Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Bachelor Thesis Mediainformatics

Machine Learning Methods to Model Monthly Precipitation over the Western Mediterranean Using Climate Indices

Davide Lussu

11th July 2022

Reviewer

Dr. Bedartha Goswami
Machine Learning in Climate Science
Cluster of Excellence "Machine Learning"
University of Tübingen

Lussu, Davide:
Machine Learning Methods to Model Monthly Precipitation over the Western Mediterranean Using Climate Indices
Bachelor Thesis Mediainformatics
Eberhard Karls Universität Tübingen
Completion period: 11th April 2022 - 11th July 2022

Machine Learning Methods to Model Monthly Precipitation over the Western Mediterranean Using Climate Indices

Davide Lussu
University of Tübingen
davide.lussu@student.uni-tuebingen.de

Dr. Bedartha Goswami
University of Tübingen
Maria-von-Linden-Str. 6
Tübingen, Germany, 72076
bedartha.goswami@uni-tuebingen.de

ABSTRACT

This study compares various models estimating monthly deseasoned precipitation based on large-scale atmospheric oscillation indices as predictors. The study area selected is the western Mediterranean and 9 climatic indices which are influential in the area are considered. Regression is performed on the time series of the climatic indices to model (gridded) precipitation time series provided by the ERA5 monthly precipitation dataset spanning over the past 40 years. The methods used are Multiple Linear Regression (MLR), Random Forest Regression (RF), M5 Model Tree (M5) and Artificial Neural Networks (ANN). The focus lies on assessing the models' abilities to reproduce the teleconnections. The models are first tested on the entire study area making predictions with all considered input variables. In a second part of the study, they are then further tested using only a selection of input variables and on seasonally separated data. The analysis highlights the conditions under which each model performs best. The results suggest that the most basic regression method MLR remains a valid option for the defined task whereas ANN shows a greater potential in modeling seasonally separated data. 3

1. INTRODUCTION

Analysing precipitation in the Mediterranean is of particular interest as this region has previously been identified as a 'hot spot' of climate change (Diffenbaugh and Giorgi, 2012 [10]). A substantial change towards higher temperatures and lower precipitation rates has long been observed and even higher degrees of change are projected by climate models for the future of the region (Seager et al., 2014 [25]). Especially the intensification of dry climate conditions poses large threats, ranging from increases in droughts and aridity to increases in fire weather which afflict ecosystems and a wide range of sectors including agriculture, forestry, and health (IPCC, 2021 [3]). On the other hand, a "paradoxical" increase in daily precipitation extremes has also been detected for the region (Alpert, P. et al., 2002 [2]). The projected threatening impact of climate change for the Mediterranean motivates the choice of the region in this study.

The Mediterranean Basin has given name to its type of climate known as the Mediterranean climate. Midlatitude and tropical atmospheric circulation patterns have their impact on its climatic conditions. Furthermore, this climate is influenced by the complex geomorphology of the area characterized by mountain ridges around the sea basin and small gulfs resulting in a great spatial variability. The winter season is primarily characterized by a mild and wet climate with the highest rainfall, whereas the summer season is hot and dry. This high precipitation variability across the seasons leads to a variety of weather conditions that can also turn into extremes. On one hand, there are dry events which cause vegetation stress, wildfires and water scarcity that afflicts regions with limited water resources. On the other hand, there is extreme precipitation which causes floods and erosion.

Given the high relevance of precipitation in the Mediterranean, a better understanding of its behaviour is desired. Being able to better predict its trend in the near and far future is important for the development of fields such as water management to ensure water and food security. Moreover, it is of interest to see how much of the change is to be attributed to changes in natural climatic oscillations and how much to the anthropogenic climate change. This requires climate simulations to accurately reproduce the main mechanisms controlling precipitation. One of such mechanisms is the connection to large-scale atmospheric circulation patterns. It is well established that said patterns have a major effect on regional hydroclimate in general and on that of the Mediterranean region. Their impact differs both spatially within the region and temporally across seasons.

The most prominent pattern in relation to the Mediterranean is the North Atlantic Oscillation (NAO). Its effect on precipitation, especially during the winter season, has been analysed in many previous studies.

However, the number of climatic patterns having their influence on Mediterranean climate goes far beyond that. In alignment with previous studies on the Mediterranean hydroclimate, the following patterns are regarded in this study: NAO, Arctic Oscillation (AO), Scandinavian Pattern (SCAND), East Atlantic pattern (EA), East Atlantic/Western Russia pattern (EAWR), Southern Oscillation index (SOI), Atlantic Multidecadal Oscillation index (AMO), Mediterranean Oscillation Index (MOI) and the Western Mediterranean Oscillation Index (WeMO).

Previous studies have utilized and compared a variety of regression methods for applications such as analyses on the teleconnec-

tions between climatic patterns and precipitation or the prediction of future precipitation based on climatic indices. These methods include Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average, Support Vector Machines, Artificial Neural Networks and Regression Trees such as Random Forests or the M5 Model Tree.
MLR is the most basic method used in this context. For example Choubin et al. (2016 [7]) make use of MLR in a comparison with multilayer perceptrons (MLP) and adaptive neuro fuzzy inference system models for predicting precipitation in Iran from various climate signals. They conclude that MLR performs worse than their implementation of an MLP.
Artificial neural networks (ANN) became popular as self-learning models that are capable of utilizing highly nonlinear data to make predictions. They have also been used in recent literature related to precipitation prediction and have proven to be effective in the task. Similar to this study, Choubin et al. (2017b [6]) applied ANNs among other methods to forecast seasonal precipitation time series. Rezaeian-Zadeh et al. (2012 [21]) compared four different ANN algorithms to predict monthly discharge volume in Iran. Hong et al. (2020 [13]) did a comparative analysis of inflow prediction through various algorithms resulting in their MLP implementation to be the best performing.
Tree regressors come in various implementations such as random forest regressors (RF) or M5 model trees (M5). These tree-based models have been implemented in various studies such as Sattari et al. (2020 [24]) to assess precipitation in Iran, or Ravinesh et al. (2017 [9]) to forecast drought in Australia, both based on climate indices as predictor variables. Many of these studies applied the models on station data and only considered small regions. Following preceding comparative studies in similar applications, we choose to investigate MLR, RF, M5 and MLP implementations for comparison in this study.
The objectives of this study are (1) to assess the comparative potential of a selection of models to estimate deseasoned precipitation in the western Mediterranean based on multiple large-scale climate signals and (2) to find the conditions under which they are optimally applied by tuning their hyperparameters and by exploring different applications of processing the training data. By doing so, the study intends to give a better insight into which methods and materials are best used for studies aimed at understanding and quantifying the relationships between precipitation trends and large-scale atmospheric variability and studies in the field of regional hydroclimate analysis.
This study is structured as follows: section 2 describes the study area and climatic data and the sources used; section 3 introduces the tools, models and evaluation methods used to perform the regression task; section 4 presents the applications used to test the methods in detail and comments on the results; section 5 summarizes and discusses the results and potential flaws.

2. STUDY AREA AND DATA

2.1 Climatic Indices

The data used for the regression models are various climate oscillation indices that describe pressure differences (with exception to AMO). In most cases they are calculated by subtracting the atmospheric pressure at one location from that at another. The time series data of the indices are provided in monthly temporal resolution, otherwise the monthly mean is taken. In the following, we introduce the indices that we presume to have a relevant impact on Mediterranean rainfall. The choices are based on previous studies on precipitation teleconnection patterns in the Mediterranean, such as that of Krichak et al. (2014 [16]), who studied the relationship between five teleconnection patterns and extreme precipitation in the area defined as "Euro-Mediterranean region" and Mathbout et al. (2019 [18]), who studied the relationship of eight such patterns and daily rainfall concentration over the Mediterranean.
The following indices are provided by the NOAA Climate Prediction Center:

- The Arctic Oscillation (AO; Thompson & Wallace, 1998 [27]) index tracks differences between sea-level pressure (SLP) anomalies in the arctic and anomalies in the mid-latitudes. The AO has its influence on mid-latitude climate by controlling the jet stream which can carry cold arctic air southwards. It is correlated with the NAO but has a different impact on Mediterranean precipitation (Krichak et al., 2014 [16]).
- The Scandinavian pattern (SCAND) index measures the pressure difference between northern and southern Europe. More precisely, the index provided by NOAA CPC has its northern action centre over Scandinavia and the southern centre spanning from western/southern Europe to eastern Russia/western Mongolia (Barnston and Livezey, 1987 [4]). A positive state of SCAND can bring above average precipitation in southern Europe.
- The East Atlantic (EA; Barnston and Livezey, 1987 [4]) pattern has its centres near 55°N, 20-35°W and 25-35°N, 0-10°W. It thus can be described as a southeastwards shifted NAO. Its key difference lies in the southern centre having a subtropical link, making it a unique index. It has an influence on precipitation in countries in the Mediterranean, such as Spain, as found by Rodriguez-Puebla et al. (1998 [22]).
- The East Atlantic/Western Russia (EAWR; Barnston and Livezey, 1987 [4]) pattern consists of four geopotential height anomaly centres over Europe, northern China, the central North Atlantic and northern Caspian sea. Krichak (2005 [15]) found its correlations with precipitation to be most significant over the eastern Atlantic and the south-eastern Mediterranean.
- The Southern Oscillation Index (SOI; Ropelewski and Jones, 1987 [23]) measures pressure differences between the western and eastern tropical Pacific by comparing the locations Tahiti and Darwin. In this study, we use the standardized data provided by NOAA (sometimes denoted soi_std).
- Atlantic Multidecadal Oscillation (AMO; Enfield et al., 2001 [11]) is the only index used in this study based on sea surface temperature anomalies. It has been included to provide an index with multidecadal periodicity and due to its effect on precipitation in the northern hemisphere, including Europe. We use the unsmoothed variant provided by the NOAA (sometimes denoted amo_us).

The remaining indices are provided by the Climatic Research Unit, University of East Anglia (CRU) unless stated differently:

- The North Atlantic Oscillation (NAO) is one of the most prominent modes influencing Mediterranean precipitation. It shows particularly interesting effects on European precipitation during winter. It is defined as the pressure difference between the Icelandic Low and the Azores High. The differences between these two points control the strength and direction of westerly winds which bring air moisture to the European region resulting in precipitation, especially along western coasts. The index provided by the CRU uses data from Gibraltar and Southwest Iceland (Reykjavik) calculated after Jones et al. (1997 [14]).

- The Mediterranean Oscillation Index (MOI, Conte et al., 1989 [8]) represents an atmospheric circulation more local to the Mediterranean. Out of the two definitions existing for this index, the one derived from the SLP differences between Algiers and Cairo is chosen in this study.

- Another circulation pattern with great influence in the region is the Western Mediterranean Oscillation (WeMO Martin-Vide and Lopez-Bustins, 2006 [17]). It is an index that has been proposed relatively recently with the purpose of being a more local teleconnection for the Mediterranean region in contrast to the NAO. It does so by comparing the pressure at Padua in northern Italy with that at San Fernando in southwestern Spain. It is provided by the Climatology Group of the University of Barcelona.

2.2 Precipitation data

Monthly precipitation is the target variable for the regression models. We use the total precipitation variable provided in the ERA5 dataset of monthly averaged data on single levels from 1979 to present (Hersbach, H. et al., 2019 [12]). The dataset has a spatial grid resolution of $0.25° \times 0.25°$ $(latitude \times longitude)$ and provides a time series of monthly averaged daily precipitation for each grid point. The study area covers the latitudes from 35°N to 47°N and longitudes from 10°W to 30°E. We only focus on the western Mediterranean as the climatic conditions over the whole Mediterranean region are vastly different and considering them all would go beyond the scope of this study. The unit of the dataset is metres of water depth. Every time series is deseasoned using moving averages (only trend + residuals are used). To get a picture of the dimensions of the study area and the spatial distribution of the rainfall features, we provide a map displaying the mean precipitation over the past 40 years in the supplementary section (see fig. 11).
In order to better differentiate between regional differences within the Mediterranean region, 6 smaller regions of dimensions $2° \times 2°$ with unique climatic conditions were selected. The selection of the regions is based on the overall mean precipitation as well as the correlation strength of the indices with precipitation at that location. A map that highlights the climatic index with the highest correlation at each grid point is shown in the appendix (see fig. 12). Five indices, namely NAO, SCAND, WeMO, AO and EA have large regions, in which they are the most dominant (indices with very small areas of influence are not shown). The selections are made such that each of the five dominant indices in the Mediterranean region has an associated small region. The regions also have distinct precipitation characteristics regarding the mean taken over a timespan of 35 years (length of the test set; see fig. 11) which are calculated from the ERA5 dataset for documentation in supplementary table 8. Northern Algeria and Eastern Spain experience relatively little precipitation of less than 2mm/day while the Balkans region has rather high precipitation of more than 3mm/day and Southern France is in between both.

3. METHODS

This section provides an overview of the methods and regression models used for the comparison. When the models are applied on a map, they are trained and evaluated on the single time series at each grid point separately, unless stated differently. The climatic indices are the input variables and the precipitation at the grid point is the target variable.

3.1 Spearman's Rank correlation

The Spearman's rank correlation coefficient ρ is used here to calculate the correlations between the time series of climatic indices and time series of precipitation. It is employed to rank the indices at each grid point by their significance at that location. Instead of calculating the correlation on absolute values, it ranks the data and calculates the Pearson correlation on the rank values. This enables it to detect nonlinear relationships which are of interest when employing nonlinear regression models.

3.2 Multiple Linear Regression

Multiple Linear Regression (MLR, labelled LR in the graphics and tables) is a common statistical method that uses multiple input variables $(X_1, X_2, \ldots, X_n)$ to predict a variable Y. The MLR equation for T observations is defined as follows:

$$y_t = a + b_1 x_{t1} + b_2 x_{t2} + \ldots + b_n x_{tn}$$

Where y_t is the t^{th} variable that is to be predicted using n independent variables x, b_n denotes the coefficients that control how much each independent variable x contributes to the predicted variable and a is the intercept of the regression line and the Y-axis. The vector b_n is the parameter that is to be estimated to fit the regression. In this study, the predictions of MLR serve as a baseline that will be compared to the other regression models.

3.3 Random Forest

Random Forests (RF) were introduced by Breiman in 2001 [5]. RF is a supervised learning method that makes use of ensemble learning to perform classification or, in this case, regression. Since single decision trees tend to overfit when grown too deep, it is preferable to use ensembles (forests) of decision trees to average their predictions, resulting in a performance boost for the final model. For each tree in the forest, a training subset is chosen out of the training set using the bagging technique. This technique selects random samples with replacement out of the training set (meaning that a sample taken for a training subset can still be used for other training subsets). An individual decision tree is grown with each subset. This is repeated until a predefined number of trees are grown. The final predictions of RF are made by taking the average prediction of all its trees. The RF provided by scikit-learn allows us to adjust the number of estimators, i.e., the number of trees in the forest.

3.4 M5

The M5 Model Tree was introduced by Quinlan (1992, [20]). It is a decision tree learner aimed at predicting continuous numerical values by fitting linear regression functions at its leaf nodes. In the first construction stage, it is built like a decision tree using the divide-and-conquer method and then pruned in the second stage. During the building stage, the training set is split into subsets based on a splitting criterion. This creates two new nodes that can either be split into more subsets using the same procedure or end up as leaf nodes. This process is repeated recursively until termination when only leaf nodes remain at the end of the branches. The used splitting criterion is the standard deviation reduction (SDR) which uses the standard deviation as an error measure. During the building process, SDR calculates the standard deviation $S(T)$ of the value subset T in the currently processed node and subtracts the standard deviation of each possible split $c \in X$ executed on that subset $S(T, X)$. The SDR calculated as follows:

$$SDR(T, X) = S(T) - S(T, X)$$

where

$$S(T, X) = \sum_{c \in X} P(c)S(c)$$

The split that results in the largest SDR (i.e., the largest expected error reduction) is applied with the purpose of ultimately reducing the prediction error at the leaf nodes. If a subset contains too few values, the building process terminates at that branch and the node is declared a leaf/terminal node. In the second stage, the branches that do not contribute much to the overall reduction of the prediction error are pruned (i.e., replaced by a leaf node). Finally, linear regression is applied on the values of the leaf nodes and smoothing is applied to smooth out the sharp discontinuities between the linear equations of neighbouring nodes.

3.5 ANN/MLP

Multilayer Perceptrons (MLP) are a common class of artificial neural networks (ANN). They consist of at least 3 layers of nodes, the first one being the input layer, the middle ones being at least one hidden layer and the last one being the output layer. The layers consist of nodes and are connected by joints with trainable weights. MLPs are feedforward networks, meaning that the nodes of each layer receive inputs from only the previous layer. The networks are trained using backpropagation, where the weights of the joints between the nodes are repeatedly adjusted to reduce the mean squared error of the output vector generated by the network (i.e., the predicted precipitation) as compared to the actual values (the observed precipitation).
There are several hyperparameters that can be adjusted to find an optimal model. We consider different numbers of layers and nodes as well as different learning rates and batch sizes. Regarding the network shape, we will consistently use a densely connected MLP shape with equally sized hidden layers, each using the Rectified Linear Unit (ReLU) as nonlinear activation function. During training, early stopping regularization is used to prevent overfitting. Its implementation in this study monitors the models' loss on a validation set after each training epoch and stops training after there has been no improvement in the loss for a previously defined number of epochs (called patience). The search for the optimal hyperparameters is part of the results section. We test two implementations of MLPs that process the training data in different ways. The first implementation (labelled NN1) follows the same procedure as the previously presented regression methods, where one MLP is trained on each time series on the respective grid point, consequently making estimations on one precipitation value at a time. The second implementation (labelled NN2) exploits the ANNs' ability to predict higher dimensional target values. This alternative implementation uses just one MLP to predict precipitation for an entire region at once. The idea here is to see whether the model can learn the climatic characteristics of entire regions and identify the distribution patterns of their precipitation and yield overall better results. However, this comes with the trade-off of needing to predict a target vector of increased size (depending on the chosen extract of the map). The increased target vector length itself poses a bottleneck which is the area size that the model can predict at once. The entire selected study area consists of 23.520 grid points which would be a too long target vector to predict from the relatively little given data. This implementation also brings the advantage of a reduction of the high computational cost of training a new MLP for every single time series. To implement this concept, the target vector at a given time contains every target data point in the selected region. A single target vector Y_t of one month t is created by flattening the 2-dimensional spatial data of a selected area of the map M_t with dimensions $M_{lat} \times M_{lon}$ into a 1-dimensional vector of length $|Y_t| = |M_{lat}| * |M_{lon}|$.
Each approach will have its individual ANN with unique parameters that have to be optimized.

3.6 Model evaluation criteria

The models' performances will primarily be evaluated using the coefficient of determination, denoted as the r^2 score. It measures the goodness of fit, that is, the ability of regression methods to approximate the observed data. The score ranges from 0 to 1 where a score of 1 would indicate a perfect fit. It indicates the proportion (percentage) of the variance of the target data that can be explained by the predictor variables of the model. r^2 is calculated as:

$$r^2 = 1–(RSS/TSS)$$

where RSS is the sum of squares of residuals (i.e., the squared differences of the predicted values and the observed values) and TSS is the total sum of squares (i.e., the squared differences between the observed values the and their mean).

3.7 Schematic overview

An overview of the data pipeline and model implementations is given in figure 1.

3.8 Tools used

The following libraries were used for the regression methods: scikit-learn for the LR and RF methods [19], a publicly available implementation of M5 [26], and TensorFlow for the implementation of the neural networks [1].

3.9 GitHub code repository

The code of the project is available at the github repository at `https://github.com/Staubsaugerbeutel/BA-Git/tree/main/code`

4. APPLICATION & RESULTS

The results section is structured as follows: in part 4.1, a generic use case is described that is created to compare all models under identical conditions, followed by the hyperparameter searches for each of the models which are then applied to the use case. In part 4.2 the models are first tested using less predictors and then in a seasonal approach.
The shown boxplots are intended to better visualize the performance of the models when evaluated on an entire map. They show the distribution of r^2 scores across all the grid points of a map. The green triangle indicates the mean performance of all its grid points and the orange line indicates the median.

4.1 Part 1: Application on a universal use case

In this section, the objective as defined in (1), to assess the selected models to estimate precipitation based on large-scale climate signals, is pursued. A generic use case is designed to compare the models under equal conditions. First, LR is run on this use case and its results serve as baseline. Then the optimal hyperparameters for the other regression models are determined and applied to this setup and later compared.

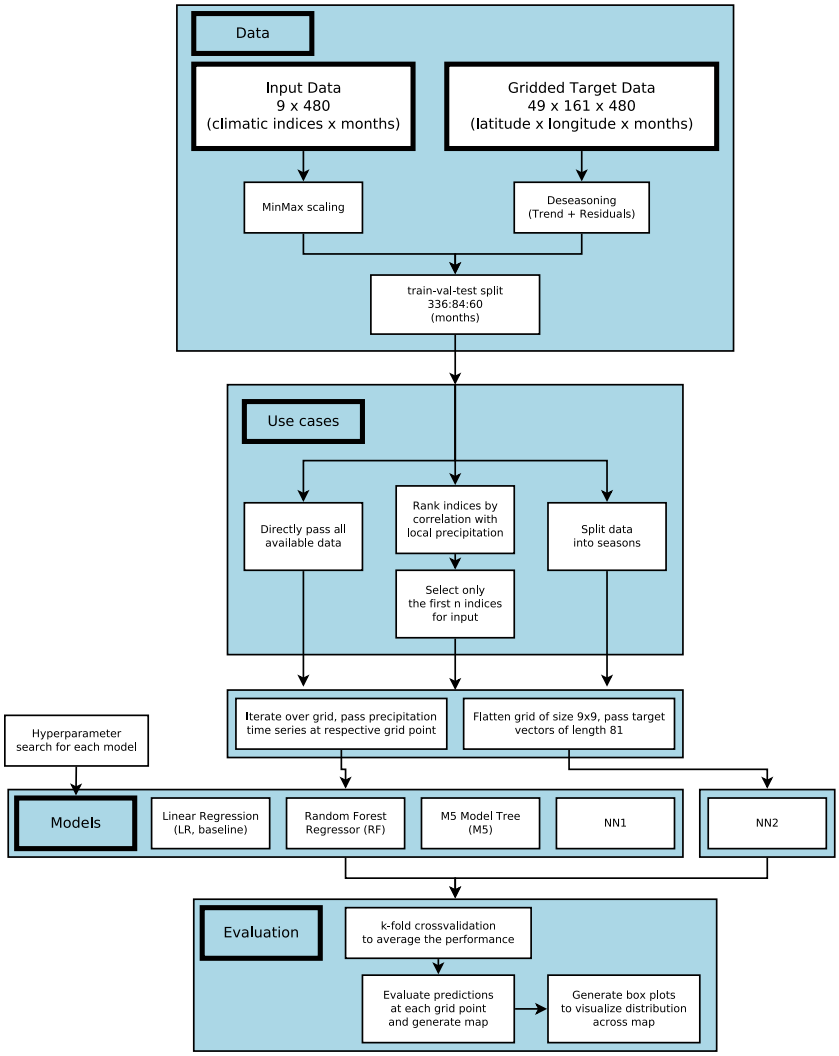

Fig. 1. Schematic overview of the project.

4.1.1 Use Case 1. Here we define the use case data, which is intended to be most universal for the Mediterranean region. For these tests, all 9 indices chosen as predictors are included in the input data. Furthermore, the whole geographic study area is being considered, i.e., the operations are performed on the time series on every grid point of the precipitation dataset.
A train-validation-test split is applied on all timeseries of the input and target data. First, the test set containing the last five years is split off the time series which corresponds to $1/8^{th}$ of the entire timespan. The test set spans the 5 years 2014-07 to 2019-06 while the remaining data spans from 1979-07 to 2014-06. This temporally separate test set is created to be able to judge how well the model generalizes "truly new" values. It is also an identical test set for all models, as opposed to the randomly shuffled validation set which is described in the following. Additionally, a validation split is created which picks samples at random times from the remaining data, generating a split of 20%. Table 1 gives an overview of the dimensions of the data for this test run.
The models are run using 5-fold crossvalidation, maintaining the validation split proportion of 20%. Using cross validation comes at a slight sacrifice of having to fit a model 5 times but with the gain of using all data points for training and of averaging the results of different validation sets each time, leading to more robust evaluation results.

Table 1. Input and target data

	Input data $(predictors \times months)$	Target data $(lat \times lon \times months)$
Training set	9×336	$49 \times 161 \times 336$
Validation set	9×84	$49 \times 161 \times 84$
Test set	9×60	$49 \times 161 \times 60$

Train-Test-Validation data split dimensions for input and target data.

4.1.2 Creating the MLR baseline. MLR needs no parameter tuning and is directly applied on time series at every grid point. The first row of fig. 6 visualizes the r^2 score of the MLR algorithm at each grid point of the validation and test set. The visualized scores

are the means taken over the 5 cross validation folds. A comparison with the mean precipitation in fig. 11 hints that the regression works particularly well in regions with higher precipitation such as the western coast of the Balkans or the western half of the Italian Peninsula. The results are further discussed in the comparison in section *4.1.8*.

4.1.3 Finding optimal parameters for the remaining models. In the following, the parameter searches for the other regression methods are shown. Because executing the algorithms multiple times with different parameters and on the entire map is computationally expensive and would take too long on the entire map, the parameter search is executed on the 6 regions specified in section 2.2. The multiple regions are also chosen to prevent a choice of parameters that is biased to the climatic conditions of just one region.

4.1.4 RF. The parameter of interest for Random Forests is the number of estimators or trees in the forest. The range of estimators that was tested for the parameter search is

$$n_{estimators} \in \{5, 10, 50, 100, 500\}$$

The validation set results of the search in the 6 regions are displayed in the boxplot charts in fig. 2. The number of estimators in-

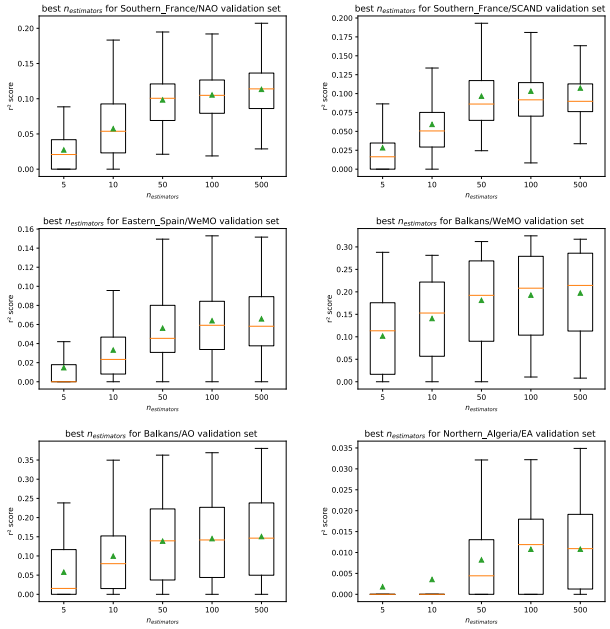

Fig. 2. RF parameter search for $n_{estimators}$ at each location. A higher number of estimators (trees in the forest) generally brings better performance stagnating at more than 50 estimators.

creases with each boxplot along the x-axis. It can be observed that an increase of the number of estimators also leads to an increase in performance. In better performing regions such as Balkans/AO, the improvement by adding more estimators starts to stagnate for more than 50 trees. On the other hand, in less well performing regions such as Eastern Spain/WeMO, increasing the number of estimators to over 50 still shows a relative increase. A drawback of having more trees is the increase of fit time per time series. On the tested system, the fit time increases by a factor of more than 10 for a model with 100 trees compared to a model with only 5 trees. Judging from the result of the parameter search in harder to predict regions such as Spain, we choose 100 estimators to perform the model on the entire map for later comparison with the other models.

4.1.5 M5. The parameter searched for in the case of M5 is its smoothing constant k. We iterate over the following smoothing constants:

$$k \in \{5, 10, 50, 100, 500, 1000\}$$

Increasing the parameter generally leads to better performance

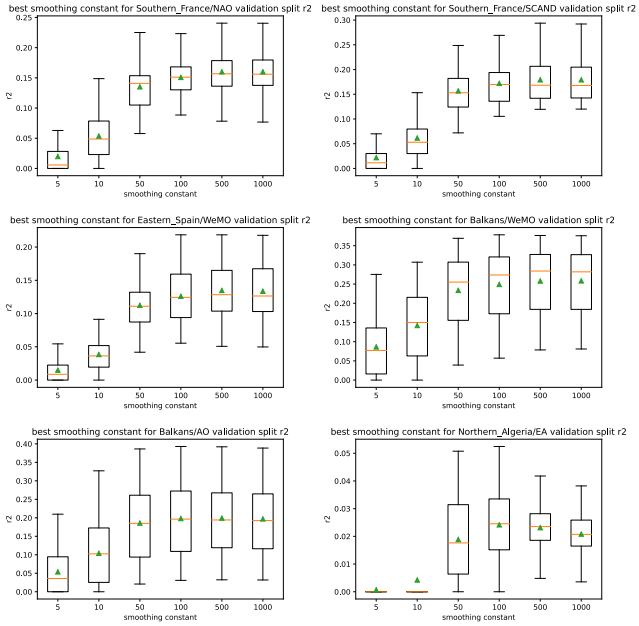

Fig. 3. M5 parameter search for best smoothing constant at each location. Smoothing constants beyond 100 do not bring much improvement.

of the model up to a certain point (see fig. 3). With k increasing from 100 to 500, not much improvement is seen anymore. Since increasing the parameter does not increase the fit time, we choose a smoothing constant of $k = 500$ for best performance in regions such as Southern France, where there is still a marginal improvement visible.

4.1.6 NN1. Since MLPs are the most computationally expensive compared to the other models, grid search and manual parameter search over a wider range of parameters was first performed on just a few single time series at chosen locations to get a rough picture of the space of parameters that works well. Parameters in the following space were tested:

- layers: $l \in [1, 20]$
- nodes: $n \in [9, 2000]$
- learning rate: $lr \in [0.001, 0.00003]$
- batch size: $bs \in [8, 128]$

Based on the results of the grid search, a set of 7 parameters is chosen to be compared on the extracts as demonstrated in the previous 2 methods. The parameter sets are shown in table 2. The parameters are chosen so that the shape of the model changes from long

Table 2. Selected parameter sets for NN1

Nodes per layer	Layers	Total parameters	Fit time (s)	multiple of LR fit times
1000	7	7,018,001	11.5	7187.5
500	7	1,759,001	5.5	3437.5
500	8	2,009,501	5.9	3687.5
500	10	2,510,501	9.3	5812.5
100	10	102,101	1.3	812.5
50	13	33,701	1.5	937.5
25	15	10,026	1.5	937.5

Nodes per hidden layer and fit times per time series. The model was trained on an Intel i7 CPU. "Multiple of LR fit times" are relative to the duration of an LR fit of $0.0016s$.

(many layers, few nodes) to wide (many nodes, less layers). Models with parameters lower than the shown ones mostly turned out not to be working at all. The fit parameters used are a learning rate of $lr = 0.0003$ and a batch size of $bs = 16$. Each model was trained until early stopping, using a patience of 20. The models displayed in the boxplots in fig. 4 increase in width and decrease in length from left to right. The models are labelled according to their number of layers and nodes. A peak is reached for the model with

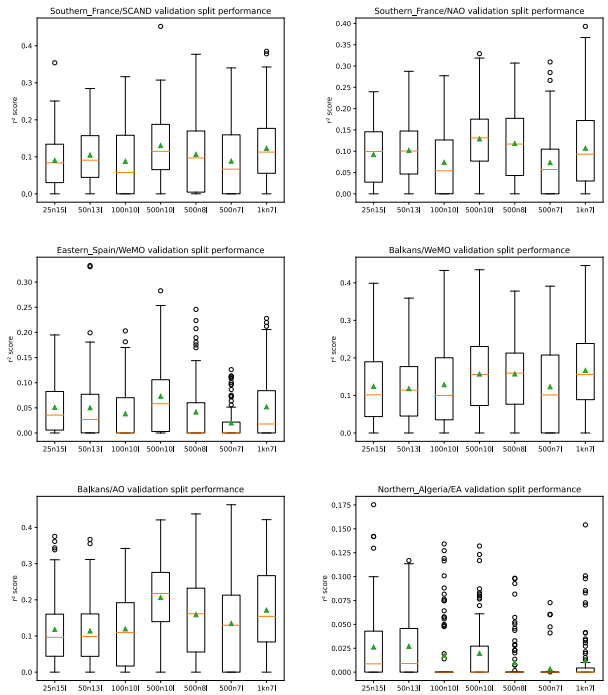

Fig. 4. NN1 hyperparameter search for 7 parameter sets of different layer (l) and node (n) combinations. The model with $500n$ and $10l$ peaks at most of the locations regarding its mean taken across the entire location.

500 nodes and 10 layers which exhibits the best performance on the validation split across almost all locations. From there on, the number of layers is further decreased while maintaining 500 nodes. This decrease is reflected in a decrease in the mean performance. The only exception to this decrease are the maximum reached performances of $500n7l$ at Balkans/AO and $500n8l$ at Southern France/SCAND which reach higher than those of $500n10l$, while their mean still sees a decrease. Doubling the number of nodes to 1000 for the model with 7 layers (seen as the right most boxplot in the figures) leads to an increase in performance that reaches up to means near the $500n10l$ model for some locations but is still worse than its $500n10l$ counterpart in Southern France/NAO, Eastern Spain/WeMO and Balkans/AO on the validation set. Interestingly, the smallest two models are the only ones that are able to explain some variance of the driest, and hardest to predict region of Northern Algeria, albeit with scores in very poor ranges.
For comparison with the other methods, we chose the MLP with $n = 500$ nodes and $l = 10$ layers as it provides the best results. Unlike the other models, the NN1 method is not cross validated on the entire study area and training is only run once per time series due to it being too computationally expensive at this size.

4.1.7 NN2. Besides the MLP implementation presented in the previous section, we provide an alternative implementation, which uses just one MLP to predict precipitation of all grid points of an entire region at once. Similar to the parameter search in the NN1 section, a set of parameter combinations is chosen from a wider grid search and then tested on all the regions to be compared with one another. The parameter sets and their approximate fit times (dependent on the system) are shown in table 3 and their performances in fig. 5. A batch size of $bs = 16$ and a learning rate of $lr = 0.0003$ is used for all training procedures. Each model was trained until early stopping, using a patience of 20.

Table 3. Selected parameter sets for NN2

Nodes per layer	Layers	Total parameters	Fit time (s)	multiple of LR fit times
2000	13	52,048,001	3.5	2187.5
1000	15	15,026,001	1.5	937.5
1000	13	13,024,001	0.9	562.5
1000	10	10,021,001	0.7	437.5
500	13	3,262,001	0.4	250
500	10	2,510,501	0.2	125

Nodes per hidden layer and fit times per time series. The fit times are calculated by executing the model on a 9×9 grid and dividing the total fit time by the number of time series included in that grid (81). The model was trained on an Intel i7 CPU. "Multiple of LR fit times" are relative to the duration of an LR fit of $0.0016s$.

As derived from the boxplots in fig. 5, the best predictions are mostly made by the model with 10 layers and 1000 nodes and the model with 13 layers and 1000 nodes. An exception to this being the parameter search at Southern France/SCAND, where this pattern is inverted.
For comparison with the other methods, we chose the relatively large MLP with 1000 nodes and 13 layers. To run the model on the entire study area, it is subdivided into regions with dimensions $2^\circ \times 2^\circ$, resulting in a total of 113 regions (excluding the maritime ones) and then one model is applied to each region. Thanks to the much shorter fit times than NN1, cross validation can be applied here too.

4.1.8 Comparison of the models' performance on the universal use case. After having optimized the hyperparameters for each model, these are applied on the entire study area to be compared with one another. The different models' performances on the time series at each grid point of the landmass are displayed for the validation and test sets separately on the maps in fig. 6. boxplots which describe the distribution of scores across the grid points on the

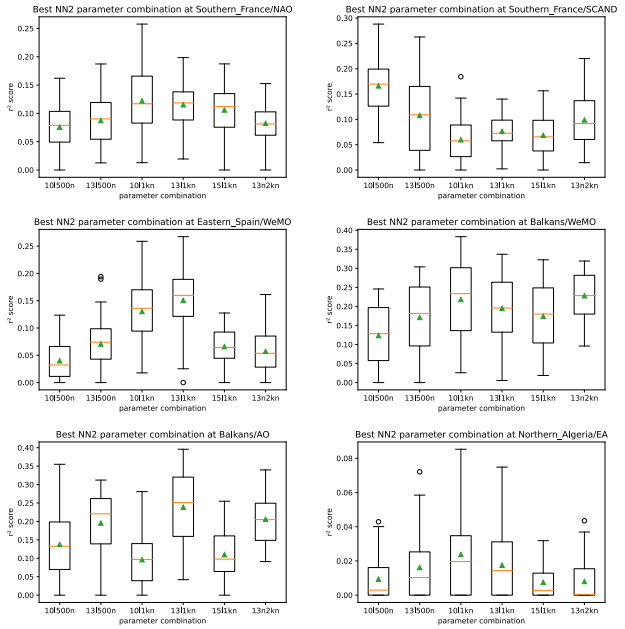

Fig. 5. NN2 parameter search for 6 parameter sets of different layer (l) and node (n) combinations. Larger MLP architectures are needed for NN2 than for NN1. The models using $10l$ and $13l$, each with $1000n$ are often best performing.

shown maps are also given in figures 7 and 8. Additionally, exact values and further measures such as fit time per time series are shown in tables 4 and 5. Since we are only interested in the fit times relative to the baseline rather than the absolute duration which depends on the system they are being run on, the duration relative to the baseline ($0.0016s$) is also provided. All the fit times have been calculated on the same system using only the Intel i7 CPU and not using a GPU for the Neural Networks, which can significantly decrease their fit times.

First, the validation set is looked at. The overall mean r^2 scores across the entire map are very low at about 0.15 for LR and M5, 0.1 for RF, 0.14 for NN1 and 0.13 for NN2. In other words, on average, the models are only able to explain about 15% of the precipitation variation using the 9 atmospheric oscillation indices as predictors. The low means of r^2 scores are due to the models commonly failing to estimate precipitation in Northern Algeria, the Southern and Eastern regions of Italy and the Balkans. When comparing this distribution with the map in fig. 12, it can be seen that these regions are mostly dominated by the EA pattern. A pattern can be observed that the regions often lie east of mountain ranges, where atmospheric oscillations controlling the westerlies (such as NAO or AO) have little influence. Predictions here could be improved by including other atmospheric oscillation indices not considered in this study. Further comparing with the maps in figures 11 and 12, it can be derived that the models work best in regions primarily influenced by the NAO and AO, independently of their precipitation amounts. For example, the models perform relatively well in regions of middle- and western Spain or the western half of the Italian Peninsula, which are primarily influenced by the NAO, while at the same time experiencing comparably low precipitation. The models also work relatively well in regions affected by the WeMO such as western France, parts of western Italy and the Balkans. Although, this does not hold up for dry regions such as eastern Spain. Comparing the models themselves on the validation sets, we see that only M5 manages to perform as well as the MLR baseline. The two models perform nearly identical in the entire region. The similarity is also confirmed by the boxplots. The similarity may be due to M5 approximating a linear regression with the multiple single linear regressions at its leaf nodes. The two models reach performances of r^2 scores above 0.24 in 25% of the regions and reach maximum scores of around 0.41. The spatial distribution of the performance of the RF models is similar to that of the latter two models. Yet, the model has an overall worse performance with a mean of only 0.10 and the model only having r^2 scores of less than 0.2 for 75% of the regions. The distribution of the scores is more contrasted, with the model reaching even lower scores in the difficult-to-predict regions, but performing very well in easier-to-predict regions and partly even reaching slightly higher scores reaching a maximum r^2 score of 0.45. Continuing with the MLP models, the performance of the tested NN1 MLP shows the noisiest results. This is due to the saving of resources in the training by skipping the cross validation, which can lead to results being shown of models that were poorly trained or that were tested on only a single randomly generated validation set with nonrepresentative values. For this model, the 75$^{\text{th}}$ percentile is located at about 0.23 which only comes close to that of LR and M5. Nevertheless, the upper 25% reach to much higher values with the 99$^{\text{th}}$ percentile being located at 0.45 and the maximum at 0.64. These values must also be taken with caution as they could just be outliers created under the circumstances of not being cross-validated as described above. NN2 also shows slightly worse performance than LR and M5, but does not fail to match the performance of NN1, only having a slightly worse mean performance of 0.13 compared to 0.14 of NN1.

When inspecting the predictions on the test set, each of the models perform overall worse than on the validation set which is to be expected (see fig. 8). The relative performances remain similar but NN1 and NN2 perform relatively worse and are only as good as RF (while they were only slightly worse than LR on the validation set). The spatial distribution of each of the models' prediction skill on the test set also differs from that on the validation set (comparing right and left column of the maps in fig. 6). For example, most of the models perform weakly on the coast of eastern Spain in the validation set but perform better in the test set. Conversely, the models perform better in northern Italy on the validation set than on the test set. This is due to the climate of the last 5 years being different from that on the 35 preceding ones, where the random samples were taken from for the validation set. While it is to be expected that the predictions themselves on the test set are inherently different, their quality should not vary as observed. These observations reveal that with the rapidly changing climate, the models are not always able to maintain their performance, making their predictions on most recent climate unreliable.

Regarding the models, M5 and LR turn out to perform nearly identical. As opposed to the validation set performances, where the RF's results had a similar spatial distribution as LR, there are some minor differences when comparing them on the test set. These differences are predominantly visible as a lack in performance of RF in locations such as North-eastern Spain or Southern France, where other models do not fail. Still, in the regions along the border between Spain and France (where the Pyrenees Mountains are) or in the Gibraltar region, RF seems to achieve better accuracy on a larger surface than the other regression models.

When comparing the fit times, LR fits a time series nearly instantly within only 0.0016 seconds. RF and M5 already need 100 and 131

Table 4. r^2 scores of each model on the validation set. (q denotes the quantiles; as seen in the boxplots).

Model	mean	q0.25	median	q0.75	q0.99	max	fit times (s)	Multiple of LR fit times
LR	0.146	0.052	0.125	0.242	0.362	0.414	0.0016	1
RF	0.104	0	0.067	0.194	0.344	0.445	0.16	100
M5	0.149	0.05	0.126	0.246	0.373	0.419	0.21	131.25
NN1	0.141	0.037	0.115	0.225	0.452	0.635	9.3	5812.5
NN2	0.132	0.035	0.101	0.222	0.365	0.438	0.248	155

Table 5. See table 4

Model	mean	q0.25	median	q0.75	q0.99	max
LR	0.103	0.033	0.096	0.161	0.305	0.414
RF	0.074	0	0.052	0.122	0.295	0.395
M5	0.105	0.036	0.096	0.162	0.311	0.413
NN1	0.074	0	0.052	0.121	0.322	0.489
NN2	0.077	0.02	0.057	0.116	0.287	0.416

times longer than LR on the same time series. The calculation time of 9.3s of the MLP chosen for NN1 takes 5812 times as long as LR. The potential solution to this could be the NN2 approach that needs fit times of only 0.25s (equivalent to 155 LR fit times) which are in the same range as RF and M5. It has to be noted for the MLP approaches, batch size has a significant impact on training duration. Batch sizes of $bs = 16$ have been chosen for the shown fit times. Choosing larger batch sizes could further decrease the fit time but could potentially affect its accuracy.
Overall, the results show that for the use case defined in section *4.1.1*, most of the models struggle to surpass the MLR baseline. Only the M5 model is on-par with MLR when it comes to applicability on the entire region. With fit times that are more than 100 times as long, it is not worth using M5 over MLR. Due to the performances being averaged on the entire study area, the differences are marginal and mean r^2 performances are poor, making comparison difficult. The fact that the performance differs spatially between the validation and test set and also between different models on the test set, further complicates drawing conclusions as to which model generally works best.

4.2 Part 2: Specific use cases

In the previous results, LR has hardly been surpassed by the other regression models despite their promising capabilities. Moreover, r^2 scores in the range of 0.2 to 0.4 are generally considered very low. The cause of this can simply be that the teleconnection relationships that are to be described are naturally not particularly strong. It further may be due to the occurrence of rain being greatly dependent on the season, while the indices oscillating independently of the season, which creates hard to predict relationships. Another factor that influences the performance can be the number of variables chosen to train a model. Some models may require less predictors to create outputs with similar accuracy or may even be negatively influenced by too many predictors.
In the following we purse the second objective of the study, namely, if more information can be extracted from the data under scenarios adapted to the aforementioned potential flaws. The results on the different scenarios may be of interest for different types of research applications, depending on their aim. These could be applications analysing teleconnections of only single climate indices, applications that aim at best predicting precipitation, independently of the input variables or applications that aim at analysing precipitation in only a single season. The following two sections are an analysis on how different models perform under different conditions in order to determine which models are best applied for different use cases.

4.2.1 Using the best correlating indices at each location. Different models perform differently depending on which and how many input variables are fed to them for training. The generic tests in the previous section fed the same set of 9 input variables to each model. This may have hampered their performance on some occasions. Additionally, given the large study area, the influences of the different climate indices vary greatly. To investigate how the models react to different inputs, tests are run again on the set of chosen regions. Every model is run on the sub-regions defined in section 2.2 multiple times, but in each iteration, another predictor is added to the input. The numbers of indices (i.e., the input vector lengths) that the models are tested on are $N_{indices} = \{1..9\}$. To choose the order in which the predictors are added, their absolute Spearman's correlation coefficient is first calculated with the precipitation time series at each grid point. Then the mean of the correlation at all grid points is taken for each region individually and the indices are sorted in descending order. The rank order of the indices is shown for each location in table 6. Iterating over $N_{indices}$, the best $n \in N_{indices}$ are added to the input vector in decreasing order. It must be noted that the correlation strength of every added predictor may decrease quicker for some regions than for others. This also means that model performance not increasing with further added indices does not necessarily imply that the model can not handle more predictors, but that the added climatic index just has a low correlation with precipitation in that region.
These tests are conducted using the parameters obtained in the parameter search section 4.1 and using 5-fold cross validation to get representative results (theoretically, the optimal parameters could change with the length of the input vector, but from thereon, the testing space grows too large). For each model, the performances using the different input vector lengths are evaluated on the validation data of all the regions. In order to visualize this, the plots in 9 show the mean r^2 score over the grid of each region with an increasing number of indices added from left to right (the value on the outer right corresponds to an input vector of length 9, like tested in the previous section). A table summarizing the best n indices for each model and location is given in table 7. The best number of indices in this table is determined for each model and location by choosing the highest mean.

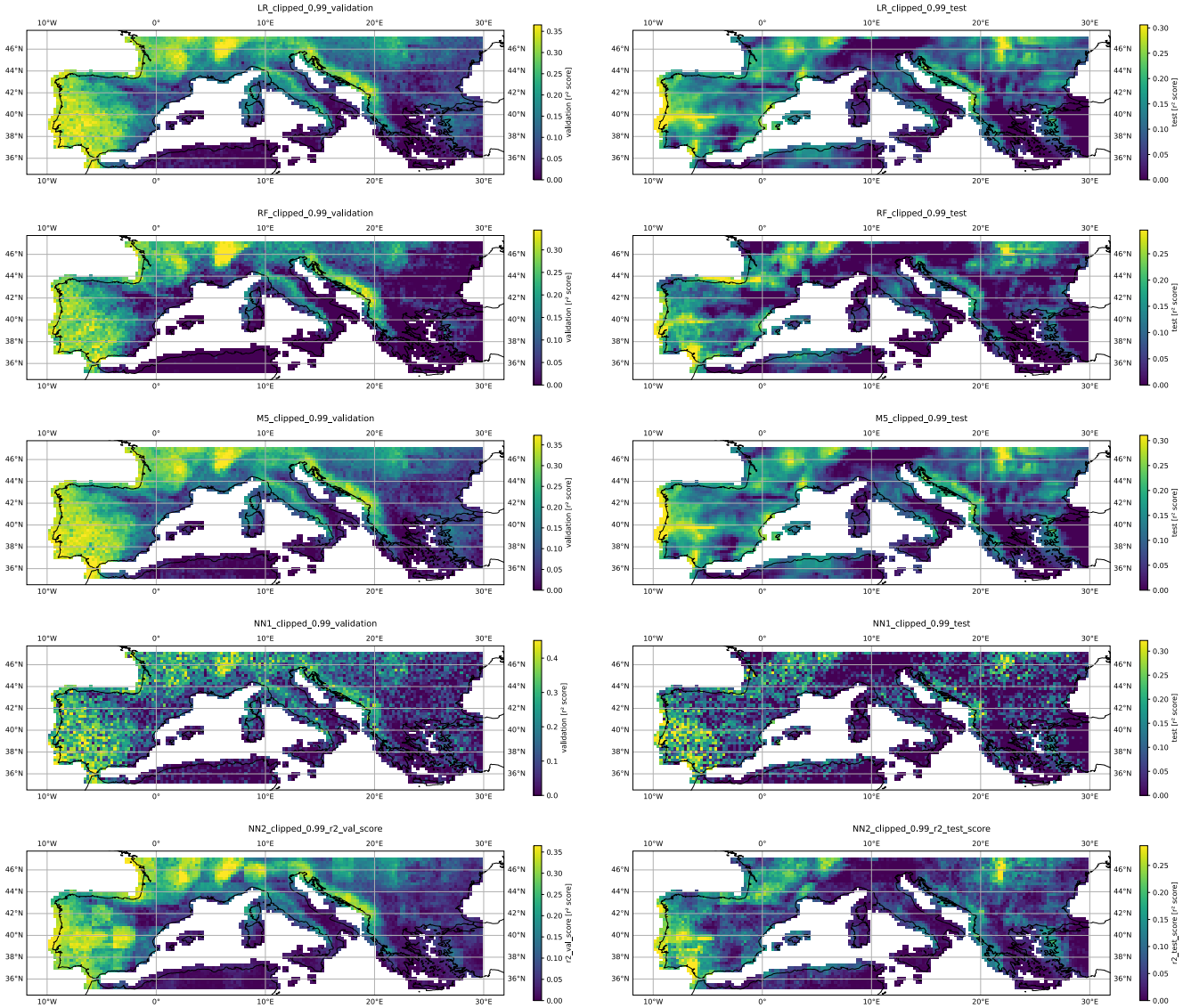

Fig. 6. Validation and test set results at each grid point scored with r^2. Only landmass is shown.

Table 6. Ranked indices and their mean absolute Spearman's ρ at each location.

Rank	Southern France/NAO		Southern France/SCAND		Eastern Spain/WeMO		Balkans/WeMO		Balkans/AO		Northern Algeria/EA	
1	scand	0.24	wemo	0.29	wemo	0.24	wemo	0.32	ao	0.21	ea	0.22
2	nao2	0.23	ao	0.24	nao2	0.24	eawr	0.24	ea	0.19	nao2	0.16
3	ao	0.22	scand	0.23	ea	0.15	ao	0.21	nao2	0.19	scand	0.11
4	wemo	0.22	nao2	0.21	scand	0.12	nao2	0.18	eawr	0.17	wemo	0.1
5	eawr	0.16	eawr	0.16	ao	0.08	scand	0.16	wemo	0.16	ao	0.1
6	ea	0.08	ea	0.07	moi1	0.06	ea	0.12	scand	0.12	eawr	0.06
7	moi1	0.04	amo_us	0.04	eawr	0.04	moi1	0.07	amo_us	0.09	moi1	0.03
8	amo_us	0.04	moi1	0.03	amo_us	0.02	amo_us	0.06	moi1	0.03	amo_us	0.03
9	soi_std	0.03	soi_std	0.03	soi_std	0.02	soi_std	0.02	soi_std	0.02	soi_std	0.02

As seen in the figures, the model performances can vary greatly depending on the number of predictors and location and there are many occasions in which the ideal number is below 9. LR and M5 behave nearly identical and generally work well with most indices added to the input. This is even more the case for RF which uses 7-9 indices for achieving best predictions. Exceptions are both South-

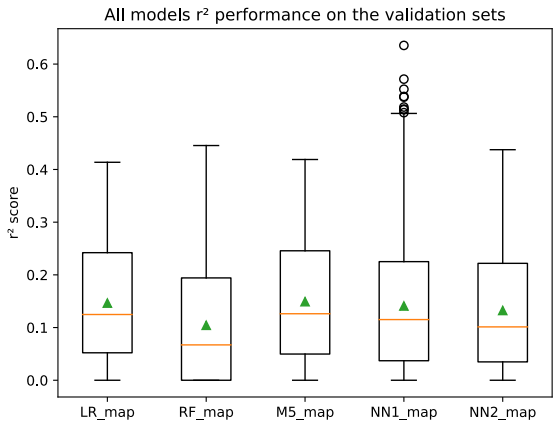

Fig. 7. Distribution of each models' r^2 scores across the grid points of the entire study area as seen in the left column of fig. 6.

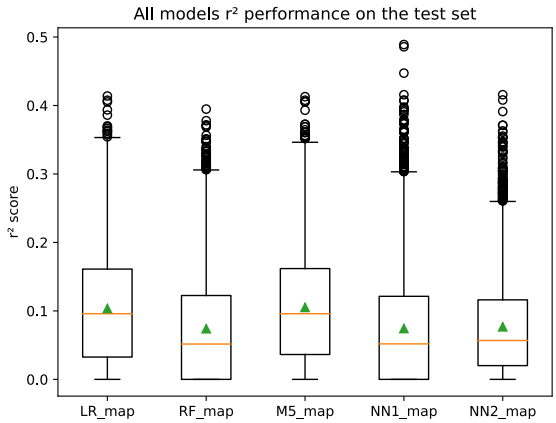

Fig. 8. See fig. 7 but on the test set.

Table 7. Best number of indices for each model and region

Region	LR	RF	M5	NN1	NN2
Southern France/NAO	7	7	7	7	8
Southern France/SCAND	8	8	8	6	7
Eastern Spain/WeMO	9	8	9	9	6
Balkans/WeMO	4	7	4	4	5
Balkans/AO	5	9	5	4	5
Northern Algeria/EA	4	7	4	6	4

ern France regions and Balkans/AO, where adding the 9th (and 8th) index results in a slight dip in the performance. In Northern Algeria/EA LR and M5 also work best with only the first four indices added to the input, although overall performance is extremely poor in this region with r^2 scores below 0.1. For the Balkan regions, where the best performance lies at only 4 or 5 indices, adding more indices barely affects the influence at all and brings only a slight decrease.

Most interestingly, there are some occasions where NN1 performs best with much less than all 9 indices. This can be best observed in both of the Balkans regions, where using more than the 4 best correlating indices (WeMO, AO, EAWR and NAO) significantly worsens the performance. In these two locations, this selection of less indices enables NN1 to reach the overall best score. The weakness of NN1 to process the longer input vectors is remarkable as ANNs are better suited to use higher dimensional input. The reason for this could lie in the lack of training samples. The examination on smaller regions gives a better insight than on the entire map (like in the previous section) and reveals that NN1 is able to take a significant lead over the other models in many regions even with all input features selected. Regions such as Balkans/AO where NN1 performs much worse with all features could be the reason of its lacking performance in the previous section.

Despite being averaged using cross-validation, the results of the NN2 MLP appear noisy and unclear. A rough upwards trend can be observed as to which number of indices is best used for this model, though using all indices never leads to the best results. It can be said that its best performance lies somewhere in the mid-ranges (using either one or all indices never leads to great results).

An interesting side note is that the multidecadal AMO adds valuable prediction skill to most models in Southern France and Eastern Spain, despite its relatively low correlation and thus being added to the input as one of the last indices.

Generally it can be said that one should be cautious when selecting input features for a model to achieve best possible predictions, especially when applying MLPs. Also, other methods than ranking by correlation are recommended for feature selection (e.g. recursive feature elimination), as even low-correlated indices (here AMO) can contribute much to the prediction skill. The optimal number of indices for each model and location are also used for the applications of the models in the next section.

4.2.2 Data split into different seasons. In the preceding sections, the models have been trained to predict precipitation for all months. As previously stated, most of the precipitation in the Mediterranean occurs during the winter months. At the same time, the climatic indices oscillate independently of the season and their time series are mostly provided already deseasoned. Furthermore, the influence of the climatic indices varies strongly across the year (e.g. NAO has the greatest influence on precipitation during winter). This means that the climatic indices could oscillate so that they appear with identical or similar values multiple times across the year, while the precipitation values (that are to be predicted) could be different for those same input values at different times of the year. Despite using deseasoned precipitation data, this occurrence may still be reflected as an impairment to the ability of the models to estimate rainfall.

To investigate whether this phenomenon has an effect on the models' performances, the input and target data is split into the four seasons, namely DJF, MAM, JJA, SON. With an overall available 480 months of data, this reduces the number of data points to just 120 months for each season. The models are then trained and tested exclusively on each of the four seasons. Like in the previous section, the parameters used are the ones determined in section 4.1 and 5-fold cross validation is applied. Additionally, the results from the previous section *4.1.8* are applied and only the best n indices are used depending on the model and location. The models are then evaluated for each region and each season to be compared next to one another in the boxplots in fig. 10.

The seasonal differences become evident in the visualizations. The winter season (DJF) consistently sees the best results across all locations and for all models. Spring (MAM) and autumn (SON) come second to that with scores in similar ranges, but different behaviour in different regions. For spring, the models perform best in Southern France/NAO, the dry regions of Eastern Spain and Northern

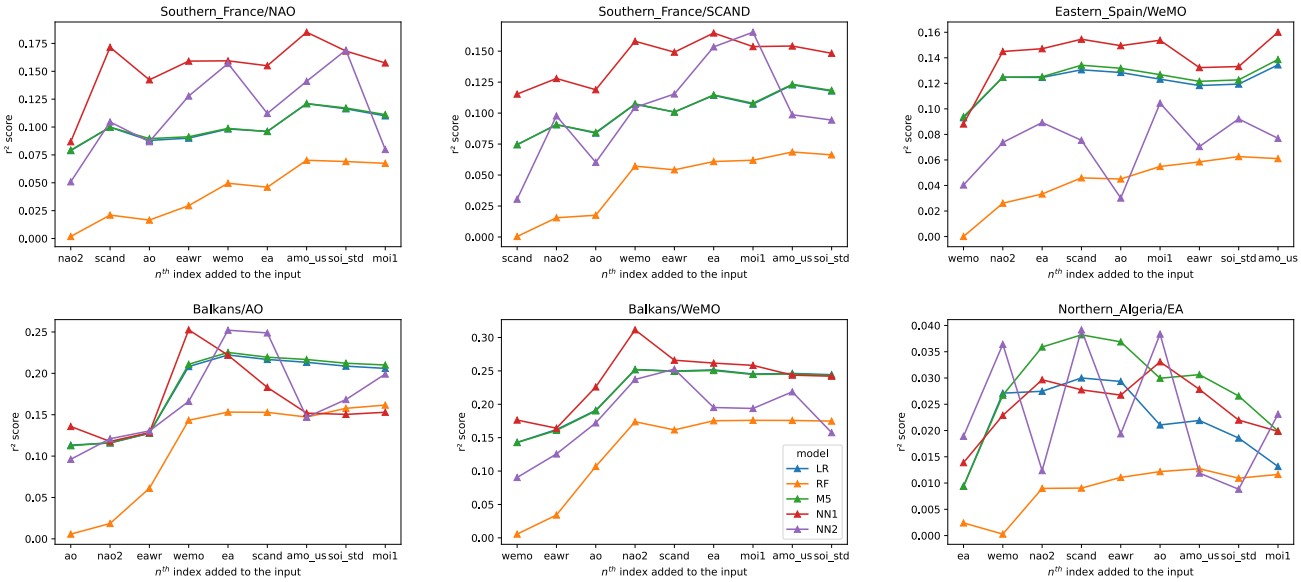

Fig. 9. Mean r^2 scores of the models' predictions on each region with increasing number of indices added to the input vector (left to right). Most notably, NN1 works best with only 4 indices as input in both Balkan regions and has much worse performance with longer input vectors. (See legend in middle bottom row plot for colors associated with the models.

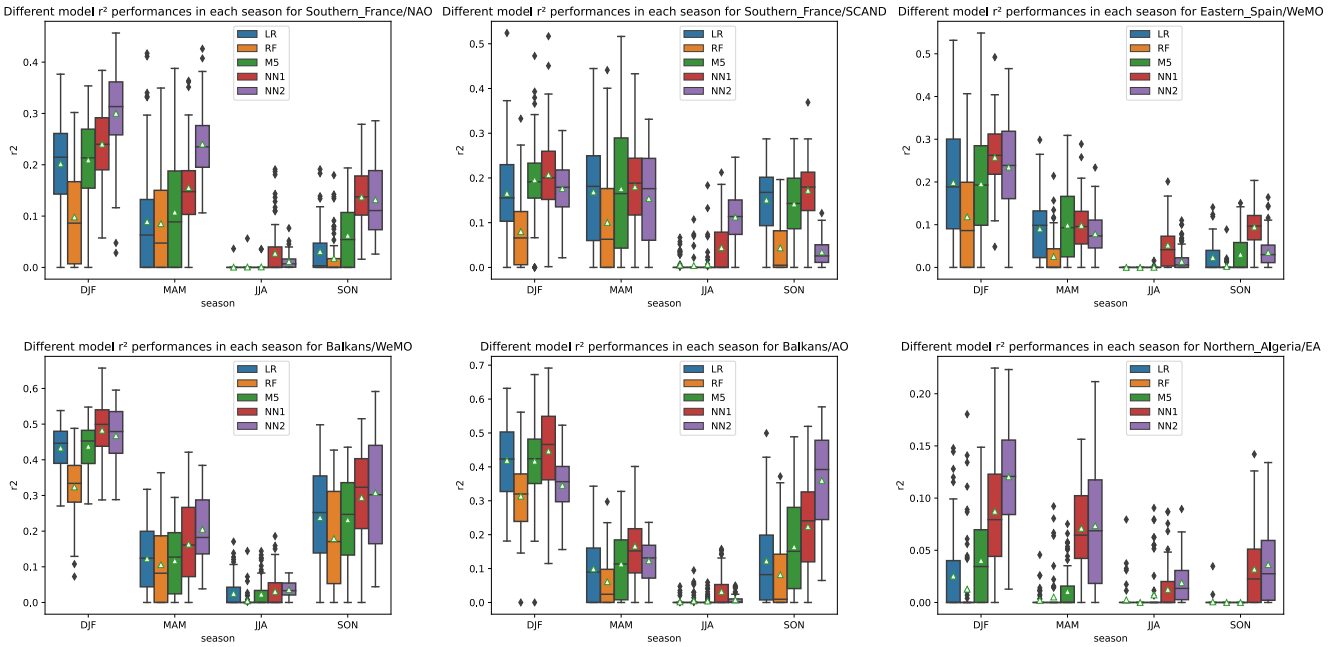

Fig. 10. Seasonal performance of each model on the validation sets of the different regions. The seasonally differing influence of the indices becomes evident with summer (JJA) being the hardest to predict and winter (DJF) being the season where the models show best results. r^2 scores of above 0.6 are reached in some regions (Balkans) which is much higher than scores reached in the previous all-yearly predictions. NN1 is shown to consistently outperform the other models apart from NN2, which also shows good performance in some locations/seasons.

Algeria (where only NN1 manages to perform at all). For Southern France/SCAND, they perform similarly during spring and autumn regarding the mean but show a much greater spread up- and downwards for spring than for autumn. For both chosen Balkan regions, all the models perform better during autumn than during spring. Lastly, almost all the models appear to be failing to pre-

dict the low precipitation during summer (JJA) in all regions. The only exception to this being the MLPs implemented in both, NN1 and NN2 which show some marginal performance for summer. For these models too, summer precipitation is the hardest to predict. The explanation for the seasonal differences lies in the correlation strengths of the predictors which vary within the years in similar patterns. To visualize this, the correlation between the most dominant atmospheric index and the precipitation of the regions has been calculated for each of the seasons separately. The correlations are shown in the appendix fig. 13. The distribution of the model accuracies across the seasons roughly resembles that of the distribution of the correlation strengths (i.e., in seasons where the correlations are the strongest (mostly DJF), the models tend to be the most accurate as well). Exceptions to this are the regions Balkans/AO and Southern France/SCAND, where the correlations during summer are not lower than correlations during other seasons, while the performance of the models during summer is still the worst. These low scores of the models on summer precipitation can further be explained with the summer mean precipitation of the non-deseasoned dataset, which is the lowest during summer for all regions (see table 8).

LR mostly ranks 3rd after M5 and NN1 for most locations and seasons and only manages to outperform RF consistently. Only in the dry region of Eastern Spain/WeMO, LR manages to perform better than M5. Furthermore, it performs very similar to M5 in the Balkans region, which also has WeMO as dominant index. RF fails to outperform any of the models in all regions and seasons with no exception. M5, applied on these conditions, manages to surpass LR in multiple occasions. Southern France/NAO is its best performing location compared to the other models with means consistently above those of LR across all seasons under the same conditions. In other locations such as Southern France/SCAND, Balkans/AO and Northern Algeria/EA it only manages to outperform LR by a small margin for some seasons. The NN1 MLP consistently beats the latter three models in nearly all locations and seasons. This holds up for all the means and most percentiles. In estimating winter precipitation in Balkans/AO, r^2 scores reach an overall maximum of 0.69. The prediction accuracies of NN2 are the least determined. In some instances, they are more accurate than all others whereas in other instances, they are only better than those of RF or even worse than that (SON in Southern France/SCAND). This is despite averaging the predictions of multiple iterations of cross-validation. NN2 outperforms all the other models by a relatively large margin for winter precipitation in Southern France/NAO and Northern Algeria/EA, spring precipitation in Southern France/NAO and Balkans/WeMO, summer precipitation in Southern France/SCAND and autumn precipitation in Balkans/W. For the other cases it performs similar to the other models, mostly still better than LR. Its worst predictions (worse than LR) are those for spring and autumn in Southern France/SCAND and Eastern Spain/WeMO and winter in Balkans/AO.

The uncertainty in the accuracies of NN2 predictions make it questionable whether this implementation of a MLP is preferable over that of NN1. Its superiority over NN1 and the other models in some seasons hints a potential that could be made more robust with a better choice of parameters, different Neural Network shapes or additional features.

Interestingly, both MLP models significantly outperform their competitors in the hardest to predict dry region of Northern Algeria, albeit only reaching scores in lower ranges (maximum reached r^2 of 0.22). The MLPs also make best predictions for summer precipitation in Eastern Spain, which is even lower than that of Algerian summer (see table 8). Furthermore, NN1 and NN2 are the only models that can estimate some of the variance of dry summer precipitation, though only reaching r^2 scores of 0.25 at best in Southern France/SCAND with NN2. The only exception to this appears in Balkans/WeMO, where LR and M5 show a marginal prediction skill. This region is also the one that receives the most precipitation during summer compared to the other regions (see table 8).

Overall, estimating precipitation trends based on climate indices as pursued in this study works best with focus set on only discrete seasons. With this approach, the best models manage to reach r^2 scores up to twice as high as those reached on estimations over the entire year. For instance, in the Balkans/WeMO region, most of the models achieved mean r^2 scores of around 0.23 (as derived from the figures of the parameter searches of the individual models). When applied seasonally, the same models reach mean scores around 0.45 for winter while maintaining good performance for autumn.

5. CONCLUSION

The present study evaluated the relative performance of various linear and non-linear regression models to model Mediterranean precipitation based on 9 climate indices. The precipitation data is deseasoned and given in monthly time scale (monthly averaged daily precipitation).

In part 1 (4.1), the models were tested on a large scenario, considering all predictors chosen for this study, the entire geographic study area and all seasons. The results show that Linear Regression is hardly surpassed in this universally set use case. M5 is the only model on-par with LR. RF is shown to be unsuited for this type of application. The MLPs of NN1 and NN2 are also struggling to make accurate predictions under the given conditions. The noisiness of NN1 results, that are possibly due to the little available training data, suggest that it is important to make ensemble predictions when using similar implementations of MLPs on monthly predictions. The biggest drawback of NN1 is the very high computation time compared to other models. It takes multiple thousand times longer computation times than LR, while RF, M5 and NN2 are still relatively quick and need about 100 times longer than LR. The significantly shorter fit times of NN2 as compared to NN1 suggest that when using MLPs on gridded data, it is recommended to exploit their capabilities to make predictions on larger targets at once instead of training one model on each time series.

The results of part 1 also highlighted some of the circumstances under which the regression models are most effective. It was found that in the Mediterranean, the models make best estimates in western parts of landmasses, where precipitation is mostly controlled by the westerlies. These are regions where NAO, AO and WeMO have the strongest correlations with precipitation. The models struggle to make rainfall estimations in regions where the aforementioned indices have less influence and where EA and SCAND are most prominent. The insights highlight the importance of a good choice of predictors depending on the observed region. Moreover, the struggle of the regression models to exceed LR despite their abilities to learn nonlinear relationships suggests that rainfall estimations are generally difficult to make in use cases that are designed too ambitious in terms of the time periods to be predicted and number of predictors. Conclusions were drawn that the models should be evaluated under different conditions that address the difficulties of the first approach to get a better picture of their behaviour and capabilities.

In part 2, models were first tested using less input features and then tested on seasonally separated data. These tests were conducted on 6 smaller regions with differing climatic conditions within the

Mediterranean.
In the first test, the models were run multiple times, iteratively using more predictors out of the selection of 9 indices. The predictors are added in order of strongest to least correlating in the respective region. The results revealed that using more predictors doesn't necessarily improve a models' performance. More specifically, RF generally works best with most predictors added to the input vector, LR and the marginally better performing M5 have strong performance independently of the number of predictors upwards of 4, NN1 sometimes works best with only few (4) selected predictors and the best choices for NN2 are not very clear. Furthermore, NN1 was found to be able to outperform all other models by a significant margin with the right choice of input.
To further address the weaknesses of the models on the universal approach of part 1, they were additionally tested on the 4 seasons exclusively. The highly varying prediction skills throughout the four seasons demonstrate the differing influences of the climatic indices on deseasoned precipitation in each season. It becomes evident that observing the teleconnections seasonally should generally be preferred. The seasonal predictions can be much more accurate as compared to those on the entire year and the seasonally differing influences of the predictors become much more visible. This holds up especially for winter season which is of particular interest in Mediterranean climate research, given that this is when the most of the annual precipitation occurs. In this seasonal approach, RF remained the worst regression model. M5, and NN2 on the other hand were able to outperform LR in many occasions and NN1 took the lead in nearly all occasions. Both MLP implementations turned out to be the best performing models for the hardest to predict dry regions such as Eastern Spain and Northern Algeria. They are also the only models to show some correct predictions for the dry Mediterranean summer seasons, though still performing very poorly. This is despite the training data being cut to just 1/4th of the original data resulting in very little data points to train on.
Overall, the results show that there is no such thing as "the" best model, but that the choice of model depends on the type of application. We recall that the scope of this study is narrowed down to estimation of gridded deseasoned precipitation data using one or more large-scale atmospheric oscillation indices as predictors with no lead times and on a monthly timescale. With reference to our findings in part 1 of the study, we suggest the use of MLR for applications disregarding the seasons/seasonality of the teleconnections and set on larger grids, where short computation times are desired. Its advantages over the other models are robust predictions and shortest fit times.
For applications that consider the seasons separately (which we also recommend) and where computation time is less important, we recommend an approach using an MLP implementation similar to that of NN1, preferably in an ensemble to make more robust predictions. Its advantages over the other models are consistently better predictions on seasonal precipitation, including predictions in dry regions, where other models fail. This could be of specific interest for the Mediterranean region, where dry conditions are expected to increase, consequently posing an increasing threat. Furthermore, Neural Networks have the largest potential as they have the largest parameter space that can be tuned.
Lastly, we want to mention the MLP implementation of NN2 which uses the same input to predict an entire extract of the gridded precipitation dataset. Its promising results in a few regions of the seasonal approach indicate its potential to beat other models in regard to fit time and accuracy. However, since the shown results are at times not very robust, it needs further improvement, for example through choices of different parameters, Neural Network shapes or perhaps by also feeding them the month of the data point as an input feature.

6. REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] P Alpert, T Ben-Gai, A Baharad, Y Benjamini, D Yekutieli, M Colacino, L Diodato, C Ramis, V Homar, R Romero, et al. The paradoxical increase of mediterranean extreme daily rainfall in spite of decrease in total values. *Geophysical research letters*, 29(11):31–1, 2002.

[3] PA Arias. Technical summary. in: climate change 2021: the physical science basis, 2021.

[4] Anthony G Barnston and Robert E Livezey. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly weather review*, 115(6):1083–1126, 1987.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] B Choubin, A Malekian, S Samadi, S Khalighi-Sigaroodi, and F Sajedi-Hosseini. An ensemble forecast of semi-arid rainfall using large-scale climate predictors. *Meteorological Applications*, 24(3):376–386, 2017.

[7] Bahram Choubin, Shahram Khalighi-Sigaroodi, Arash Malekian, and Özgür Kişi. Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. *Hydrological sciences journal*, 61(6):1001–1009, 2016.

[8] M Conte, A Giuffrida, and S Tedesco. Mediterranean oscillation: Impact on precipitation and hydrology in italy. In *Conference on Climate and Water.*, volume 1, 1989.

[9] Ravinesh C Deo, Ozgur Kisi, and Vijay P Singh. Drought forecasting in eastern australia using multivariate adaptive regression spline, least square support vector machine and m5tree model. *Atmospheric Research*, 184:149–175, 2017.

[10] Noah S Diffenbaugh and Filippo Giorgi. Climate change hotspots in the cmip5 global climate model ensemble. *Climatic change*, 114(3):813–822, 2012.

[11] David B Enfield, Alberto M Mestas-Nuñez, and Paul J Trimble. The atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental us. *Geophysical Research Letters*, 28(10):2077–2080, 2001.

[12] H Hersbach, B Bell, P Berrisford, G Biavati, A Horányi, J Muñoz Sabater, J Nicolas, C Peubey, R Radu, and D. Simmons A. Soci C. Dee D. Thépaut J-N. Rozum, Schepers. Era5 monthly averaged data on single levels from 1979 to present, 2019. (Accessed on 08-07-2021), 10.24381/cds.f17050d7.

[13] Jiyeong Hong, Seoro Lee, Joo Hyun Bae, Jimin Lee, Woon Ji Park, Dongjun Lee, Jonggun Kim, and Kyoung Jae Lim. Development and evaluation of the combined machine learning models for the prediction of dam inflow. *Water*, 12(10):2927, 2020.

[14] Phil D Jones, Trausti Jónsson, and Dennis Wheeler. Extension to the north atlantic oscillation using early instrumental pressure observations from gibraltar and south-west iceland. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 17(13):1433–1450, 1997.

[15] Simon O Krichak and Pinhas Alpert. Decadal trends in the east atlantic–west russia pattern and mediterranean precipitation. *International journal of climatology: a journal of the Royal Meteorological Society*, 25(2):183–192, 2005.

[16] Simon O Krichak, Joseph S Breitgand, Silvio Gualdi, and Steven B Feldstein. Teleconnection–extreme precipitation relationships over the mediterranean region. *Theoretical and applied climatology*, 117(3):679–692, 2014.

[17] Javier Martin-Vide and Joan-Albert Lopez-Bustins. The western mediterranean oscillation and rainfall in the iberian peninsula. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 26(11):1455–1475, 2006.

[18] Shifa Mathbout, Joan Albert Lopez-Bustins, Dominic Royé, Javier Martin-Vide, and Aziz Benhamrouche. Spatiotemporal variability of daily precipitation concentration and its relationship to teleconnection patterns over the mediterranean during 1975–2015. *International Journal of Climatology*, 40(3):1435–1455, 2020.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] John R Quinlan et al. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific, 1992.

[21] Mehdi Rezaeian-Zadeh, Hossein Tabari, and Hirad Abghari. Prediction of monthly discharge volume by different artificial neural network algorithms in semi-arid regions. *Arabian Journal of Geosciences*, 6(7):2529–2537, 2013.

[22] Concepción Rodriguez-Puebla, AH Encinas, S Nieto, and J Garmendia. Spatial and temporal patterns of annual precipitation variability over the iberian peninsula. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 18(3):299–316, 1998.

[23] Chester F Ropelewski and Phil D Jones. An extension of the tahiti-darwin southern oscillation index. *Monthly weather review*, 115(9):2161–2165, 1987.

[24] Mohammad Taghi Sattari, Fatemeh Shaker Sureh, and Ercan Kahya. Monthly precipitation assessments in association with atmospheric circulation indices by using tree-based models. *Natural Hazards*, 102(3):1077–1094, 2020.

[25] Richard Seager, Haibo Liu, Naomi Henderson, Isla Simpson, Colin Kelley, Tiffany Shaw, Yochanan Kushnir, and Mingfang Ting. Causes of increasing aridification of the mediterranean region in response to rising greenhouse gases. *Journal of Climate*, 27(12):4655–4676, 2014.

[26] smarie. m5py, 2022. Software available from https://smarie.github.io/python-m5p/.

[27] David WJ Thompson and John M Wallace. The arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical research letters*, 25(9):1297–1300, 1998.

7. SUPPLEMENTARY MATERIAL

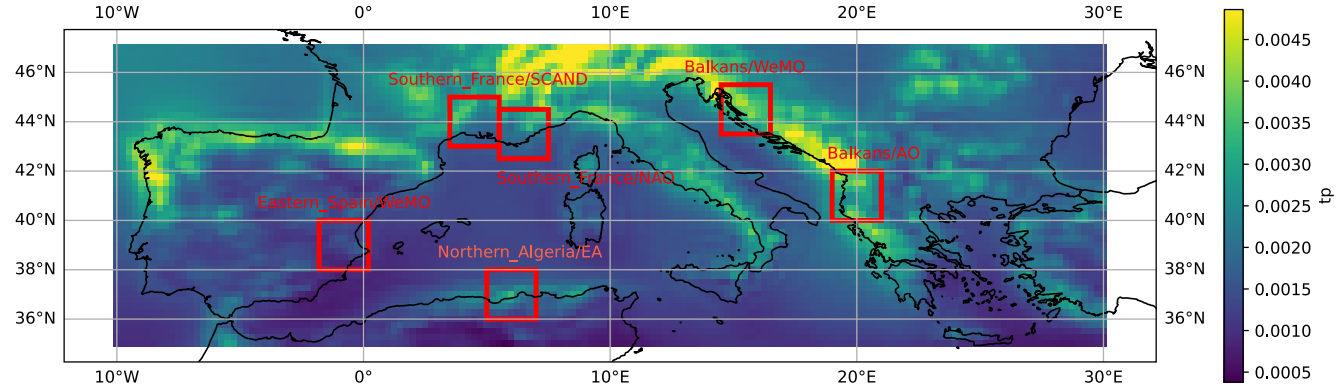

Fig. 11. All time mean monthly deseasoned total precipitation (tp, unit in metres). The values are capped at the 99th quantile to remove very high precipitation for visualization purposes.

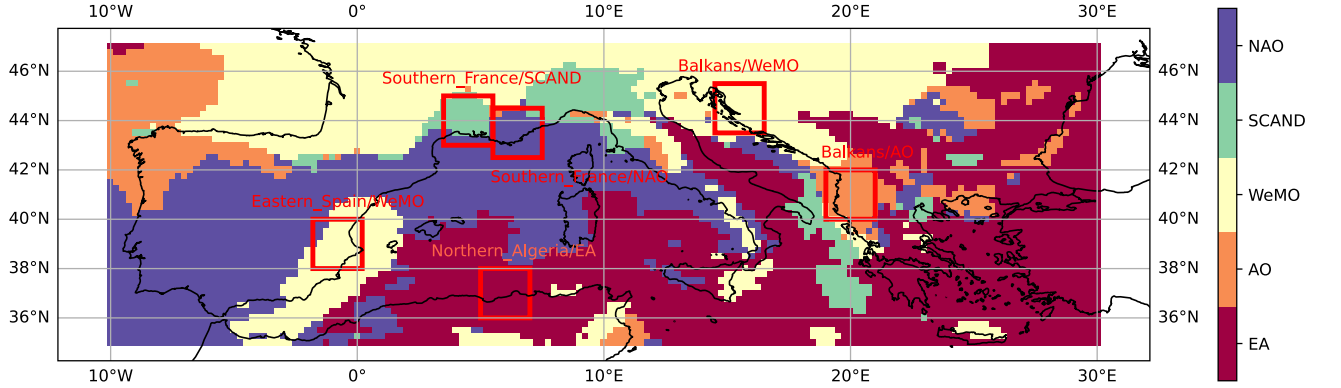

Fig. 12. Indices with highest correlation with deseasoned precipitation time series at each grid point.

Table 8. Precipitation features of the 6 regions

Region	Dominant index	all-time mean	DJF	MAM	JJA	SON
Southern_France	NAO	2.24	2.26	2.36	1.37	3.02
Southern_France	SCAND	2.43	2.31	2.42	1.47	3.51
Eastern_Spain	WeMO	1.17	1.25	1.14	0.5	1.84
Balkans	WeMO	3.19	3.42	3.14	2.28	4.02
Balkans	AO	3.13	4.19	3.18	1.37	4.03
Northern_Algeria	EA	1.79	2.79	1.85	0.37	2.14

Note that the seasonal means are calculated on the non-deseasoned dataset. All units are monthly averaged daily precipitation given in millimetres (mm).

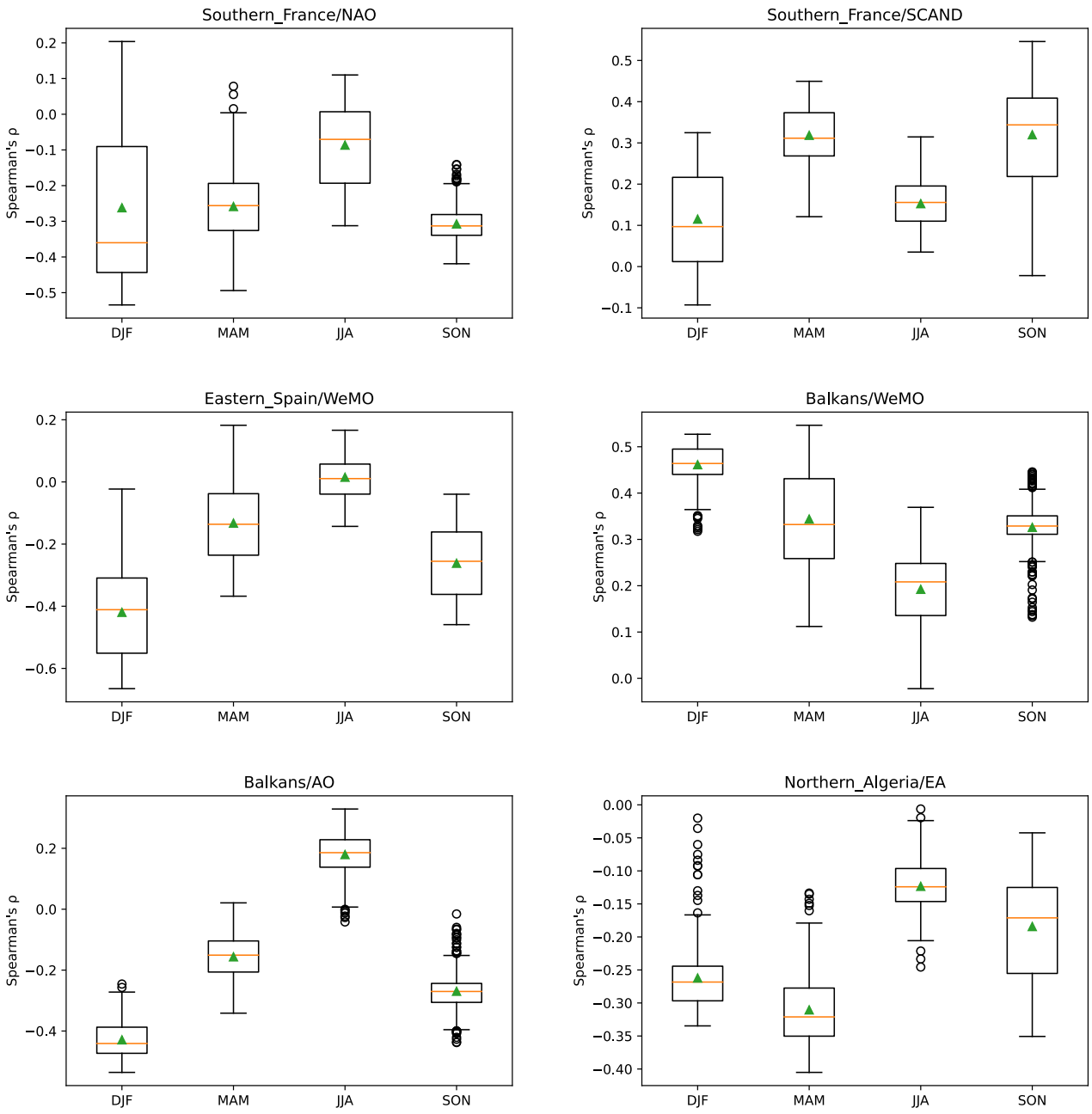

Fig. 13. Seasonal correlation for each region with its dominant climatic index.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum Unterschrift