Supplement of Clim. Past, 20, 573–595, 2024
https://doi.org/10.5194/cp-20-573-2024-supplement

Climate of the Past
Open Access
EGU

Supplement of

A past and present perspective on the European summer vapor pressure deficit

Viorica Nagavciuc et al.

Correspondence to: Viorica Nagavciuc (viorica.nagavciuc@awi.de)

1. Random Forest Reconstruction

We use a Random Forest approach (a simplified illustration is shown in Figure S1) to reconstruct VPD at each grid point. RF is a machine learning method, which was introduced by (Breiman, 2001). Since RF encompass a large variety of regression and classification methods, we use the random forest regression known as random-input random forests (Breiman, 2001). This approach is used among others within the publication of (Michel et al., 2020) and is performed using a model evaluation and optimization framework by adding hold-out sampling and cross validation techniques to the initial RF algorithm. In this section, we first detail the general workflow with which a reconstructed timeseries is evaluated within the reconstruction framework. In a second step, we give details on how a RF model is built and optimized for a given pair of VPD and proxy data.

1.1 General workflow for the reconstruction and evaluation

We here consider as predictands the observed fields of the climate quantity which we want to reconstruct. In our case, we independently reconstruct each of the observation time series from the VPD fields over the period 1900-1994 as predictand Y (applied to each grid point of the initial VPD data). The predictors which are needed to reconstruct Y are the $\delta^{18}O$ time series from our $\delta^{18}O$ network (denoted X1,...,X26). No proxy selection is made prior to the RF models fittings, since we use a relatively modest number of predictors (26) well geographically constrained to the area of interest (Europe). Nevertheless, it is worth noting that for the RF-based reconstruction of a given grid point, the RF method naturally attribute largest weights to $\delta^{18}O$ time series with the closest variations with the target VPD time series.

To evaluate the reconstruction of each grid point, the common period of observational data and proxy data is randomly divided (i.e., a bootstrap sample) R=30 times into two distinct groups (i.e., hold-out sampling (Michel et al., 2020). RF models will then be set up using the first groups of each pair of bootstrap splits (called training or calibration periods), each covering 75% of the data. The remaining 25% of the data for each bootstrap split (called the validation periods) will allow evaluating the quality of the RF models. The ratio between the validation and the training window is a compromise to have enough years to train the model as well as to calculate meaningful validation statistics. A too small time window for the validation period would make robust statements impossible, as a too short training period may result in the model not being built well due to missing or overlooked data. In other studies, the size of the training period varies between two-thirds and three-quarters of the total test period.

To quantify the RF model efficiency, the actual values of VPD are compared with the reconstructed VPD over the testing periods, that are given by RF models trained over the corresponding training periods from the same bootstrap split. In our study, we use the Nash–Sutcliffe model efficiency coefficient (S_{CE}; (Nash and Sutcliffe, 1970) which is primarily used in hydrologic modelling and is an adapted version of the common R^2. The CE is defined as:

$$CE = 1 - \frac{\sum_{t=1}^{T}(Q_m^t - Q_0^t)^2}{\sum_{t=1}^{T}(Q_0^t - \overline{Q_0})^2}$$

Where $\overline{Q_0}$ is the RF modelled VPD and Q_0^t is RF modelled VPD and Q_0^t the observed VPD at time t.

It represents the following cases which are presented in the publication (Nash and Sutcliffe, 1970). If CE=1, the RF model fits perfectly which is characterized by an error variance equal to zero; if CE=0, it indicates that the RF model has the same predictive skill as the mean of the time-series in terms of the sum of the squared error; if CE<0, it indicates that the observed mean is a better predictor than the model. Therefore, a CE=1 suggest a model with more predictive skill it is useful and meaningful to use. In our study, the hypothesis, if CE is significantly greater than 0 at the 95% confidence level, is tested with a one sided t-test.

The final gridded reconstruction is obtained by reapplying the RF model over the whole period 1900-1994 for a given window (nest), and its final score is obtained as the averaged CE scores over the 30 training/testing splits. In our study, a given grid point will be considered as robustly reconstructed if its CE scores over the 30 training/testing splits are significantly greater than 0 at the 95% confidence, based on a one sided Student t-test.

1.2 Regression tree and random forest

For all RF models either based on training samples or complete time series over the historical period, we here denote Y the VPD time series for a given grid point, $X = (X_j)_{1\leq j\leq p}$ are the *p proxy records* over the same period than Y (either a training period or the full 1900-1994 period), and $X' = (X_j^i)_{1\leq j\leq p}$ are the same *p proxy records*, but for the period over which Y is reconstructed (either a testing period or a period prior to 1900).

The RF is an aggregate of the outputs given by different regression trees (Figure S1), which are based on randomly drawn sub- sample of the initial p predictors. To build a single regression tree the X and Y data are gathered at the root of the tree as an initial node. This node is then cut into two child nodes. Any cut can be written as:

$$\{X^j \leq d\} \cup \{X^j \geq d\}$$

where j = {1, ..., p} is the proxy (variable) index and d ∈ R is the value of the cut. As a consequence, all observations with a jth variable greater than d are placed in one child node and all observations with a jth variable lesser than d are placed in another child node. The optimal (j,d) pair is determined by minimizing the sum of the intra-nodes variance of Y within the two the child nodes (Michel et al., 2020). Therefore, (j,d) is found by minimizing the following convex problem:

$$(j,d) = \arg\min_{\substack{d\in\mathbb{R} \\ 1\leq j\leq p}} \sum_{i:X_j^i<d} (Y_i - \frac{\sum_{i:X_j^i<d} Y_i}{\#(\{i: X_j^i < d\})})^2 + \sum_{i:X_j^i>d} (Y_i - \frac{\sum_{i:X_j^i>d} Y_i}{\#(\{i: X_j^i > d\})})^2$$

where p is the number of proxy records used which varies with the inferior boundary of the nested time frame (section 2.1), X_j^i denotes the ith value of the jth proxy, and # is the cardinal

operator. We then recursively apply the same procedure to the next child nodes with the same variables until one node contains less than s=5 observations. Based on this method, a decision or regression tree is created. Hence, the RF model is obtained by building a number of N such trees, but each are based on m randomly picked variables from the p initial ones, with m<p. The proxy records values over the reconstruction period of interest (i.e., X') are then used to browse each of the RF trees according to the cuts previously made during the model fitting.

The RF reconstruction is then given as the average of the outputs given by the N trees.

It is worth noting that there are inherent parameters to the RF method (i.e., control parameters). These parameters are the number of randomly used proxies in each tree (m), the number of trees (N) and the amount of data required in a node to stop the construction of regression trees (s). The optimization of three parameters is very consuming in terms of time and energy, but a very large advantage of the RF method is that it only needs one to be effectively optimized (Breiman, 2001; Oshiro et al., 2012). Indeed, for a large suite of simulated datasets, (Oshiro et al., 2012) have shown that using more than N=128 trees have a very low likelihood to gain in robustness, which they then recommended to use whatever the source of modelled data. In the same way, (Breiman, 2001) has shown that s=5 always give very satisfying results, while the gain of robustness by its optimisation does not worth the time it takes to. Therefore, we here further optimise m, for each RF model built in this study. To estimate the optimal m for each of the RF models, we use the K-fold cross-validation approach (Geisser, 1975; Stone, 1974) with K=5 similarly to (Michel et al., 2020). It consists in building RF models for every possible value of m ($0<m<p$, $m \in \mathbb{N}$) over 5 further training/testing samples determined with each of the K folds being used as the testing sample (Michel et al., 2020) for more details on the cross validation procedure). The resulting optimised RF model is then applied to the entire data of X and Y.

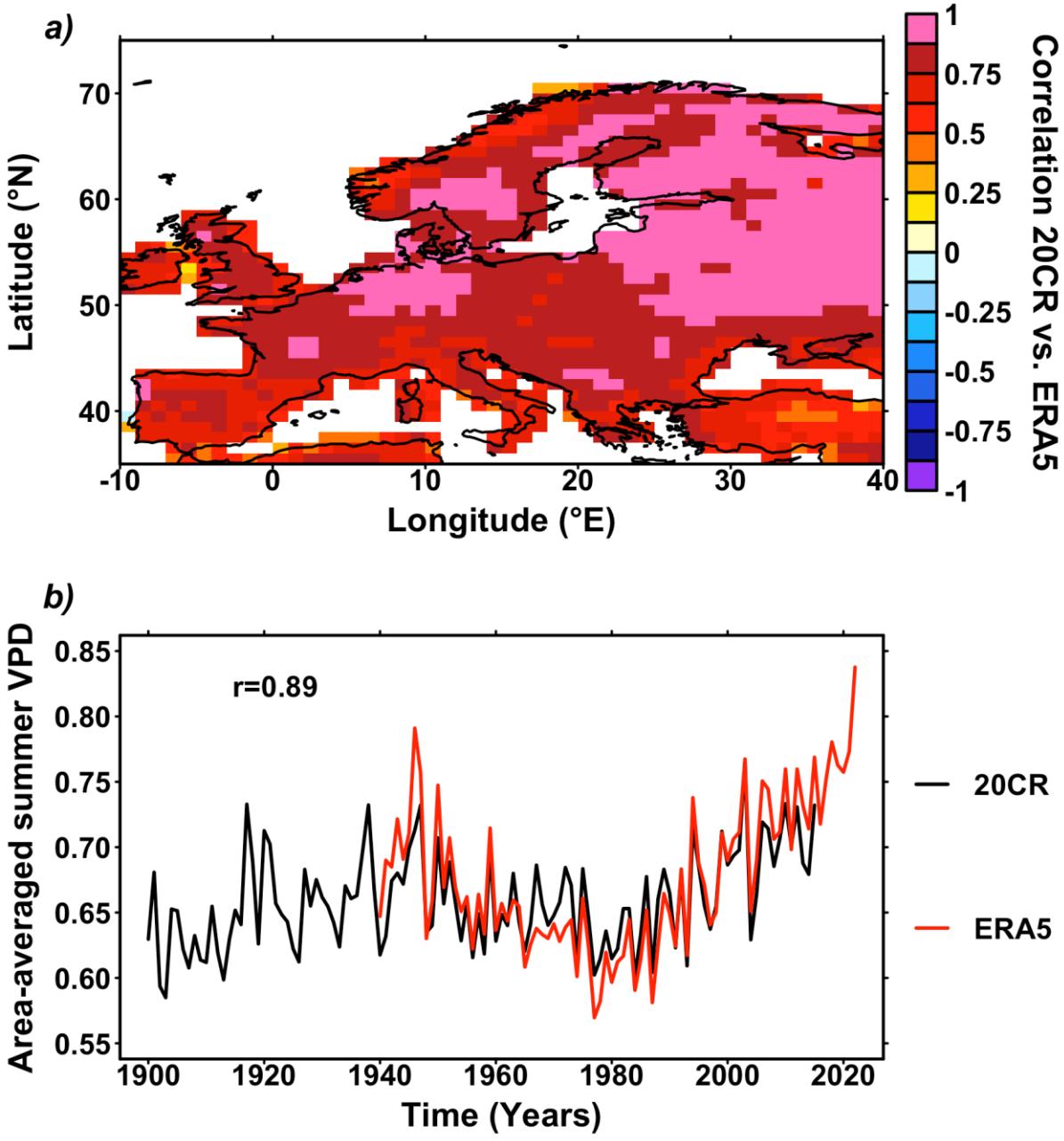

Figure S1. a) Spatial distribution of local correlations between summer VPD from ERA5 and 20CRv3 over the period 1940-2015. b) Averaged summer VPD from ERA5 (1940-2022, red line) and 20CRv3 (1900-2015, black line) over the area from a).

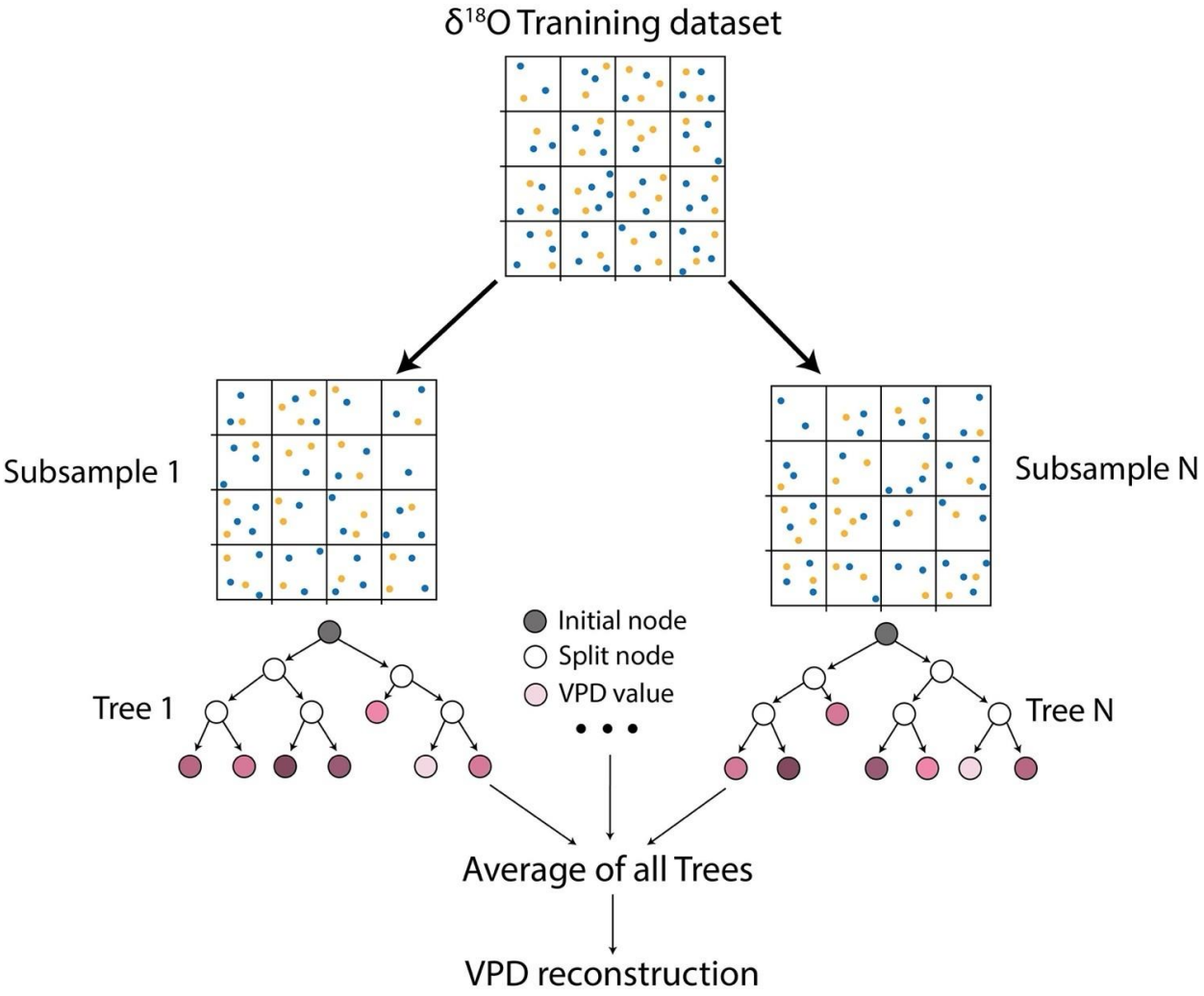

Figure S2: Simplified illustration of the RF approach for our VPD reconstruction

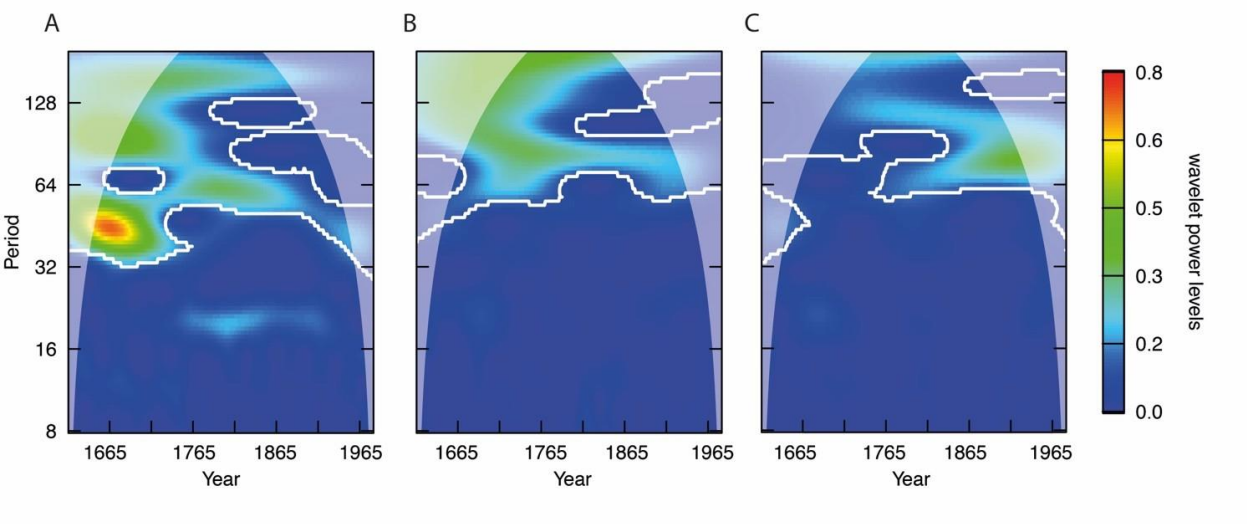

Figure S3: Investigation of the different spectra of the rolling mean time series of the three reconstructed areas. A, North Europe, **B**, Central Europe and **C**, Mediterranean region. The white contour lines indicate significant frequencies ($p < 0.01$). All the plots are done with the *WaveletComp* package in R (Rosch and Schmidbauer, 2018).

References

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, doi:10.1109/ICCECE51280.2021.9342376, 2001.

Geisser, S.: The predictive sample reuse method with applications, J. Am. Stat. Assoc., 70(350), doi:10.1080/01621459.1975.10479865, 1975.

Michel, S., Swingedouw, D., Chavent, M., Ortega, P., Mignot, J. and Khodri, M.: Reconstructing climatic modes of variability from proxy records using ClimIndRec version 1.0, Geosci. Model Dev., 13(2), doi:10.5194/gmd-13-841-2020, 2020.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, J. Hydrol., 10(3), 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Oshiro, T. M., Perez, P. S. and Baranauskas, J. A.: How many trees in a random forest?, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7376 LNAI., 2012.

Stone, M.: Cross validation and multinomial prediction, Biometrika, 61(3), doi:10.2307/2334733, 1974.