SUPPLEMENTARY MATERAL

COMPARISON OF METHODS TO AGGREGATE CLIMATE DATA TO PREDICT CROP YIELD: AN APPLICATION TO SOYBEAN

Mathilde CHEN [a]*, Nicolas GUILPART [b], David MAKOWSKI [a]

[a] Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA PS, 91120 Palaiseau, France

[b] Université Paris-Saclay, AgroParisTech, INRAE, UMR Agronomie, 91120 Palaiseau, France

*** Correspondence:**

Corresponding author

mathilde.chen@inrae.fr

List of Supplementary Material

List of Supplementary Figures

List of Supplementary Tables

Supplementary Material 1. Procedure of yields data detrending

To avoid any confusion with technological progress due to improved cultivars and technological progress, yield data were detrended. For a given site, yield time series are used to fit a cubic smoothing spline *f(t)*. Yield for each year *t* of this site is then expressed as:

$$Detrended\ yield(t) = f(t_{max}) + A(t)$$

with $f(t_{max})$, defined as the expected yield value at the most recent year of available data for this site (generally 2015 or 2016), and *A(t)*, the yield anomaly, calculated as the difference between expected yield *f(t)* and actual yield in year *t*.

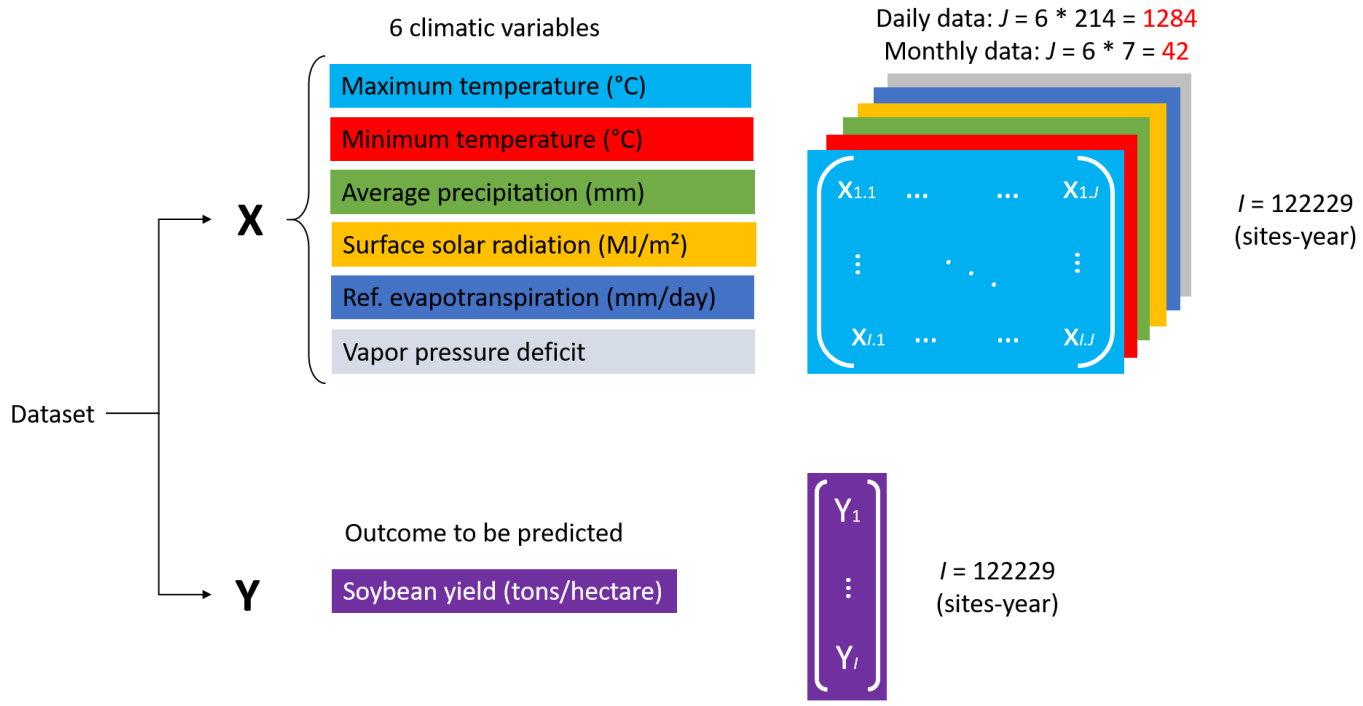
Supplementary Material 2. Dimension reduction techniques applied on daily and monthly climatic predictors

0. Common notations

Each unique combination of site and year is defined by a set of climatic predictors (**X**) and an outcome (**Y**) to be predicted.

The dataset of predictors comprises I observations described by J climatic predictors and it is represented by the $I \times J$ matrix **X**, whose generic element is $x_{i,j}$. The vector of outcome contains also I observations, each individual element is noted as $y_{i,}$ (see below). For each reduction dimension technique, **X** is centered and standardized, so that each column has a mean of 0 and a standard deviation of 1.

Figure. Illustration of data structure and notations



1. Principal component analysis (PCA)

PCA analyzes a data table representing observations described by several dependent variables or predictors **X**, which are, in general, inter-correlated. Its goal is to extract the important information from **X** and to express this information as a set of new orthogonal variables called *principal components*.[1]

The central concept of PCA is to compute new variables called principal components from linear combinations of the original predictors. The first principal component is required to have the largest possible variance and

therefore this component will explain the largest part of the variance in **X**. The second component is computed under the constraint of being orthogonal to the first component and to have the largest possible explained variance.

Briefly, PCA looks for linear combinations of predictors:

$$f_i = \sum_{j=1}^{J} \beta_j x_{ij}, \qquad with\ i = 1, \ldots, I \tag{1}$$

where β_j is a weighting coefficient applied to observed values of $x_{i,jj}$ of the *j*th predictor. The weights are chosen to display the highest source of variation in **X**, following a stepwise procedure:

Step 1: find the vector $\xi_1 = (\xi_{11}, \xi_{J1})'$ for which the linear combination values:

$$f_{i1} = \sum_{j=1} \xi_{j1} x_{ij} = \xi'_1 x_i$$

have the largest possible mean square $\frac{1}{I}\sum_i {f_{1i}}^2$ and are subject to the constraint: $\sum_{j=1} {\xi_1}^2 = ||\xi_1||^2 = 1$

Step 2 to *m*: carry out second and subsequent steps, possibly up to a limit of the number of variables *J*. On the m^{th} step, compute a new weight vector ξ_m with components ξ_{jm} and new values $f_{im} = \xi'_m x_i$. Thus, the values f_{im} have maximum mean square, subject to the constraints:

$$||\xi_m||^2 = 1 \qquad \text{and} \qquad \sum_{j=1} \xi_{jk}\, \xi_m = 0, \qquad k < m$$

By maximizing the mean square, the strongest and most important mode of variation in the variables is identified. On second and subsequent steps, the most important mode of variation is seek again, but it requires that the weights are orthogonal to those identified previously, so that they are indicating something new. The amount of variation measured declines on each step. At some point, usually well short of the maximum index *J*, we expect to lose interest in modes of variation thus defined.

The values of the linear combinations f_{im} are called *principal component scores*, and these scores are used as the new inputs of the predictive models.

In this study, PCA is performed using the *prcomp()* function of R (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp).

2. Functional principal component analysis (FPCA)

Several papers showed that the computation of PCA was a key issue when data are sparse and in a high-dimensional space. In this case, the sample covariance matrix estimated by PCA is sometimes a poor estimate of the population covariance matrix. FPCA, a popular multivariate analysis technique for extracting information from functional data,[2-4] can be used to overcome this difficulty.

Similarly to PCA, FPCA reduces the dimensions of a data set in which there are a large number of interrelated variables, while still holding as much of the total variation as possible. Functional **X** is transformed in a new set

of principal components functions that are uncorrelated and ordered so that the first few retain most of the variation present in all of the original variables. Compared to usual PCA, FPCA takes into account the order of the observations in time.

In the context of functional data, observations are now function values noted $x_i(s)$, where the discrete index j in the PCA context has been replaced by the continuous index *s* (in °C, mm, MJ/m², or mm/day depending on the considered climatic variable). Similarly, the weights β_j become functions with values $\beta_j(s)$. As former vectors β and x are now functions $\beta_j(s)$ and $x(s)$, the appropriate way to combine them to compute the principal components is to replace summations over *j* by integrations over the continuous index *s* in equation (1):

$$f_i = \int \beta x_i = \int \beta(s)\, x_i(s)\, ds \qquad (2)$$

The following procedure is then applied to determine the weights:

Step 1: In an analogous way to classical PCA, find the weight function $\xi_1(s)$ as:

$$f_i = \int \xi_1(s)\, x_i(s)\, ds$$

which maximizes the mean square $\frac{1}{I} \sum_i {f_{1i}}^2$ and is subject to the constraint: $\int \xi_1(s)^2 ds = 1$

Step 2 to *m*: Next, compute weight function $\xi_m(s)$ and principal component scores maximizing $\frac{1}{I} \sum_i {f_{im}}^2$, subject to the constraint $\int \xi_m(s)^2 ds = 1$ and the additional constraint $\int \xi_{m-1}(s)\, \xi_m(s)\, ds = 0$ (to guarantee orthogonality) and so on as required.

Several approaches have been proposed to overcome integration in FPCA computation. One of them is to express each functional observation as a linear combination of basis functions, and to approximate each function by a finite number of basis functions.[2] The B-spline basis is very flexible and able to approximate complex non-periodic functions by polynomial segments.

The standardized time series of daily cumulative climate data were projected onto a B-spline basis to get values at regular and identical time intervals in the range of 1 to 214 using 150 splines of order 4. The number of B-splines for analyses conducted on daily data was chosen based on previous work.[5] Standardized time series of monthly data were projected in the range of 1 to 7 using 7 splines of order 4.

The projection was performed using the *create.bspline.basis()* function (https://www.rdocumentation.org/packages/fda/versions/6.1.4/topics/create.bspline.basis) of the fda R package (version 6.1.4; https://cran.r-project.org/web/packages/fda/index.html). FPCA was performed using the *pca.fd()* function (https://www.rdocumentation.org/packages/fda/versions/6.1.4/topics/pca.fd) of the same package.

3. Multivariate functional principal component analysis (MPCA)

In recent years, the study of multivariate functional data that takes multiple functions at the same time into account, has led to new insights. Each observation unit here consists of a fixed number *P* of functions $X = (X^{(1)}, \ldots, X^{(P)})$.[6] In our case, *P*=6, as 6 climatic variables are considered.

The basic idea of MFPCA is to extend functional principal component analysis to multivariate functional data on different dimensional domains. Multivariate functional principal components and scores are estimated based on their univariate functional counterparts, as presented in the article of Happ and Greven.[6]

4. Partial least square regression (PLSR)

PLSR[7] generalizes and combines PCA and multiple regression, by transforming initial predictor matrix (**X**) and outcome vector (**Y**) into a reduced number of latent variables, which are included one by one in an iterative process. These factors will be the new variables of a usual linear regression. The main idea is to perform successive regressions by projections onto latent structures to reveal hidden or latent underlying biological effects.

PLSR looks for a decomposition of centered data matrices **X** and **Y** in terms of component scores, called latent variables: ($\xi_1,\ldots, \xi_H$) and ($\omega_1,\ldots, \omega_H$), which are linear combinations of the columns of **X** and **Y** respectively, and associated loading vectors: ($\mathbf{u}_1,\ldots, \mathbf{u}_H$) and ($\mathbf{v}_1,\ldots, \mathbf{v}_H$), where H is the number of latent variables retained in the final model. However, the regression coefficients that define these components are not linear, as they are solved via successive local regressions on the latent variables. The loading vectors are estimated to solve the following optimization problem:

$$max_{u_h\,=\,1, v_h\,=\,1}\; cov(\,X_{h-1}\,u_h\,, Y\,v_h\,)$$

where $\mathbf{X}_{h-1}$ is the residual **X** matrix in the regression of **Y** on ($\xi_1,\ldots, \xi_{h-1}$) for each dimension $h = 1, \ldots, H$, and the associated latent variables are denoted $\xi_h = \mathbf{X}_{h-1}\,\mathbf{u}_h$ and $\boldsymbol{\omega}_h = \mathbf{Y}\mathbf{v}_h$.

As in principal component analysis, the loading vectors and the latent variables are directly interpretable. The loading vectors $\mathbf{u}_h$ and $\mathbf{v}_h$ indicate how the x_j and y_i variables explain the relationship between **X** and **Y**. The latent variables contain information regarding similarities or dissimilarities between individuals.

In this study, PLSR is performed using the *plsr()* function (https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf) of the pls R package (version 2.8-2, https://cran.r-project.org/web/packages/pls/index.html). The optimal number of components is chosen based on a 10-fold cross-validation procedure implemented in the *pls()* function.

5. Functional partial least square regression (FPLSR)

PLSR has recently been generalized to the case of a functional predictor, here after referred as FPLSR.[8] It consists of building PLSR components as linear functionals of the predictor $\{X(t):t\in T\}$

$$t \;=\; \int_0^T X(t)\;\omega\,(t)\;dt$$

obtained by maximizing the squared covariance:

$$max_{\omega,||\omega||^2=1}\; Cov^2(\int_0^T X(t)\;\omega(t)\;dt, Y)$$

In the multivariate context this maximization problem is known as Tucker's criterion. FPLS algorithm and computational details are provided in the articles of Aguilera et al.[8] and Febrero-Bande and de la Fuente.[9]

FPLSR is performed using the *fdata2pls()* function (https://www.rdocumentation.org/packages/fda.usc/versions/0.9.4/topics/fdata2pls) of the fda.usc R package (version 0.9.4, https://cran.r-project.org/web/packages/fda.usc/).

References:

1 Jolliffe, I. in *International Encyclopedia of Statistical Science* (ed Miodrag Lovric) 1094-1096 (Springer Berlin Heidelberg, 2011).

2 Ramsay, J. O. & Silverman, B. W. *Functional Data Analysis*. 428 (Springer-Verlag New York, 2005).

3 Ullah, S. & Finch, C. F. Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology* **13**, 43, doi:10.1186/1471-2288-13-43 (2013).

4 Wang, J.-L., Chiou, J.-M. & Müller, H.-G. Functional Data Analysis. *Annual Review of Statistics and Its Application* **3**, 257-295, doi:10.1146/annurev-statistics-041715-033624 (2016).

5 Bonneu, F., Makowski, D., Joly, J. & Allard, D. Machine Learning Based on Functional Principal Component Analysis to Identify Major Influential Factors of Wheat Yield. *SSRN Electronic Journal*, doi: 10.2139/ssrn.4207476 (2022).

6 Happ, C. & Greven, S. Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association* **113**, 649-659, doi:10.1080/01621459.2016.1273115 (2018).

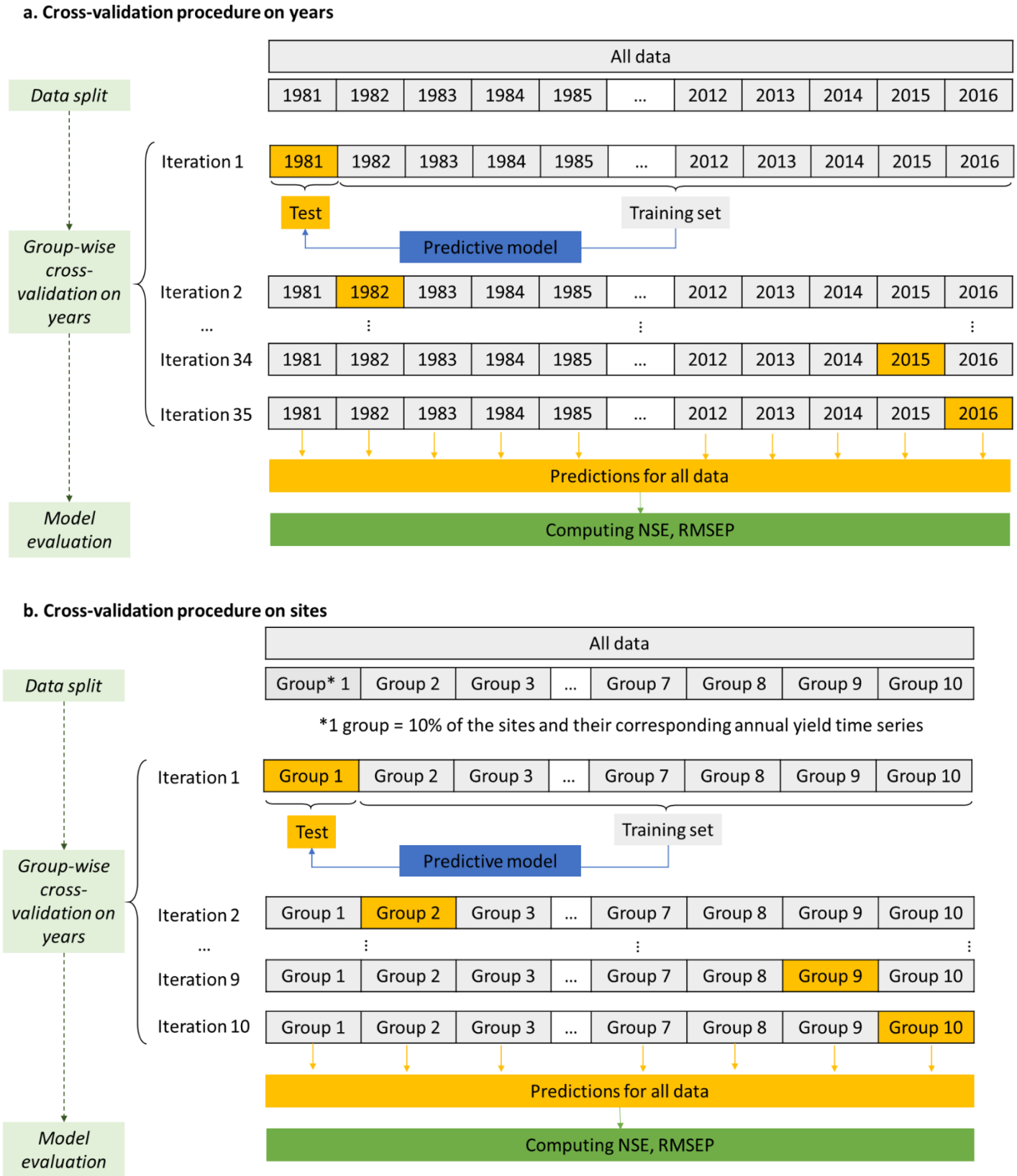
7 Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**, 109-130, doi:https://doi.org/10.1016/S0169-7439(01)00155-1 (2001).

8 Aguilera, A. M., Escabias, M., Preda, C. & Saporta, G. Using basis expansions for estimating functional PLS regression: Applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems* **104**, 289-305, doi:https://doi.org/10.1016/j.chemolab.2010.09.007 (2010).

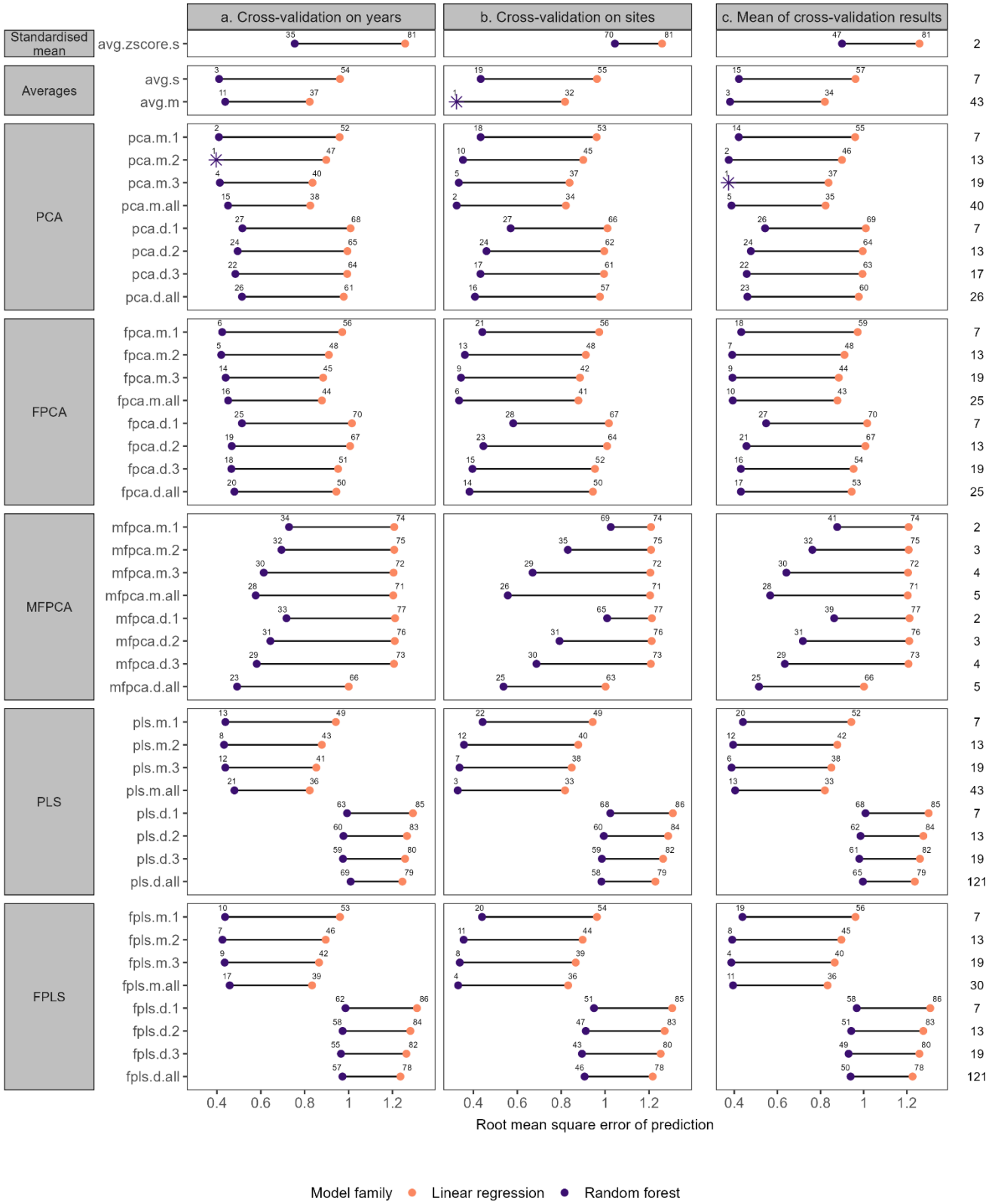
9 Febrero-Bande, M. & de la Fuente, M. O. Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software* **51**, 1 - 28, doi:10.18637/jss.v051.i04 (2012).

SUPPLEMENTARY FIGURES

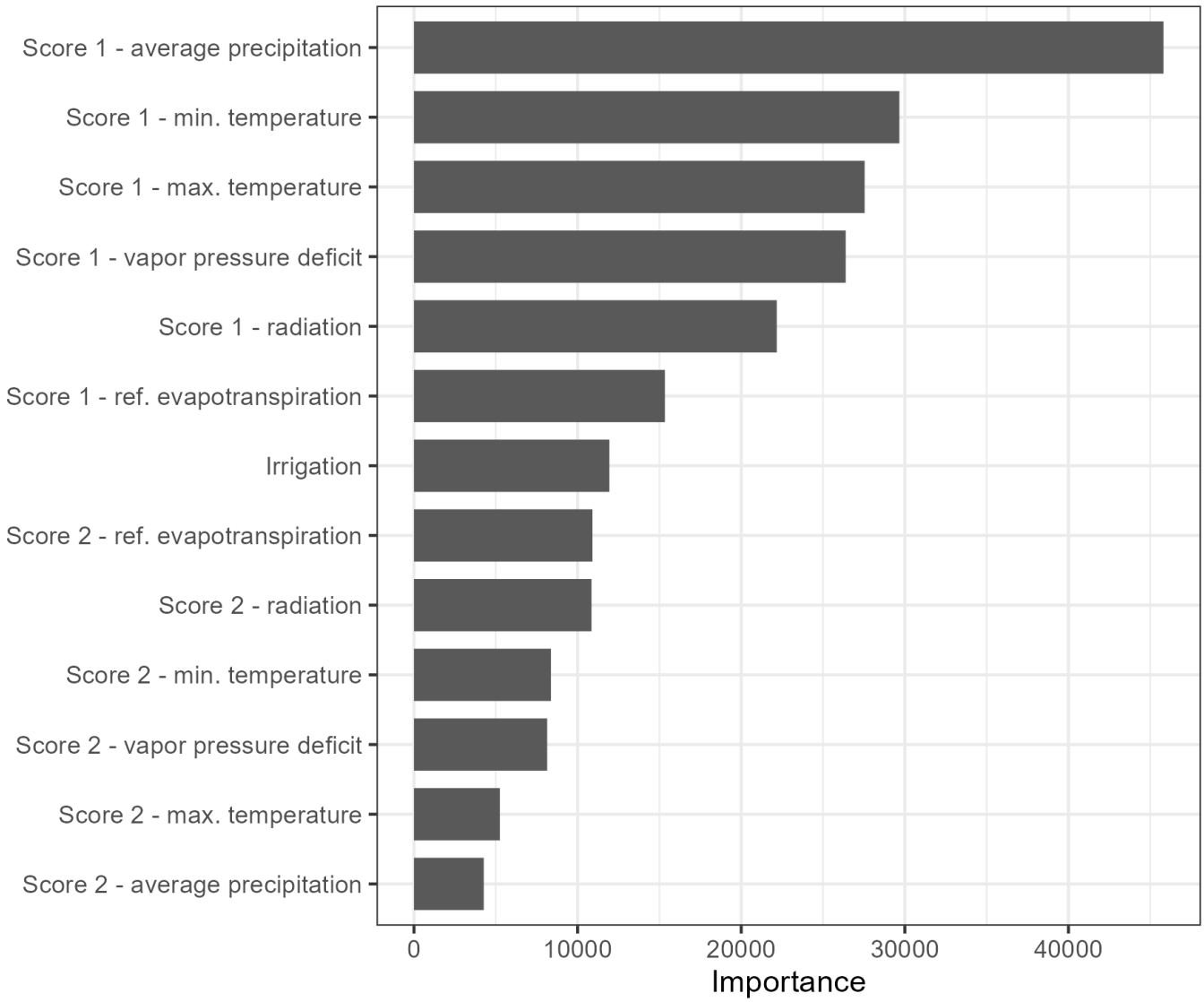
Supplementary Figure 1. Schematic representation of cross-validation procedures applied to evaluate predictive performances of tested models.



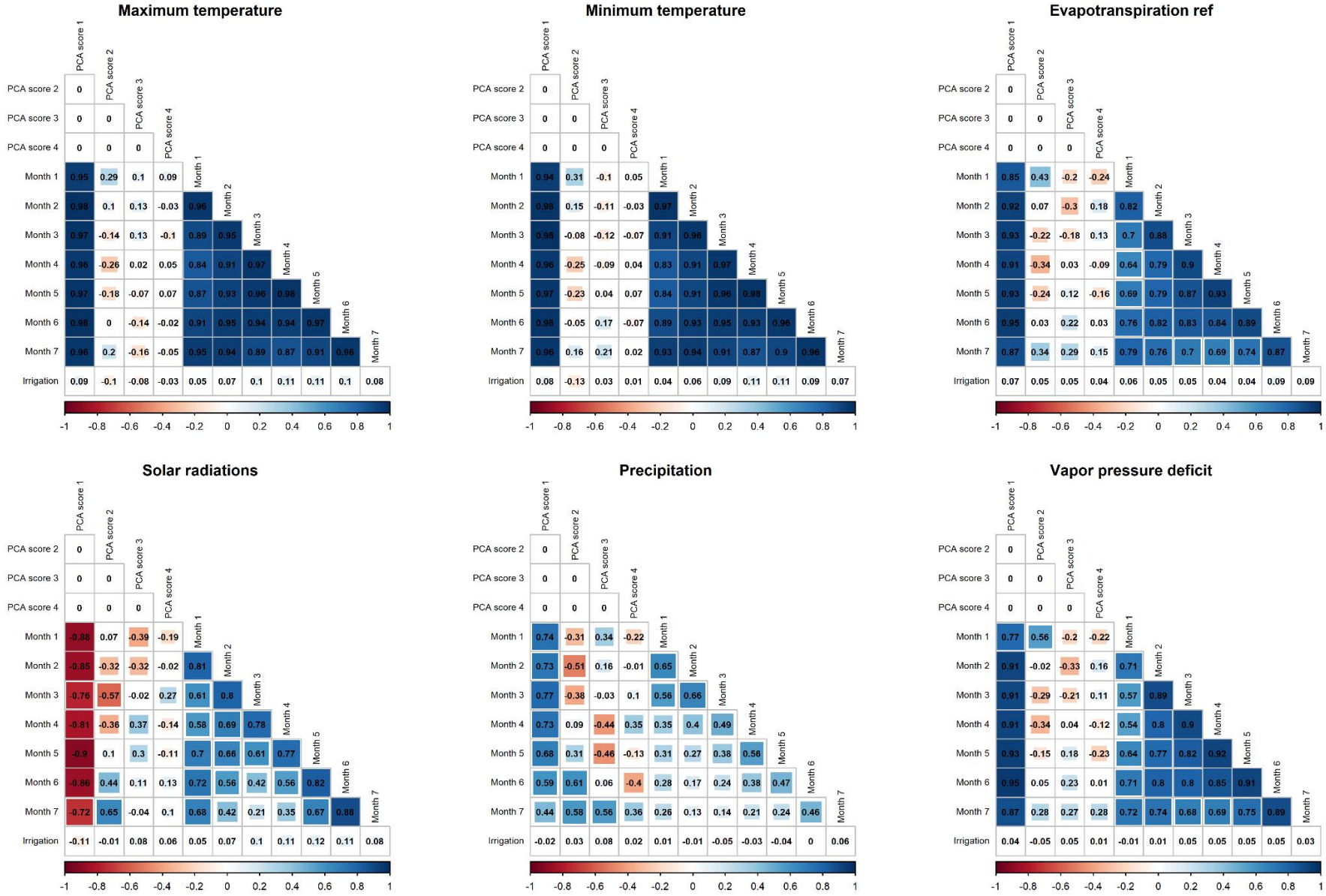
Supplementary Figure 2. Root mean square error of climate-based models predicting soybean yield estimated by cross-validation on years (a), sites (b), and on average (c). Lower value indicates better performance. For each column, model's ranking is indicated above corresponding dots and the best model is highlighted by a *. Number of predictors is indicated on the right of the figure. See Table 1 for models' name abbreviations.



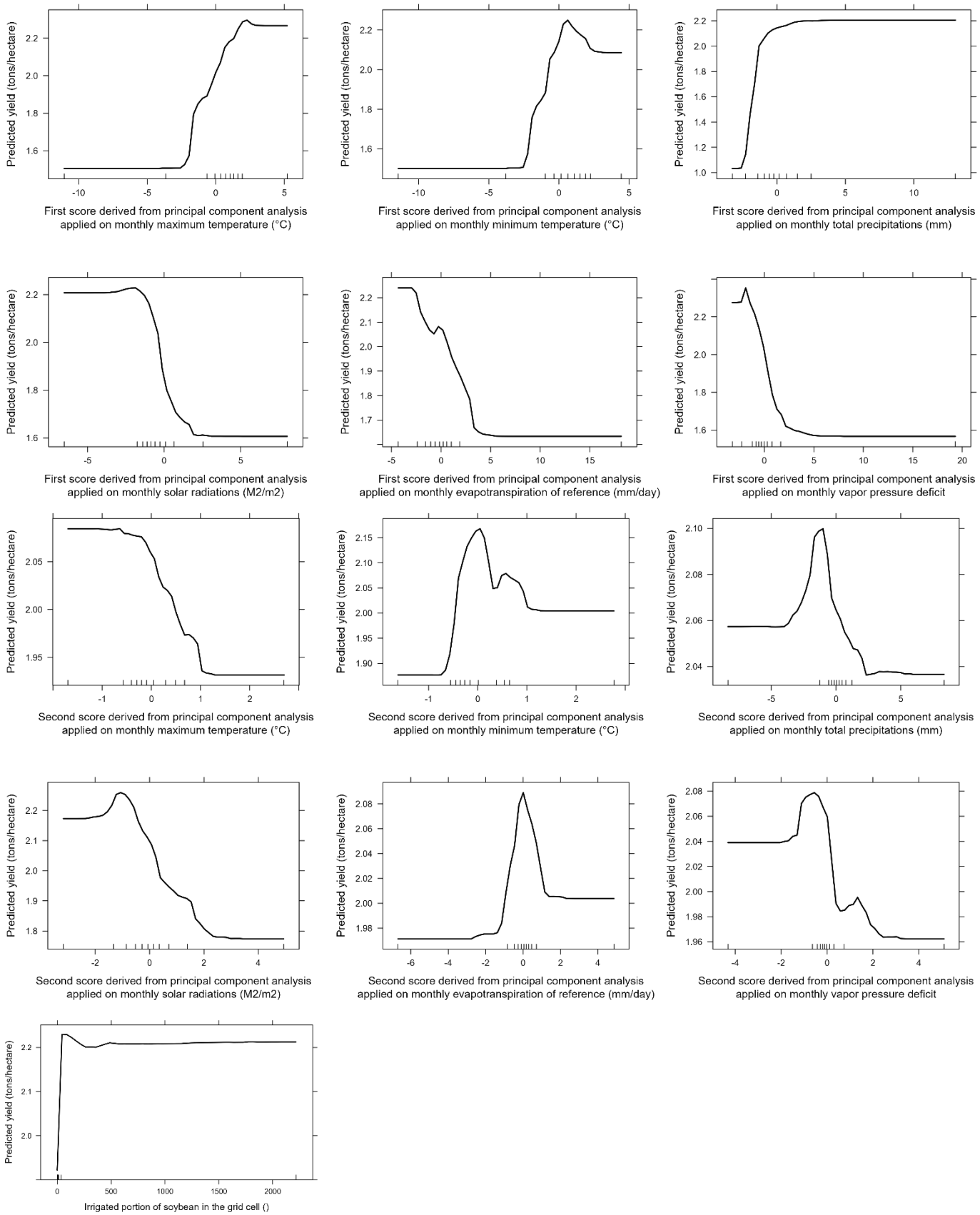
Supplementary Figure 3. Importance of predictors in the model showing best performance to predict soybean yield at global scale. Best model was the random forest model including the first ('score 1') and second ('score 2) components derived from principal component analysis applied on monthly climate variables. Predictors were ranked in decreasing order of importance in the model.



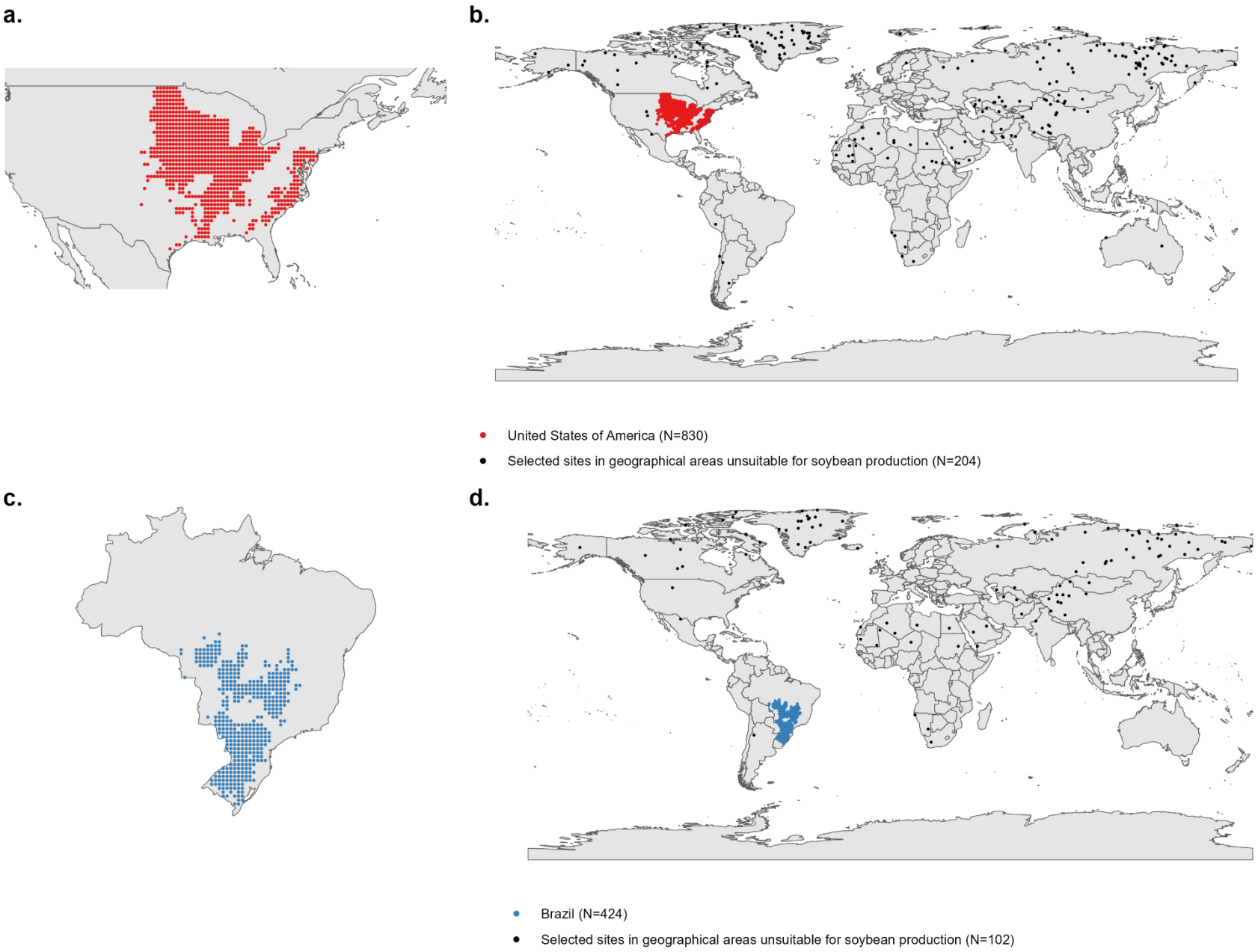
Supplementary Figure 4. Pearson correlation coefficient between monthly averages of climate data, proportion of irrigated soybean ("Irrigation"), and scores derived from principal components analysis (PCA) applied on monthly averages. Coefficients were computed on the full dataset (N=122,229).



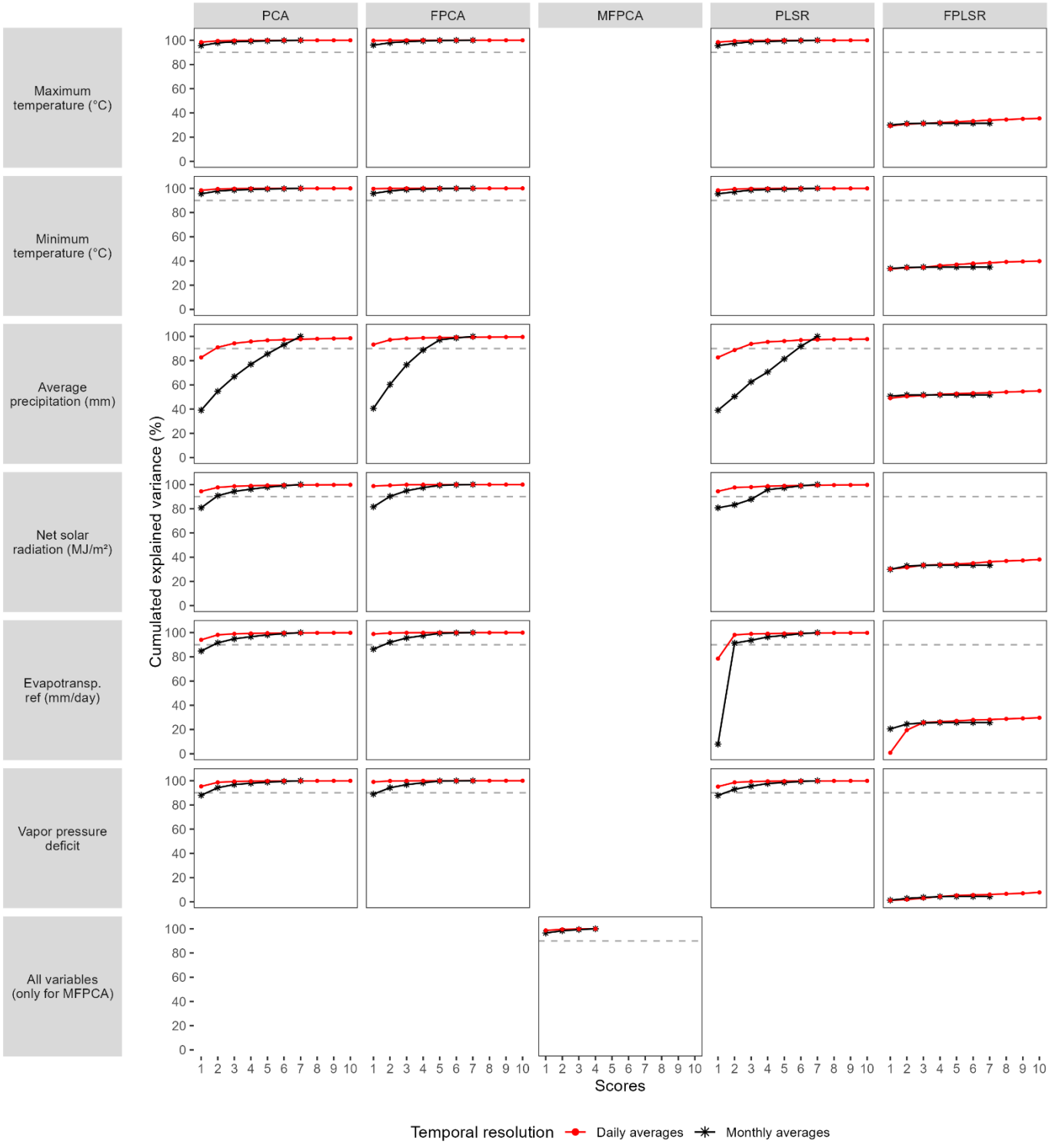
Supplementary Figure 5. Partial dependency plot for each variable of the model showing best predictive performances at global scale. Best model was the random forest model including the first and second components derived from principal component analysis applied on monthly climate variables.



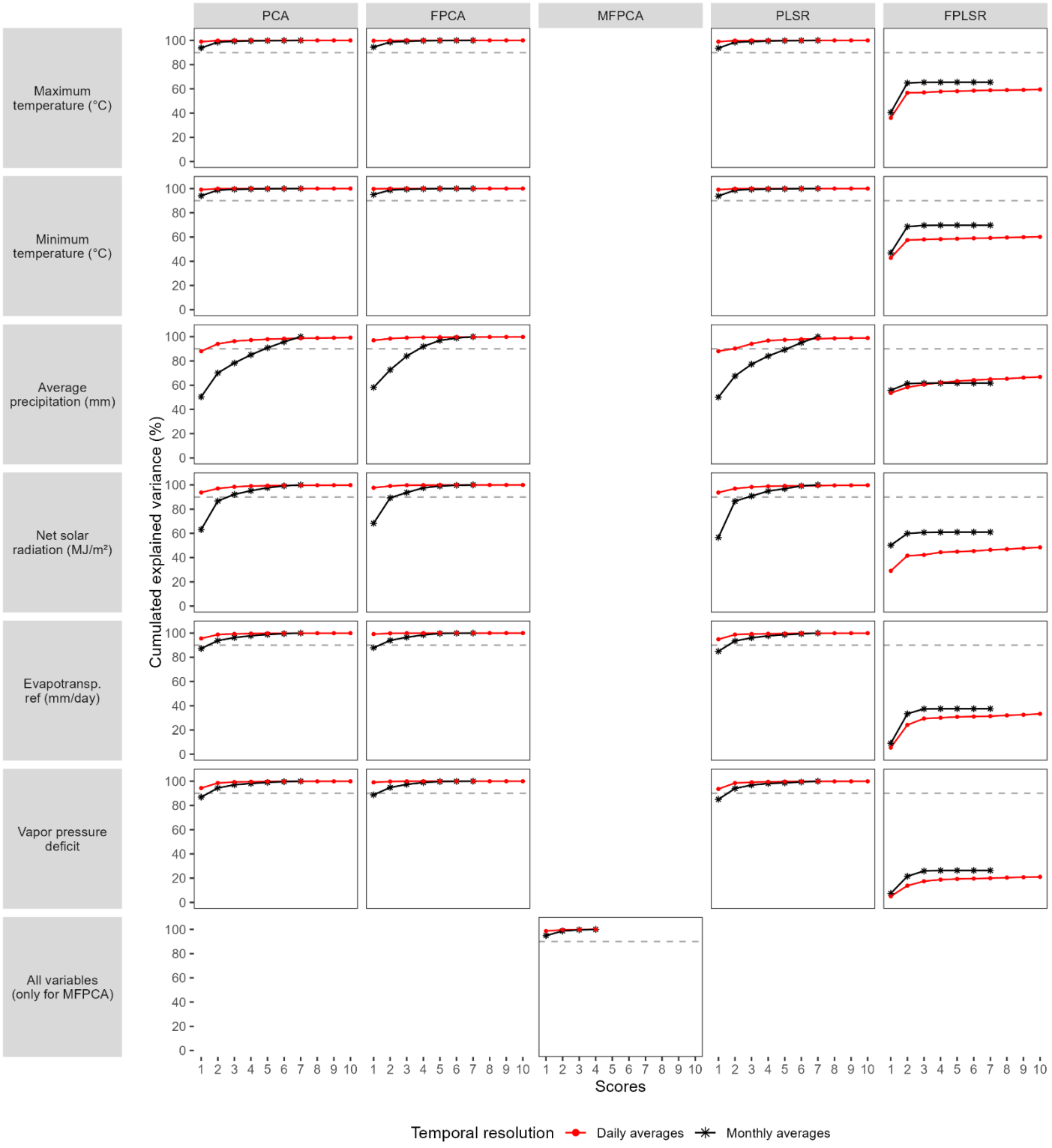
Supplementary Figure 6. Sites distribution for the analyses conducted (a and b) in the US and (c and d) in Brazil.



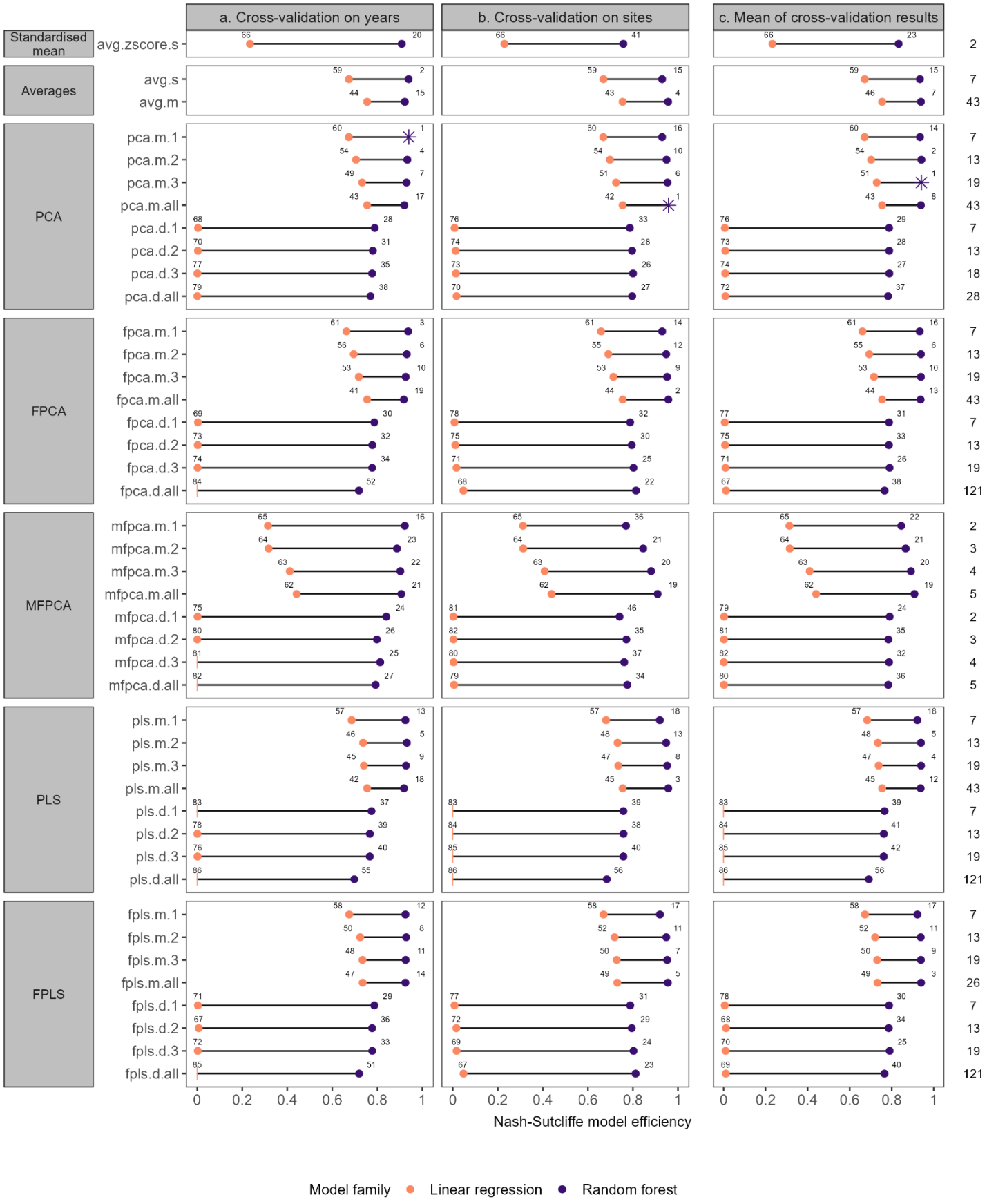
Supplementary Figure 7. Percentage of variance of climatic data in the US explained by the different dimension reduction techniques applied. Representation limited to the first 10 components. Abbreviations: PCA: principal component analysis; FPCA: functional principal component analysis; MFPCA: multivariate functional principal component analysis; PLSR: partial least square regression; FPLSR: functional partial least square regression.



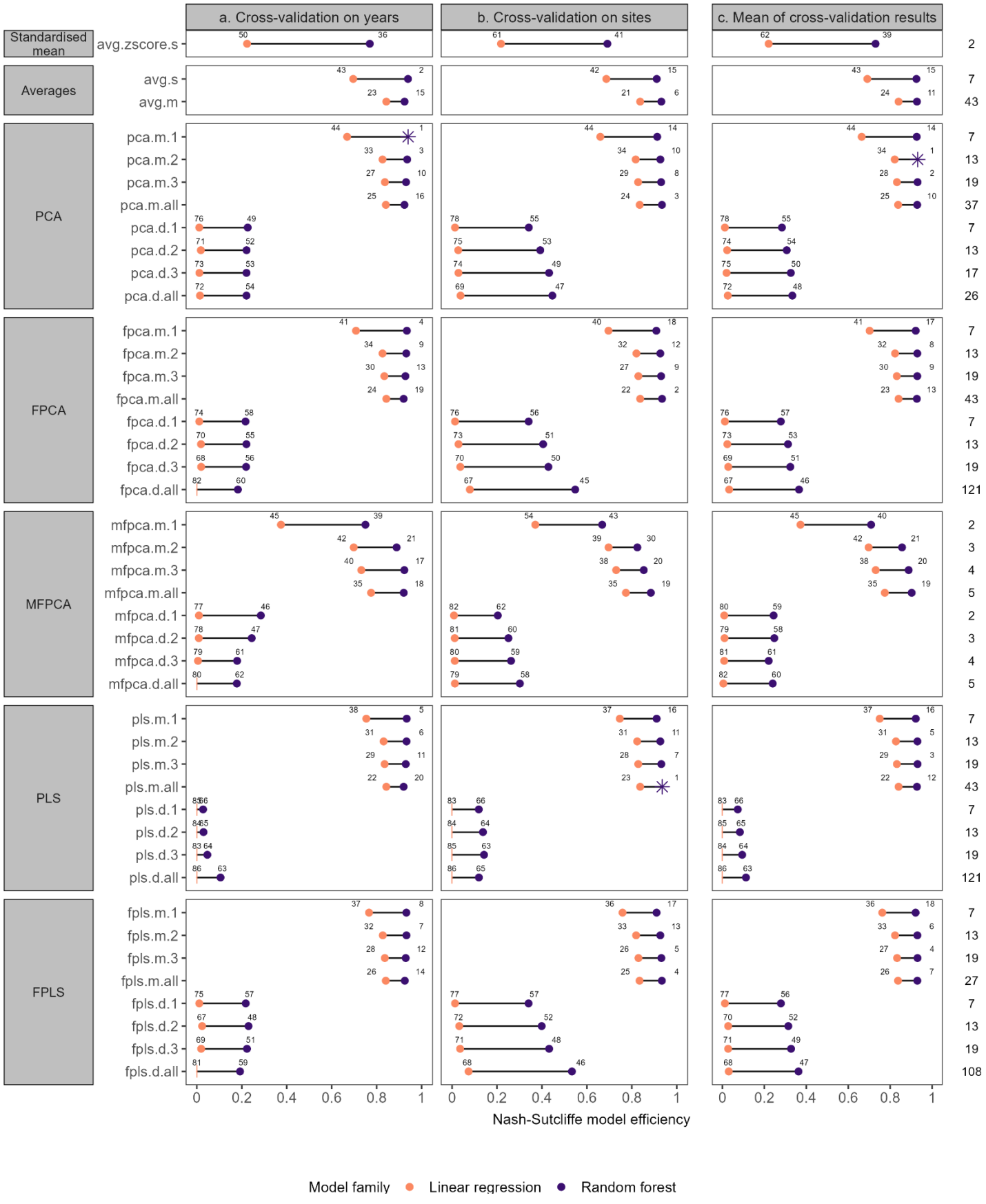
Supplementary Figure 8. Percentage of variance of climatic data in Brazil explained by the different dimension reduction techniques applied. Representation limited to the first 10 components. Abbreviations: PCA: principal component analysis; FPCA: functional principal component analysis; MFPCA: multivariate functional principal component analysis; PLSR: partial least square regression; FPLSR: functional partial least square regression.



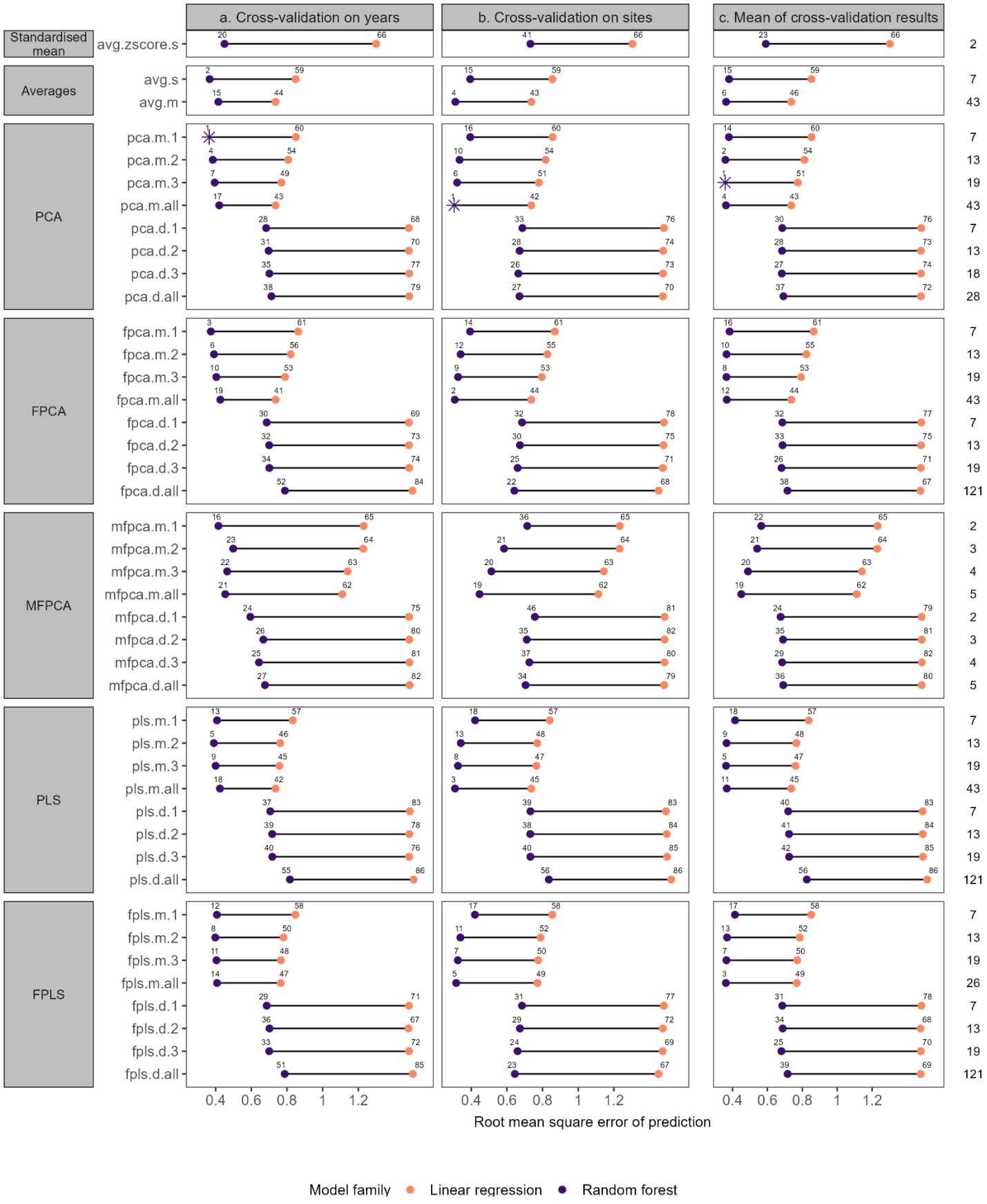
Supplementary Figure 9. Nash-Sutcliffe efficiency of climate-based models predicting soybean yield in the United-States estimated by cross-validation on (a) years, (b) sites, and (c) on average.



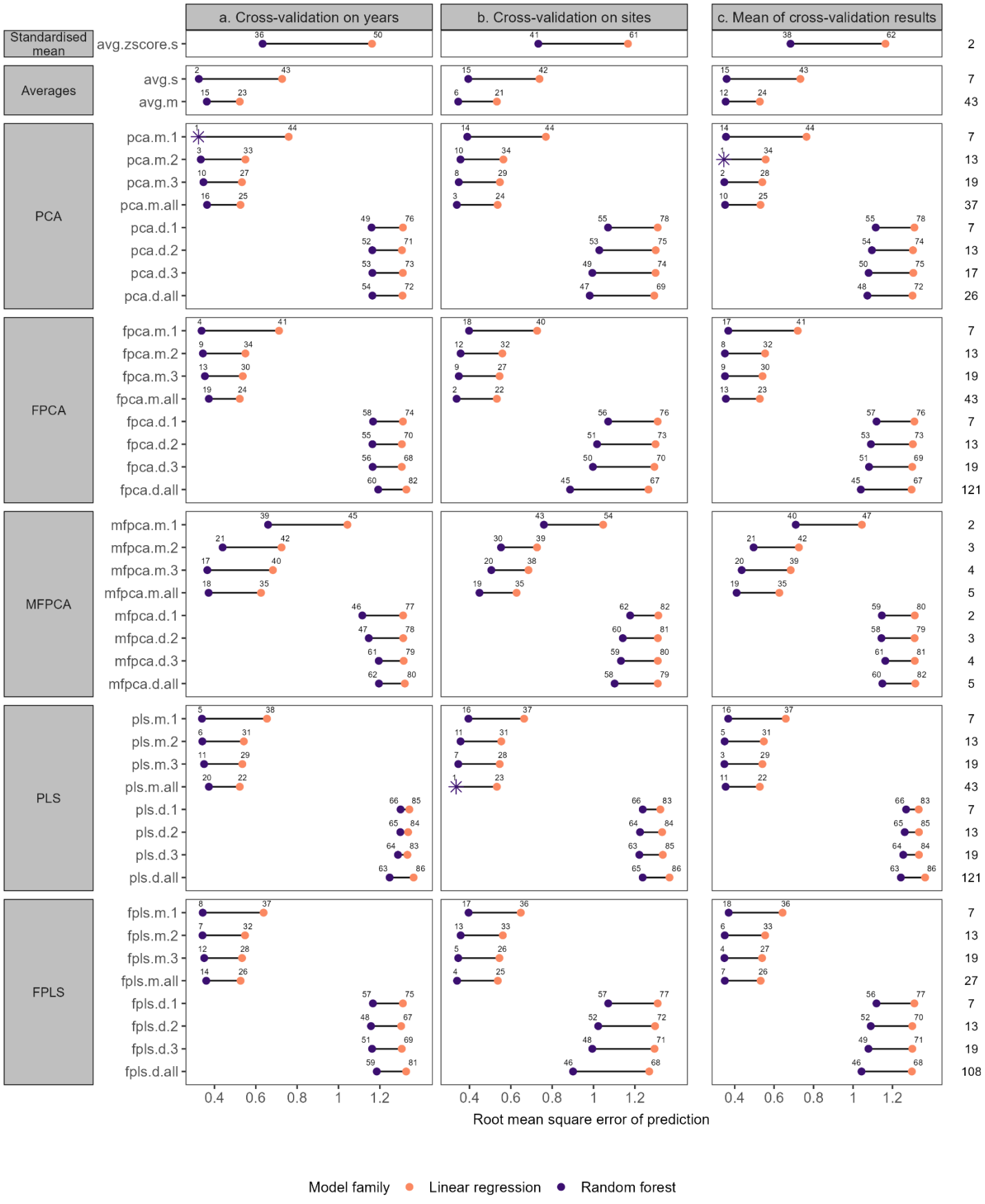
Supplementary Figure 10. Nash-Sutcliffe efficiency of climate-based models predicting soybean yield in Brazil estimated by cross-validation on (a) years, (b) sites, and (c) on average.



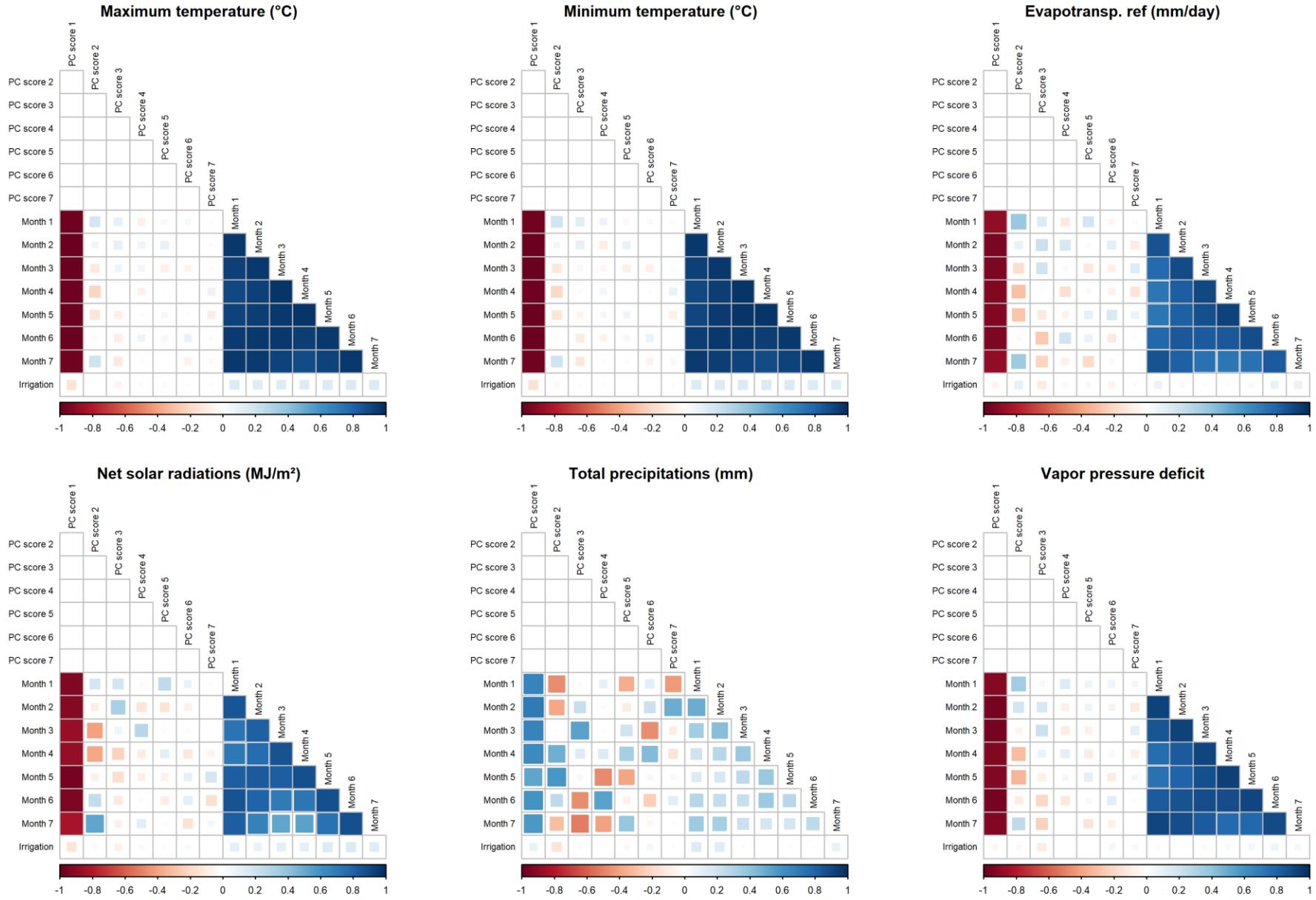
Supplementary Figure 11. Root mean square error of climate-based models predicting soybean yield in the US estimated by cross-validation on (a) years, (b) sites, and (c) on average.



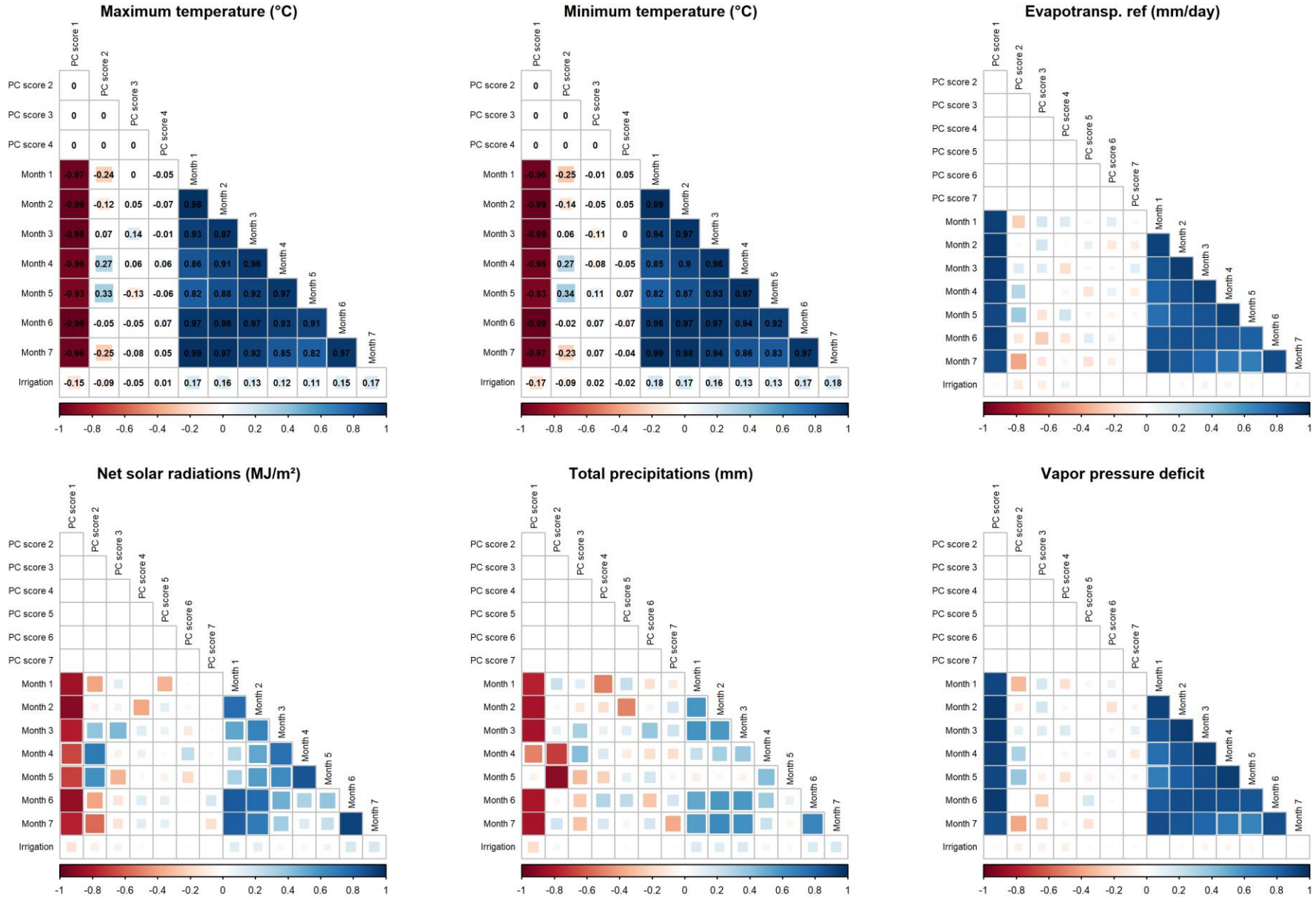
Supplementary Figure 12. Root mean square error of climate-based models predicting soybean yield in Brazil assessed estimated by cross-validation on (a) years, (b) sites, and (c) on average.



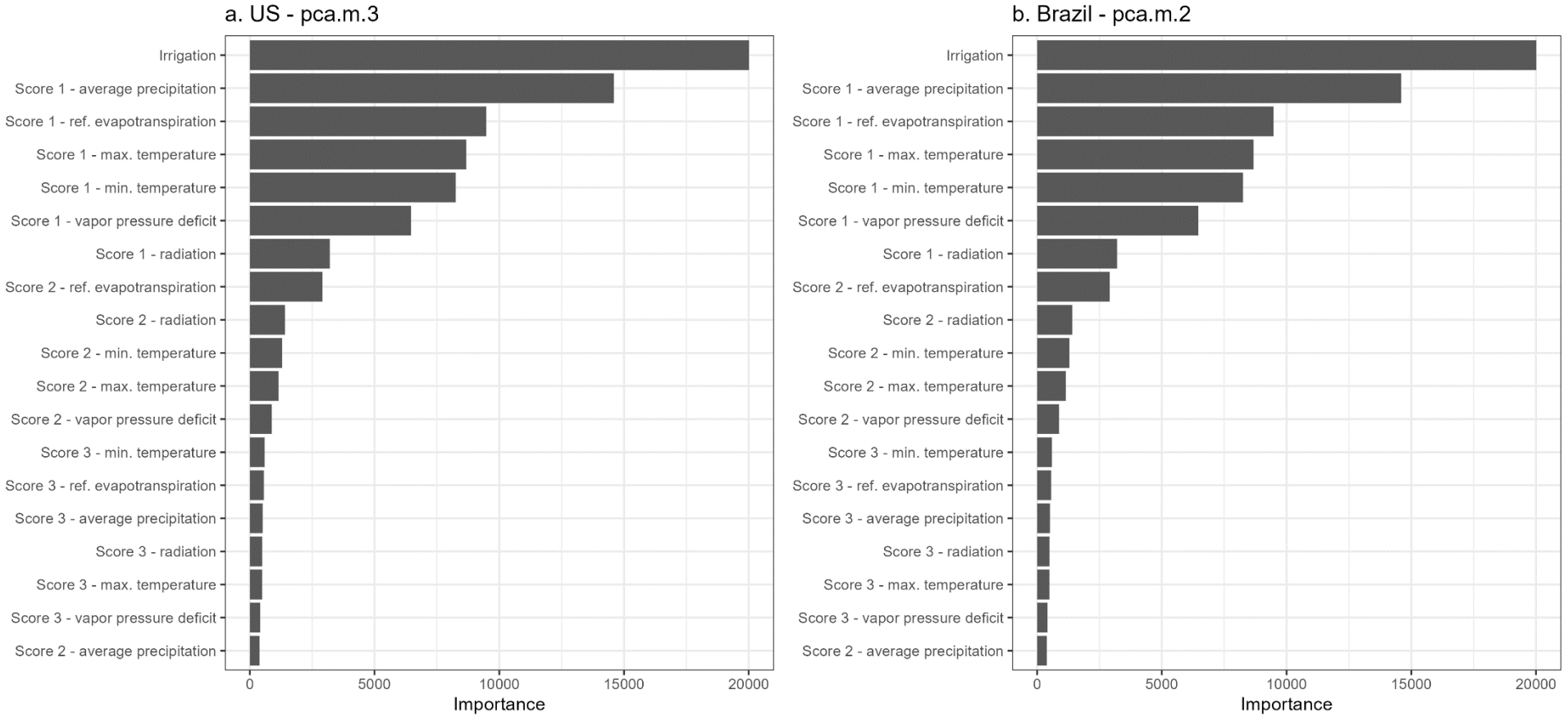
Supplementary Figure 13. Correlation between monthly averages of climate data, proportion of irrigated soybean ("Irrigation"), and scores derived from principal component analysis (PCA) applied on monthly averages in the US.



Supplementary Figure 14. Correlation between monthly averages of climate data, proportion of irrigated soybean ("Irrigation"), and scores derived from principal component analysis (PCA) applied on monthly averages in Brazil.



Supplementary Figure 15. Importance of predictors in the random forest models showing best performance to predict soybean yield (a) in the US and (b) in Brazil. Best models for at US and Brazil scales were the RF based on three or two scores ("scores 1", "score 2", or "score 3") derived from principal component analysis based on monthly climate data, respectively.



SUPPLEMENTARY TABLES

Supplementary Table 1. Variables describing climate during the growing season from ERA5-land dataset.

Variable	Short description	Pre-processing	Code name of Copernicus Climate Data Store variables used to compute climatic predictors	Unit
Maximum and minimum temperatures	Temperature of air at 2m above the surface of land, sea or in-land waters.	Initially in K, converted in °C	max_2m_temperature min_2m_temperature	°C
Average precipitation	Average of precipitation height, accumulated liquid and frozen water, comprising rain and snow, that falls to the Earth's surface. This is not cumulated precipitation over the day, even if the name of the variable used to extract data from CDS refers to "total precipitation".	Initially in m, converted in mm	total_precipitation	mm
Surface net solar radiation	Amount of solar radiation (also known as shortwave radiation) that reaches a horizontal plane at the surface of the Earth (both direct and diffuse) minus the amount reflected by the Earth's surface (which is governed by the albedo).	Initially in $J.m^{-2}$, converted in $MJ.m^{-2}$	surface_net_solar_radiation	$MJ.m^{-2}$
Vapor pressure deficit	Difference (deficit) between the amount of moisture in the air and how much moisture the air can hold when it is saturated.	Computed from mean 2m temperature and mean 2m dewpoint temperature	2m_temperature (min and max) 2m_dewpoint_temperature (min and max)	kPa

Variable	Short description	Pre-processing	Code name of Copernicus Climate Data Store variables used to compute climatic predictors	Unit
Reference evapotranspiration	The evapotranspiration rate from a reference surface (amount of water lost from a cropped surface), not short of water. The reference surface is a hypothetical grass reference crop with specific characteristics.	Computed based on Penman-Monteith method based on radiation, air temperature, humidity and wind speed.	2m_temperature (min and max), 2m_dewpoint_temperature (min and max) 10m_u_component_of_wind, 10m_v_component_of_wind, surface_net_solar_radiation, surface_pressure	$mm.day^{-1}$

Supplementary Table 2. Mean and standard deviation of soybean yield and climate data over the 1981-2016 period in the US and Brazil datasets.

	US dataset		Brazil dataset	
	Sites located in		Sites located in	
	soybean producing areas only (29,803 sites-year)	soybean producing & unsuitable areas (37,147 sites-year)	soybean producing areas only (14,757 sites-year)	soybean producing & unsuitable areas (18,429 sites-year)
Soybean production				
Yield, in tons/ha	3.38 (0.70)	2.71 (1.49)	2.99 (0.62)	2.4 (1.32)
Seasonal climate data				
Maximum temperature, in °C	24.80 (2.90)	21.96 (9.61)	28.27 (1.82)	24.7 (10.45)
Minimum temperature, in °C	14.10 (3.00)	11.58 (8.25)	19.14 (2.10)	15.6 (9.51)
Average precipitations,[1] in mm	0.13 (0.03)	0.11 (0.05)	0.23 (0.05)	0.2 (0.09)
Solar radiations, in MJ/m^2	0.66 (0.04)	0.62 (0.13)	0.66 (0.03)	0.62 (0.13)
Reference evapotranspiration, in mm/day	2.010 (0.34)	1.96 (0.84)	1.36 (0.25)	1.45 (0.85)
Vapor pressure deficit, kPa	0.75 (0.18)	0.78 (0.56)	0.63 (0.15)	0.68 (0.57)

Notes: Yield was detrended to remove the increasing trends of soybean yield time series due to improved cultivars and technological progress.

[1] Not cumulated.