reduction. *Proceedings of the twenty-first international conference on Machine learning*, (p. 106).

Wendler-Bosco, V., & Nicholson, C. (2022). Modeling the economic impact of incoming tropical cyclones using machine learning. *Natural Hazards, 110*, 487–518.

Wijnands, J. S., Qian, G., & Kuleshov, Y. (2016). Variable selection for tropical cyclogenesis predictive modeling. *Monthly Weather Review, 144*, 4605–4619.

Wimmers, A., Velden, C., & Cossuth, J. H. (2019). Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Monthly Weather Review, 147*, 2261–2282.

Wu, Z., Yin, H., He, H., & Li, Y. (2022). Dynamic-LSTM hybrid models to improve seasonal drought predictions over China. *Journal of Hydrology, 615*, 128706. doi:https://doi.org/10.1016/j.jhydrol.2022.128706

Xie, L., Li, Z., Zhou, Y., He, Y., & Zhu, J. (2020, November). Computational Diagnostic Techniques for Electrocardiogram Signal Analysis. *Sensors*.

Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., & Wu, M. (2006). Supervised probabilistic principal component analysis. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 464–473).

Zaki, M. J., Meira Jr, W., & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms.* Cambridge University Press.

Zampieri, M., Ceglar, A., Dentener, F. J., & Toreti, A. (2017). Wheat yield loss attributable to heat waves, drought and water excess at the global, national and subnational scales. *Environmental Research Letters, 12*.

Zaniolo, M., Giuliani, M., & Castelletti, A. (2019). Data-driven modeling and control of droughts. *IFAC-PapersOnLine, 52*, 54–60.

Zaniolo, M., Giuliani, M., Castelletti, A. F., & Pulido-Velazquez, M. (2018). Automatic design of basin-specific drought indexes for highly regulated water systems. *Hydrology and Earth System Sciences, 22*, 2409–2424.

Zhang, B., Abu Salem, F. K., Hayes, M., & Tadesse, T. (2020). Quantitative Assessment of Drought Impacts Using XGBoost based on the Drought Impact Reporter. *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning.* Retrieved from https://www.climatechange.ai/papers/neurips2020/18

Zhang, R., Chen, Z.-Y., Xu, L.-J., & Ou, C.-Q. (2019). Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China. *Science of the Total Environment, 665*, 338–346.

Zhang, S., Ma, M., Li, M., Chen, J., & Bai, C. (2022). The role of Atlantic variability in modulating the tropical cyclone formation in the Australian region.

Zhang, S.-q. (2009). Enhanced supervised locally linear embedding. *Pattern Recognition Letters, 30*, 1208–1218.

Zhang, T., Lin, W., Lin, Y., Zhang, M., Yu, H., Cao, K., & Xue, W. (2019). Prediction of tropical cyclone genesis from mesoscale convective systems using machine learning. *Weather and Forecasting, 34*, 1035–1049.

Zhang, W., Fu, B., Peng, M. S., & Li, T. (2015). Discriminating developing versus nondeveloping tropical disturbances in the western North Pacific through decision tree analysis. *Weather and Forecasting, 30*, 446–454.

Zhang, W., Leung, Y., & Chan, J. C. (2013). The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part I: Tropical cyclone recurvature. *Journal of applied meteorology and climatology, 52*, 1394–1416.

Zhang, W., Murakami, H., Khouakhi, A., & Luo, M. (2021). Compound Climate Extremes in the Present and Future Climates: Machine Learning, Statistical Methods and Dynamical Modelling. *Compound Climate Extremes in the Present and Future Climates: Machine Learning, Statistical Methods and Dynamical Modelling, 9*, 807224. Frontiers Media SA.

Zhang, X., Chen, G., Cai, L., Jiao, H., Hua, J., Luo, X., & Wei, X. (2021). Impact assessments of Typhoon Lekima on forest damages in subtropical China using machine learning methods and landsat 8 OLI imagery. *Sustainability, 13*, 4893.

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, Y., Zhang, Z., Qin, J., Zhang, L., Li, B., & Li, F. (2018). Semi-supervised local multi-manifold isomap by linear embedding for feature extraction. *Pattern Recognition, 76*, 662–678.

Zhang, Z., & Zha, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing, 26*, 313–338.

Zhao, Z., & Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. *Proceedings of the 24th international conference on Machine learning*, (pp. 1151–1157).

Zscheischler, J., Westra, S., Van Den Hurk, B. J., Seneviratne, S. I., Ward, P. J., Pitman, A., . . . others. (2018). Future climate risk from compound events. *Nature Climate Change, 8*, 469–477.

This project is part of the H2020 Programme supported by the European Union, having received funding from it under Grant Agreement No 101003876