Supplementary Information: Regression methods

In the general case, we denote $Y \in \mathbb{R}^{n}$ the predictand, which corresponds to the AMV index, and $x \in \mathbb{R}^{n\times p}$, which corresponds to the matrix of observations from p proxy records associated to the observations of Y. We then denote $x' \in \mathbb{R}^{m\times p}$, other observations for which we search to reconstruct Y. Each regression method needs to optimise one or more statistical parameters. For example, for PCR, the reconstruction is highly sensitive to the number of Principal Components used for regressing the climate index. Hence, each of these statistical parameters are optimized using K-Fold cross validation metric (KFCF, with K=10 in this study) (*cf*. Methods, section "K-Fold cross-validation (KFCV)")[49,97,98].

- Principal Components Regression (PCR):

The PCR[49] consists in finding the best linear combination to regress Y using the principal components of x. We denote $S = x^T x \in \mathbb{R}^{p\times p}$, the variance-covariance matrix of x, where x^T is the transposed vector of x. The eigenvectors of S, or Empirical Orthogonal Functions (EOFs), denoted $V = (v^j)_{1\le j\le p}$, are obtained by diagonalising S, which is equivalent to calculate the EOFs such that the variance of the projection of x on themselves is maximized. The Principal Components (PCs), denoted $U = (u^j)_{1\le j\le p}$, are calculated by projecting x on V: $U = xV$. Using KFCV (*cf*. Methods, section "K-Fold cross-validation (KFCV)"), we determine $q \le p$ PCs kept for the regression. The PCR model is constructed by estimating the best linear regression between $(U^1, \dots, U^q)$ and Y. The linear regression model is:

$Y = \beta_0 + \beta_1 U^1 + \dots + \beta_q U^q + \varepsilon$, where ε is a gaussian white noise.

$\beta = (\beta_k)_{0\le k\le q}$ is estimated by $\hat{\beta} = (\hat{\beta}_k)_{0\le k\le q}$, defined as the Ordinary Least Squares estimator. The extended AMV index is obtained by applying the estimated regression coefficients to the projected new observations on the EOFs:

$$\hat{Y}_q = \hat{\beta}_0 + \hat{\beta}_1 U^1 + \dots + \hat{\beta}_q U^q$$

- Partial Least Squares (PLS):

PLS regression[50] is an alternative to the PCR, where the EOFs, are calculated such that they are orthogonal and that the covariance between Y and the projection of x on the EOFs is maximized. To do so, we need to resolve the following dependent equations:

$$
\begin{gathered}
v^1 = arg \max_{\substack{v \in \mathbb{R} \\ |v|=1}} Cov(Y, Xv) \\
v^2 = arg \max_{\substack{v \in \mathbb{R} \\ |v|=1 \\ v^T v^1 = 0}} Cov(Y, Xv) \\
\ldots \\
v^p = arg \max_{\substack{v \in \mathbb{R} \\ |v|=1 \\ v^T v^1 = 0 \\ \ldots \\ v^T v^{p-1} = 0}} Cov(Y, Xv)
\end{gathered}
$$

Analogously to the PCR, the latent variables (LVs; PCs analog in PLS) are calculated by projecting X on the matrix $V = (v^j)_{1 \le j \le p}$: $U = xV$. Using KFCV (*cf*. Methods, section "K-Fold Cross-Validation (KFCV)"), we determine $l \le p$ LVs kept for the regression. We then construct the regression model by estimating the best linear regression between $(U^1, \ldots, U^l)$and Y:

$Y = \beta_0 + \beta_1 U^1 + \ldots + \beta_q U^q + \varepsilon$ where ε is a gaussian white noise.

$\beta = (\beta_k)_{0 \le k \le q}$is estimated by $\hat{\beta} = (\hat{\beta}_k)_{0 \le k \le q}$, the Ordinary Least Squares estimator. The extended AMV index is then obtained by applying the estimated regression coefficients to the projected new observations on the EOFs:

$$\hat{Y}_q = \hat{\beta}_0 + \hat{\beta}_1 U^1 + \ldots + \hat{\beta}_q U^q$$

- Elastic Net regression (Enet):

In the multiple regression case, $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$is estimated by the Ordinary Least Squares estimator. This usual regression is known to often result in a poor reconstruction accuracy due to several assumptions made on the original data, such as homoscedasticity. Previous studies developed regularized regression approaches to overcome the OLS defaults. The Elastic Net regression[51] is a combination of the Ridge regression and the Lasso regression.

The Ridge regression shrinks towards zero the estimated coefficients associated to predictors unlinked to the predictand. By contrast, Lasso also reduces the variability of the estimates, but in this case by shrinking to zero the estimated coefficients associated to unreliable variables. Hence, a selection is made by rejecting variables associated to coefficients shrunk to zero.

A regularized regression adds a threshold constraint using the $l_k$norm of $\beta: \|\beta\|_k^k = \sum_{j=1}^{k} |\beta_j|^k$. With k=1 for Lasso and k=2 for Ridge. The loss function of Elastic Net is given by:

$$L^{enet}(\beta) = \left\| Y - \sum_{j=1}^{p} \beta_j X^j \right\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

Where, λ_1 and λ_2 are penalty factors.

Let $w = (w_j)_{1 \le j \le p} = (sgn(\beta_j))_{1 \le j \le p}$, where sgn is the sign function. The loss function can then be denoted as:

$$L^{enet} = \|\mathrm{Y} - \mathrm{X}\beta\|_2^2 + \lambda_1 w^T \beta + \lambda_2 \beta^T \beta$$

The estimated coefficients by minimizing the loss functions are:

$$\hat{\beta}^{enet} = (X^T X + (1-\alpha)\lambda I)^{-1}(X^T Y - \frac{\alpha\lambda}{2} w)$$

Where $\alpha \in [0,1]$. If $\alpha = 1$, a Ridge regression is performed, and if $\alpha = 0$, a Lasso regression is performed.

The reconstruction is obtained by applying the estimated regression coefficients $\hat{\beta}^{enet}$ on $x^1, \dots, x^p$.

$$\hat{Y}_{\lambda,\alpha} = \sum_{j=1}^{p} x^j \, \hat{\beta}_j^{enet}$$

The optimization of α and λ is performed using KFCV (*cf*. Methods, section "K-Fold cross-validation (KFCV)") for both by crossing different possible values for each of them. As they respectively take their values in the continuous sets $[0,1]$ and $\mathbb{R}^p$, they have to be discretized. The more they are, the more robust the reconstruction will be, at the expanse of the computational time.

Random Forest (RF):

The RF regression has been introduced in the early 21st century[52] and consists in aggregating regression trees.

We denote each set of predictand/predictor $\{(Y_i, x_i)\}_{1 \leq i \leq n}$ put on the root of the tree. The first step consists in cutting that root in two child nodes. A cut is defined as: $\{x^j \leq d\} \cup \{x^j \geq d\}$. Where $j \in \{1, ..., p\}$and $d \in R$. Cutting a node with $\{x^j \leq d\} \cup \{x^j \geq d\}$, means that the years of observations for which the j^{th} value of the proxy record is lower than d are placed in the left child node $c_1$and the others in the right child node c_2. The method selects the best pair (j, d) that minimizes a loss function. Here, we aim at minimizing the variance of Yin each child node. The variance of a given node c is defined as:

$$\sum_{i:\, X_i \in c} (Y_i - \bar{Y}_c)^2$$

Where $\bar{Y}_c = \frac{1}{card(c)} \sum_{i: X_i \in c} Y_i$.

Two subsets of $\{(Y_i, x_i)_{1 \leq i \leq n}\}$ are thus obtained through the optimal cut: $\{(Y_i, x_i)_{i \in c_1}\}$ and $\{(Y_i, x_i)_{i \in c_2}\}$.

The same procedure is recursively applied to the child nodes $c_1$and c_2. We then stop these recursive calculations when a chosen depth of the tree is reached.

There exists different kind of regression trees[14] in random forest, the commonly used regression trees are called random-input regression trees. It consists in only a randomly selected set of $q < p$variables between $(x^j)_{1 \leq j \leq p}$ used for constructing the tree. A large number of K random-input regression trees is computed. For each tree, $q < p$ proxy records are randomly selected with probability $\frac{1}{p}$and the method is applied until the depth of the tree reaches m. It should be noted that for each tree, the q selected variables can contain variables not used if it does not give any optimal cut through the different nodes. Thereby, a single variable can be used more than one time in the same tree.

The reconstruction is obtained by splitting each testing series in the different random input regression trees previous constructed. In each tree, the estimation attributed to an observation is the empirical average of Y inside the node where the corresponding observation ends up, given the cuts made on the corresponding predictors. For each testing series, the K reconstructions are averaged to give the final reconstruction.

A priori, this method requires the optimization of two parameters: the number of trees K and the number of proxy records for each tree m. In practice, K does not require to be tuned, as long as K is large given p, which guarantees convergent reconstructions for a given m. m is the only parameter to optimize here. KFCV (*cf*. Methods, section "K-Fold Cross-Validation (KFCV)") is then applied to optimize m, with a high K (here set to 200, different K than the KFCV one), to empirically select the best RF model.

We here simplified the Random Forests theory such that only the main steps are presented. However, this theory is complex and for more information, the reader can refer to ref. 52.