ENVIRONMENTAL RESEARCH
LETTERS

PURPOSE-LED PUBLISHING™ Part of the ENVIRONMENTAL RESEARCH SERIES

LETTER • OPEN ACCESS

Comparison of methods to aggregate climate data to predict crop yield: an application to soybean

To cite this article: Mathilde Chen *et al* 2024 *Environ. Res. Lett.* **19** 054049

IOP Publishing Environ. Res. Lett. 19 (2024) 054049 https://doi.org/10.1088/1748-9326/ad42b5

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
7 December 2023

REVISED
16 April 2024

ACCEPTED FOR PUBLICATION
24 April 2024

PUBLISHED
3 May 2024

LETTER

Comparison of methods to aggregate climate data to predict crop yield: an application to soybean

Mathilde Chen[1,2,3,*], **Nicolas Guilpart**[4] and **David Makowski**[1]

[1] Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA PS, 91120 Palaiseau, France
[2] CIRAD, UMR PHIM, F-34398 Montpellier, France
[3] PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France
[4] Université Paris-Saclay, AgroParisTech, INRAE, UMR Agronomie, 91120 Palaiseau, France
[*] Author to whom any correspondence should be addressed.

E-mail: mathilde.chen@cirad.fr

Keywords: agriculture, soybean yield prediction, climatic predictors, dimensions reduction, model comparison

Supplementary material for this article is available online

Abstract

High-dimensional climate data collected on a daily, monthly, or seasonal time step are now commonly used to predict crop yields worldwide with standard statistical models or machine learning models. Since the use of all available individual climate variables generally leads to calculation problems, over-fitting, and over-parameterization, it is necessary to aggregate the climate data used as predictors. However, there is no consensus on the best way to perform this task, and little is known about the impacts of the type of aggregation method used and of the temporal resolution of weather data on model performances. Based on historical data from 1981 to 2016 of soybean yield and climate on 3447 sites worldwide, this study compares different temporal resolutions (daily, monthly, or seasonal) and dimension reduction techniques (principal component analysis (PCA), partial least square regression, and their functional counterparts) to aggregate climate data used as inputs of machine learning and linear regression (LR) models predicting yields. Results showed that random forest models outperformed and were less sensitive to climate aggregation methods than LRs when predicting soybean yields. With our models, the use of daily climate data did not improve predictive performance compared to monthly data. Models based on PCA or averages of monthly data showed better predictive performance compared to those relying on more sophisticated dimension reduction techniques. By highlighting the high sensitivity of projected impact of climate on crop yields to the temporal resolution and aggregation of climate input data, this study reveals that model performances can be improved by choosing the most appropriate time resolution and aggregation techniques. Practical recommendations are formulated in this article based on our results.

1. Introduction

Large scale crop yield predictions play a significant role for commodity trading and implementation of food security policies [1]. At national scale, they are essential to prevent food shortages arising from harvest losses or failures, while forecasts spanning continental and global scales are used in projecting the impact of climate change on crops yields [2]. Accurate prediction of yield variations, which can impact the price of commodities traded intensively on international markets, is particularly strategic [3]. Notably, Brazil and the United States (US) emerge as primary producers [4] and exporters [5] of soybean, a crop upon which China [6] and the European Union [7] heavily rely. For these countries, forecasting soybean yields in major producers is crucial for addressing disruption in their respective supply chains [8].

The time resolution of climate variables used for yield prediction presents a challenge: while daily global-scale datasets, such as ERA5-land [9], provide numerous potential predictors (e.g. temperature, precipitation, solar radiation), only one crop yield observation is usually available every year. There is thus a sharp contrast between the number of yield data and the number of potential predictors available. An

additional issue is that climate predictors are often correlated either across dates for a given variable (e.g. temperatures during successive days) or between different types of climate variables at a given date (e.g. rainfall and solar radiation).

In order to reduce the number of predictors of annual crop yields, daily climate data are often aggregated over temporal intervals such as months [8, 10] or seasons [11], or condensed into indices based on *a priori* knowledge of soybean physiology [12]. While this approach effectively diminishes the number of climate predictors, facilitating parameter estimation and mitigating overfitting risks, it is somewhat arbitrary in determining the optimal temporal resolution for deriving average climate variables. An alternative approach is to transform the high-dimensional climate data into a low-dimensional representation which retains some meaningful properties of the original (high-dimensional) data. For instance, principal component analysis (PCA) produces linear combinations of original predictors, summarizing data without significant loss of information [13]. Another technique for dimensionality reduction is partial least square regression (PLSR), which replaces both the initial predictors and predicted outcome by a reduced number of latent variables added iteratively [14].

Climate data are commonly regarded as finite, discrete, and independent observations, thereby overlooking the temporal structures inherent in climatic patterns which can also influence crop yield [15]. Compared to usual PCA and PLSR, techniques based on functional data analysis such as functional PCA (FPCA) [16], multivariate FPCA (MFPCA) [17], and functional PLSR (FPLSR) [18] explicitly incorporate the temporal sequencing of observations. By assuming that observed discrete time series arise from smooth functions of time, these functional techniques can be seen as continuous counterparts of PCA and PLSR [19].

The influence of daily climate time series aggregation method on crop yield predictions was explored in several studies [10, 20, 21] which focus on a limited number of dimension reduction techniques, consider specific temporal resolutions of climate data, and do not compare functional to more traditional approaches to take into account the temporal structure of climate data to predict crop yield. To date, these approaches were only assessed in France [22] or in the US [23] at the region or county level. In addition, none of these studies was conducted at a larger scale, thereby constraining our comprehension of the impending challenges confronting agricultural production. This limitation is particularly critical as production zones are anticipated to undergo shifts in response to climate change [24].

To fill this gap, we simultaneously evaluate (i) the impact of the temporal resolution of climate data, (ii) a large range of dimension reduction techniques to aggregate climate data, and (iii) modeling techniques to predict soybean yields from climate data. These analyses were first conducted at the global scale and then at a national scale (i.e. separately in the US and in Brazil), in order to examine the robustness of our conclusions. Using data of historical soybean yields [25] and climate [9] on 3447 sites worldwide from 1981 to 2016, we identified the best data-driven approach to predict soybean yields from climate inputs and to examine the impact of climate data aggregation method on predictive performances.

2. Material and methods

2.1. Data

2.1.1. Soybean yield

Using the Global Dataset of Historical Yields [25], grid-wise data covering the 1981–2016 period were derived for locations representative of global soybean production (Argentina, Brazil, Canada, China, India, Italy, and US). A set of grid-cells located in these major soybean producers and with substantial soybean area was constituted. To avoid any confusion with technological progress due to improved cultivars and technological progress, yield data were detrended. See further details on data sources and pre-processing steps in supplementary material 1. In soybean producing sites, detrended yield ranged from 0.1 to 6.7 $t.ha^{-1}$. Mean (standard deviation) yield was 2.6 (1.1) $t.ha^{-1}$.

Previous work showed that increasing the range of environmental conditions in the training dataset improves predictive models accuracy [26]. Following the procedure employed in previous work [8], several grid-cells located in areas characterized by climate deemed inappropriate for soybean cultivation and resulting to zero yields (such as deserts and arctic areas) were included. The selection process was designed to ensure a balanced distribution of sites across climate zones, with the objective of including 20% of zero yield values in the global yield dataset. Consequently, the final dataset covered 3447 sites, including 663 located in unsuitable climate conditions for soybean production (figure 1).

2.1.2. Irrigation fraction

Using the SPAM dataset [27], the proportion of soybean cultivation under irrigation (i.e. fractional area) was retrieved for each grid-cell. A fractional area of 0 indicates that 100% of soybean grown within the considered grid-cell is rainfed. Rainfed soybean was exclusively grown in 856 sites (i.e. fractional area of irrigated soybean in 2010 is equal to 0% in these sites). Within sites with fractional area of irrigated soybean exceeding 0%, irrigated soybean production covered 54.4% of soybean production in average.

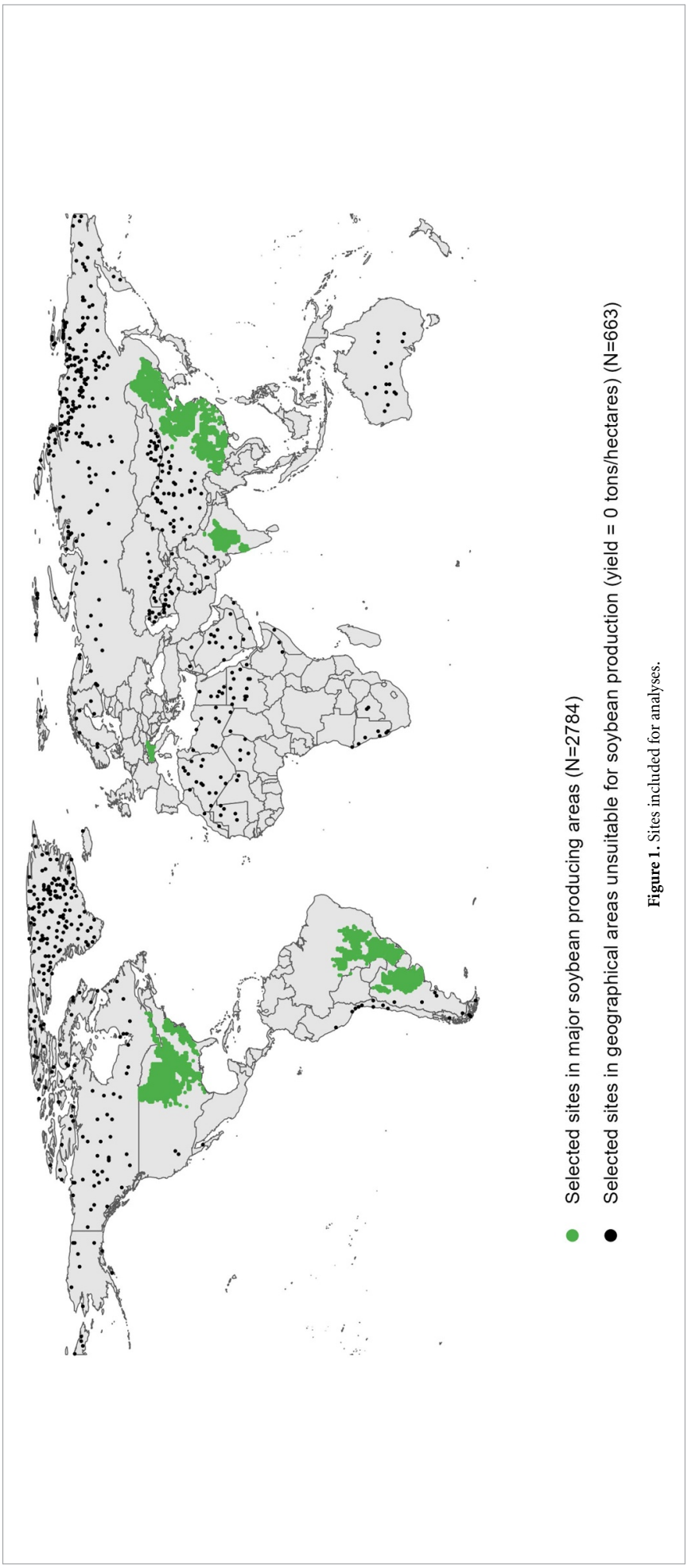

Figure 1. Sites included for analyses.

Table 1. Mean (standard deviation) and range [minimum–maximum] of seasonal climate data over the 1981–2016 period.

	Soybean producing areas (98 361 sites-year)	Full dataset (122 229 sites-year)
Seasonal climate data		
Maximum temperature, °C	25.37 (3.64) [14.26–34.85]	22.16 (10.15) [−20.81–42.42]
Minimum temperature, °C	15.69 (4.21) [3.34–25.19]	12.68 (9.21) [−27.60–28.41]
Cumulated average precipitation, mm	35.84 (14.55) [4.09–123.35]	30.84 (17.09) [0.00–155.01]
Cumulated net surface solar radiation, MJ m^{-2}	135.20 (12.33) [87.79–180.83]	126.64 (28.99) [21.20–213.59]
Reference evapotranspiration, $mm.day^{-1}$	1.73 (0.44) [0.69–5.07]	1.73 (0.86) [0.19–8.17]
Vapor pressure deficit, kPA	0.74 (0.22) [0.24–2.34]	0.75 (0.53) [0.02–5.04]

Notes: Sites located in soybean producing areas are displayed as green dots in figure 1. Seasonal refers to the mean over soybean growing season.

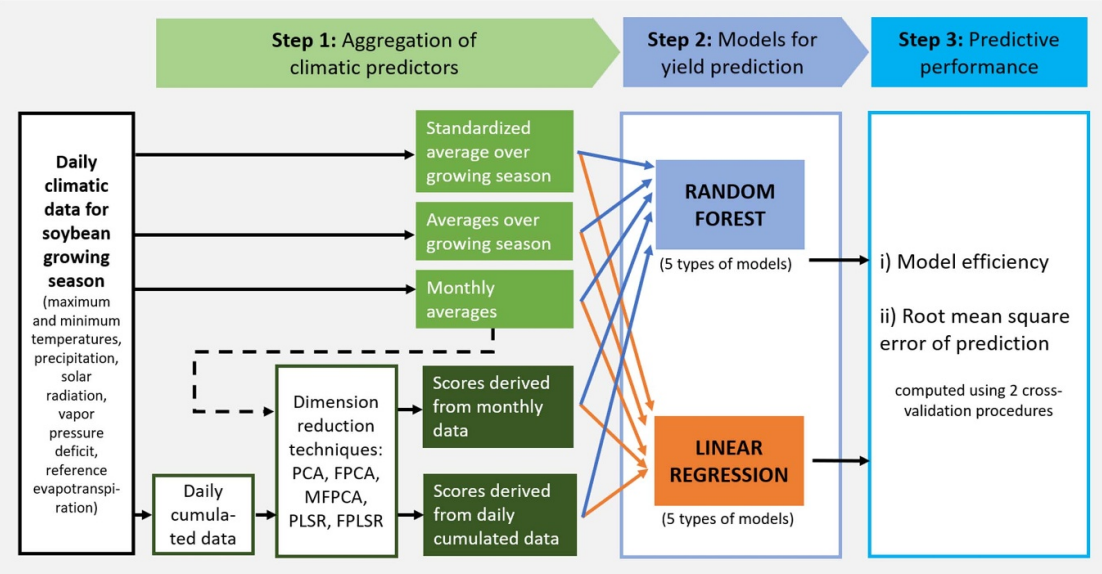

Figure 2. Modeling framework implemented in this study. Abbreviations: PCA: principal component analysis; FPCA: functional principal component analysis; MFPCA: multivariate functional principal component analysis; PLSR: partial least square regression; FPLSR: functional partial least square regression.

2.1.3. Climate data

Climate data aggregated at daily frequency from 1981 to 2016 and realigned with the yield dataset was derived from ERA5-land dataset, a product of the European Centre for Medium-Range Weather Forecasts atmospheric reanalysis of the global climate [9]. Six variables were computed for each grid-cell: maximum and minimum temperatures (both in °C), net surface solar radiation ($MJ.m^{-2}$), average precipitation (mm), reference evapotranspiration ($mm.d^{-1}$), and vapor pressure deficit (kPa). Detailed information regarding on climate variables computation is provided in supplementary table 1. Data were computed for each day of soybean growing season, delineated on a country-specific basis according to crop calendars sourced from the Agricultural Market Information System [28]. Given that the number of days in the growing season varied from 1272 to 1284 depending on the country (i.e. Brazil/Argentina vs. other countries) and the type of year (i.e. bissextile vs non-bissextile years) and considering that six climate variables were considered in this study, each site-year was characterized by numerous climate predictors (ranging from 7632 to 7704).

The range of climatic conditions covered by the dataset is showed in table 1.

2.2. Statistical analyses

Analyses were conducted in three steps, presented in figure 2 and in the following paragraphs: 1) aggregation of climate predictors; 2) prediction of soybean yield based on predictors obtained from the

Table 2. Climate data aggregation methods considered in this study.

	Temporal resolution of climate data			
	Standardized seasonal	Seasonal	Monthly	Cumulative daily
Dimension reduction technique				
No dimension reduction	avg.zscore.s	avg.s	avg.m	
Principal component analysis (PCA)			pca.m	pca.d
Functional PCA			fpca.m	fpca.d
Multivariate functional PCA			mfpca.m	mfpca.d
Partial least square regression (PLSR)			plsr.m	plsr.d
Functional PLSR			fplsr.m	fplsr.d

previous step; 3) assessment of models' predictive performances.

2.2.1. Aggregation of daily climate predictors

Several temporal resolutions were considered. Cumulative daily values over the growing season were computed for each climate variable, as done in previous work [22]. Additionally, non-cumulative daily climate data were averaged, either on a monthly basis or over the entire soybean growing season. Finally, an alternative approach involved rescaling climatic data, so that each climate variable had a mean of zero and a standard deviation of one, and to use the mean of all rescaled data. The four temporal resolutions are hereafter denoted as 'daily', 'monthly', 'seasonal', and 'standardized seasonal', respectively.

Five different dimension-reduction techniques were then applied to cumulative daily data and monthly averages: PCA, FPCA, MFPCA, PLSR, and FPLSR (see supplementary material for methodological and computational details). The different combinations of temporal resolutions and dimension reduction techniques examined in this study are shown in table 2.

The cumulative variance explained by the components obtained through each dimension reduction technique is presented in figure 3. Across PCA, FPCA, MFPCA, or PLSR, the two first components generally accounted for 90% of the variability both in daily and monthly climate data, excluding monthly precipitations. Data transformation by FPLSR was less effective than other techniques of dimension reduction (figure 3).

2.2.2. Models for soybean yield prediction

In this study, two modeling approaches were considered for soybean yield prediction: linear regression (LR) and random forest (RF). LR assumes that the relationship between climate predictors and soybean yield is linear (supplementary material 3). On the contrary, RF [29] is a tree-based machine-learning method that makes no assumption regarding the distribution and relationship between predictors and yield. Briefly, RF algorithm consists in building an ensemble of independent decision trees from bootstrapped samples. Individual trees have the properties to have low bias but high variance, and when combined together, produce an output with lower variance. RF was chosen because this machine learning algorithm showed good performance compared to other algorithms in predicting yield of crops [30], especially soybean [8].

For each approach, predictive models based on aggregation methods presented in table 2 were fitted. For each dimension-reduction method (i.e. PCA, FPCA, MFPCA, PLSR, and FPLSR) applied on daily or monthly climate data, models including the scores associated with one, two, three, or all components as predictors were considered. In total 43 RF models and 43 LR models were compared.

2.2.3. Evaluation of predictive performances

Nash–Sutcliffe model efficiency (NSE, unitless) and root mean square error (RMSE, in $t.ha^{-1}$) were used as measures of predictive performances, as commonly used for agricultural systems and crop models [8, 31, 32]. An efficiency of one corresponds to a perfect match of predictions to observed data, an efficiency of zero indicates that predictions are as accurate as the mean of observed data, whereas an efficiency lower than zero occurs when the observed mean is a better predictor than the tested model. The lower the RMSE, the lower the difference between predictions and observations, which corresponds to a better performance of the model.

Previous articles emphasized the importance of rigorous cross-validation strategies to ensure that the predictive performance of a given model is evaluated on a dataset independent from the one used to train that algorithm [30, 33]. NSE and RMSE were computed for each model following two separate cross-validation procedures. First, a year-by-year cross-validation was performed, to assess model's capability in predicting yields in a new year, not included in the training dataset (temporal extrapolation). Secondly, a group-wise cross-validation was employed, wherein 10 randomly selected site groups were used to evaluate the model's ability to forecast yields in novel geographic regions not encompassed within the training dataset (spatial extrapolation). A visual representation of these cross-validation procedures is provided in supplementary figure 1.

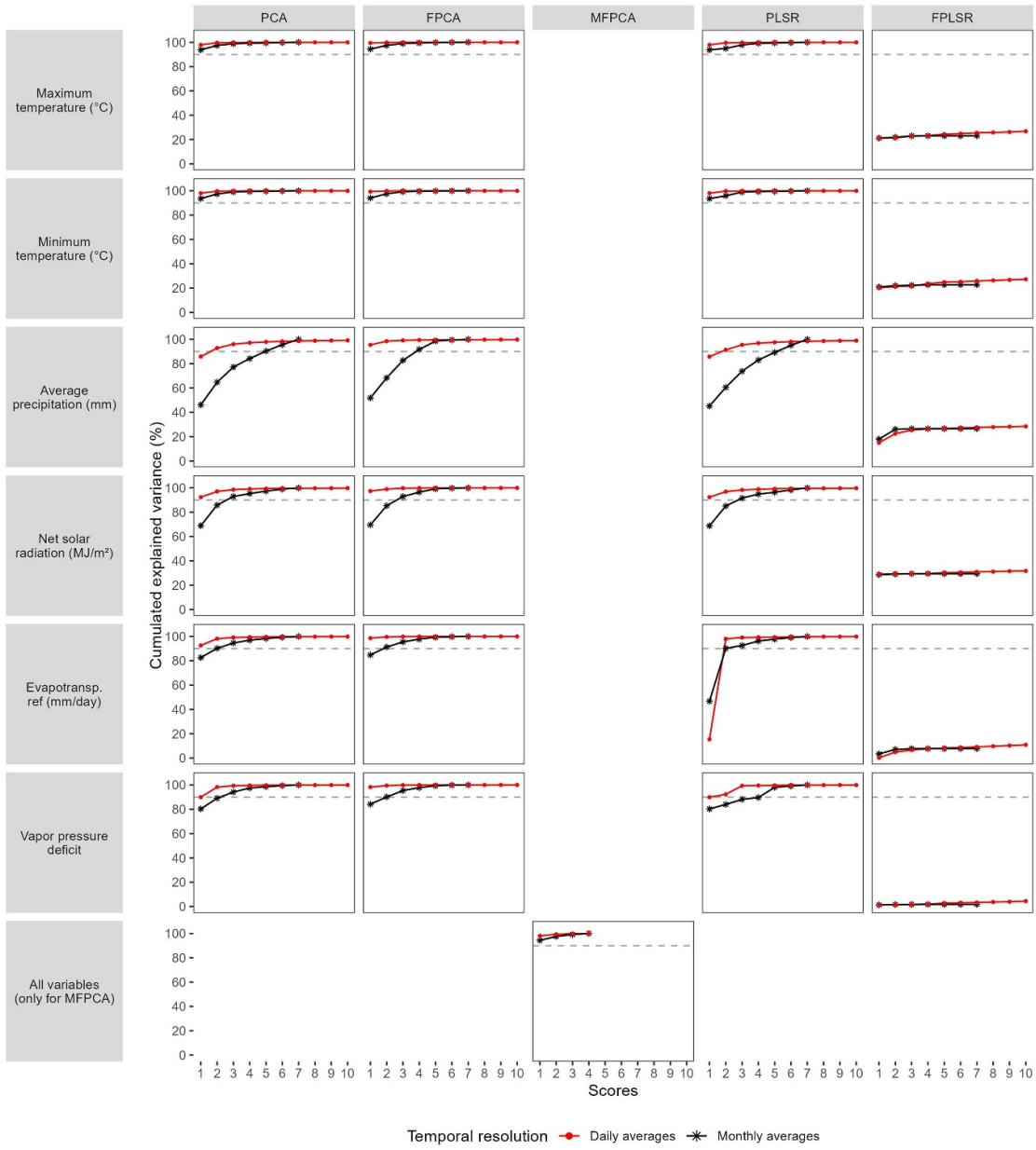

Figure 3. Cumulative variance of climate data explained by the different dimension reduction techniques. Representation limited to the first 10 components. Abbreviations: PCA: principal component analysis; FPCA: functional principal component analysis; MFPCA: multivariate functional principal component analysis; PLSR: partial least square regression; FPLSR: functional partial least square regression. Horizontal dotted line corresponds to 90% of explained variance.

The variations in predictive performance according to models' characteristics were examined using a separate LR model, which related NSE to model family, temporal resolution of climate predictors, dimension reduction techniques (all included as categorical variables), and number of predictors. This separate model was fitted using results from both cross-validation procedures (on years and on sites). Thus, to estimate the independent effect of all other factor on NSE, the model additionally included cross-validation procedure as confounding factor.

2.2.4. Interpretation of the best model

To examine the compatibility of the optimal data-driven model with existing knowledge regarding the influence of climate on soybean physiology, the importance of predictors and their partial dependency profiles were computed. The importance of a predictor in a model is measured by the increases of prediction error resulting from a random permutation of its values [34]. Partial dependence plot summarizes the effect of a particular predictor on the predicted outcome, showing how the predicted value of a model behaves as a function of a given predictor [35].

2.3. Sensitivity analyses

Sensitivity analyses were conducted to assess (i) the robustness of the ranking of the different approaches to the geographical region considered and (ii) the sensitivity of yield projections under global warming based on the employed modeling techniques. The full procedure presented in figure 2 was repeated separately at the scale of the US and Brazil. Distinct sets of grid-cells located areas with unsuitable climate conditions were selected, ensuring that the total number of zero-yield values constituted 20% of the respective yield datasets in each country (see procedure detailed previously and in supplementary material 1). Models' ranking was compared to the one obtained in the analysis conducted at the global scale. The model that performed the best in each analysis was identified as the 'reference model'. Subsequently, four variants of this model, varying in temporal resolution of climate data, dimension reduction techniques (functional or otherwise), or both aspects, were selected for further investigation (supplementary material 4). These models were used to project soybean yields from climate inputs re-computed in simulated climate scenarios where daily temperatures from 1981 to 2016 were incrementally increased by +1, 2, 3, or 4 °C while keeping other climate parameters unchanged. For each site, the relative difference between the median of predictions in each climate scenario and the median of estimations under historical climate conditions (i.e. climate spanning 1981–2016) was computed. The resulting predictions were also compared to those from one RF and one LR model based on monthly averages, an approach which has been used in previous studies to predict soybean yield [8].

All relevant scripts and documentation are available via the project repository (https://github.com/MathildeChen/SOYBEAN_PRED_COMP).

3. Results

3.1. Predictive performance of soybean yield forecasting models

Figure 4 shows the performance of predictive models according to model family, climate data aggregation technique, and temporal resolution of climate data. NSE values of each individual model are displayed in figure 5, along with mean NSE values over the two cross-validation procedures.

Overall, predictive performance was systematically higher for RF models (mean [standard deviation (SD)] NSE: 0.82 [0.14]) compared to LR models (mean [SD] NSE: 0.44 [0.17]) irrespective of temporal resolution of climate data and dimension reduction technique (figure 4(a)). Models incorporating climate data as averages, or as scores derived from PCA, FPCA, or PLSR showed higher NSE compared to those relying on standardized means, MFPCA, or FPLSR scores (figure 4(b)). Generally, models based on monthly climate data performed better compared to those based on daily cumulative or seasonal climate data (figure 4(c)), but the reverse was observed among models based on MFPCA (figures 5(a) and (b)). FPCA and MFPCA methods showed equivalent or lower performance compared to PCA (figures 4(b) and 5). Similarly, FPLSR exhibited lower and highly variable predictive accuracy compared to PLSR (figures 4(b) and 5). Among RF models employing the same dimension reduction technique (i.e. averages, PCA, FPCA, MFPCA, PLS, or FPLS applied on daily or monthly climate data), most parsimonious RF models demonstrated better performances in the year-by-year cross-validation (figure 5(a)), while the contrary was observed in cross-validation across groups of sites (figure 5(b)). The statistical analysis of NSE differences according to models' characteristics confirmed most of the trends observed in figure 4, with the exception of the number of predictors which showed not significant effect on NSE (p-value = 0.252, supplementary table 2).

3.2. Characteristics of the best predictive model

At the global scale, the *pca.m.2, pca.m.3*, and *avg.m* RF models showed equivalent predictive performance (mean NSE: 0.92 for the three models; figure 5(c)). Ranking remained consistent when considering RMSE as predictive performance metric (mean RMSE ranging from 0.37 to 0.38 for the three models; supplementary figure 2). The *pca.m.2* model, demonstrating the highest parsimony among these alternatives, was selected as the 'best' predictive model. Predictions strongly correlated with observed yields with both year-by-year and site-by-site cross-validation procedures (Pearson correlation coefficients between observed yields and model predictions in year-by-year [ρ_{year}] and site-by-site [ρ_{site}] cross-validation procedures: 0.957 and 0.967, respectively; supplementary figure 3). In contrast, *pca.m.2* LR model showed lower and sometimes suboptimal performance in predicting soybean yields (ρ_{year} and ρ_{site}: 0.753 and 0.752, respectively). Notably, in sites where no production had been recorded, frequent negative yield values were predicted by this model, which falls outside the realistic yield range. Similar results were obtained for the *pca.d.2* LR model. Predictions from the *pca.d.2* RF model were more accurate (ρ_{year} and ρ_{site}: 0.933 and 0.942, respectively) compared to LR model with similar temporal resolution, but did not reach the precision achieved by the *pca.m.2* RF model.

Although both models tended to underestimate higher yields while overestimating lower values (figures 6(a) and (b), supplementary figure 3), this phenomenon was more pronounced for the *pca.d.2* model. Finally, the residuals of the *pca.m.2* RF model were symmetrically distributed (figure 6(c)).

Predictors showing highest importance in the *pca.m.2* model were the scores associated with the first

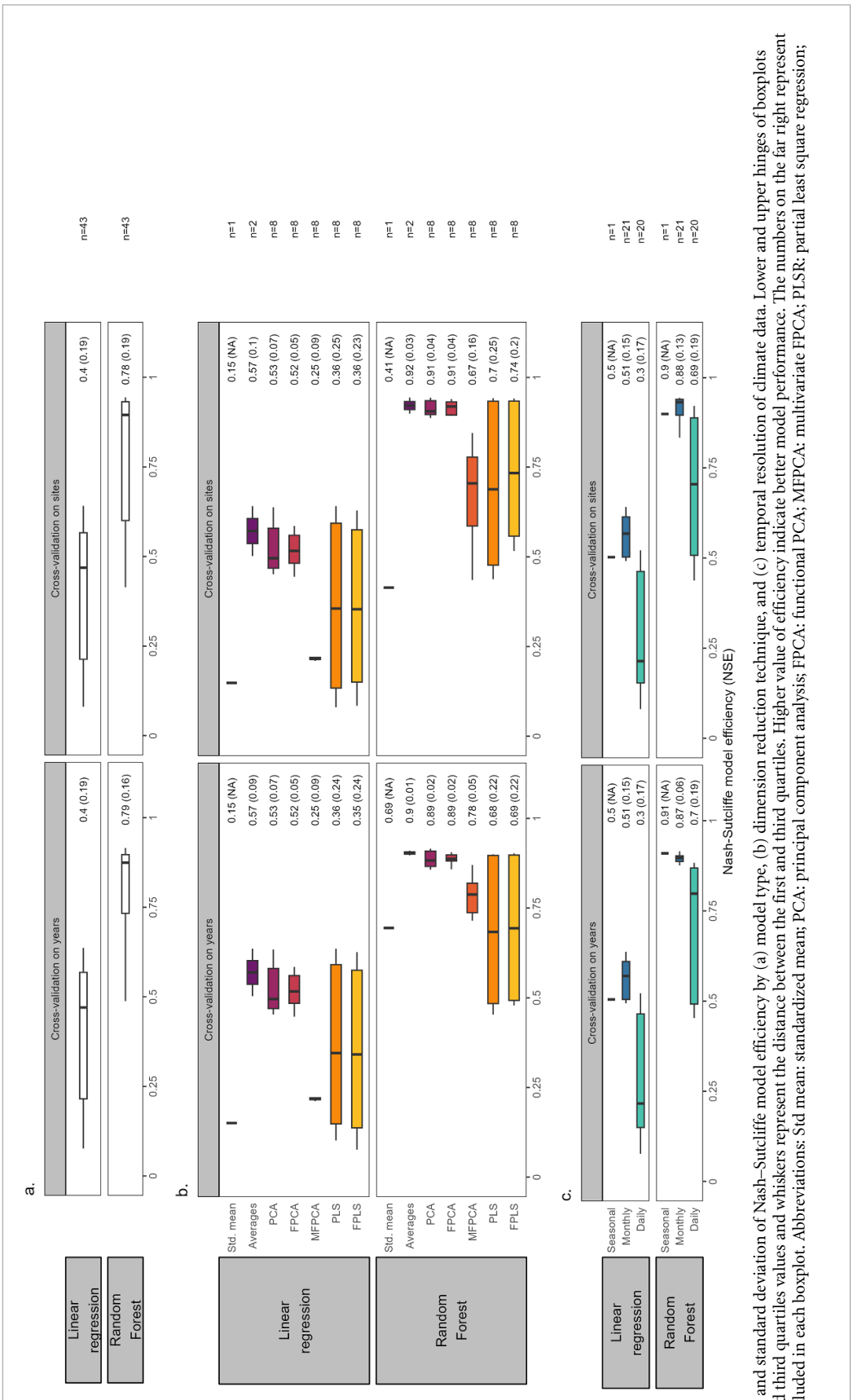

Figure 4. Boxplots, mean, and standard deviation of Nash–Sutcliffe model efficiency by (a) model type, (b) dimension reduction technique, and (c) temporal resolution of climate data. Lower and upper hinges of boxplots correspond to the first and third quartiles values and whiskers represent the distance between the first and third quartiles. Higher value of efficiency indicate better model performance. The numbers on the far right represent the number of models included in each boxplot. Abbreviations: Std mean: standardized mean; PCA: principal component analysis; FPCA: functional PCA; MFPCA: multivariate FPCA; PLSR: partial least square regression; FPLSR: functional PLSR.

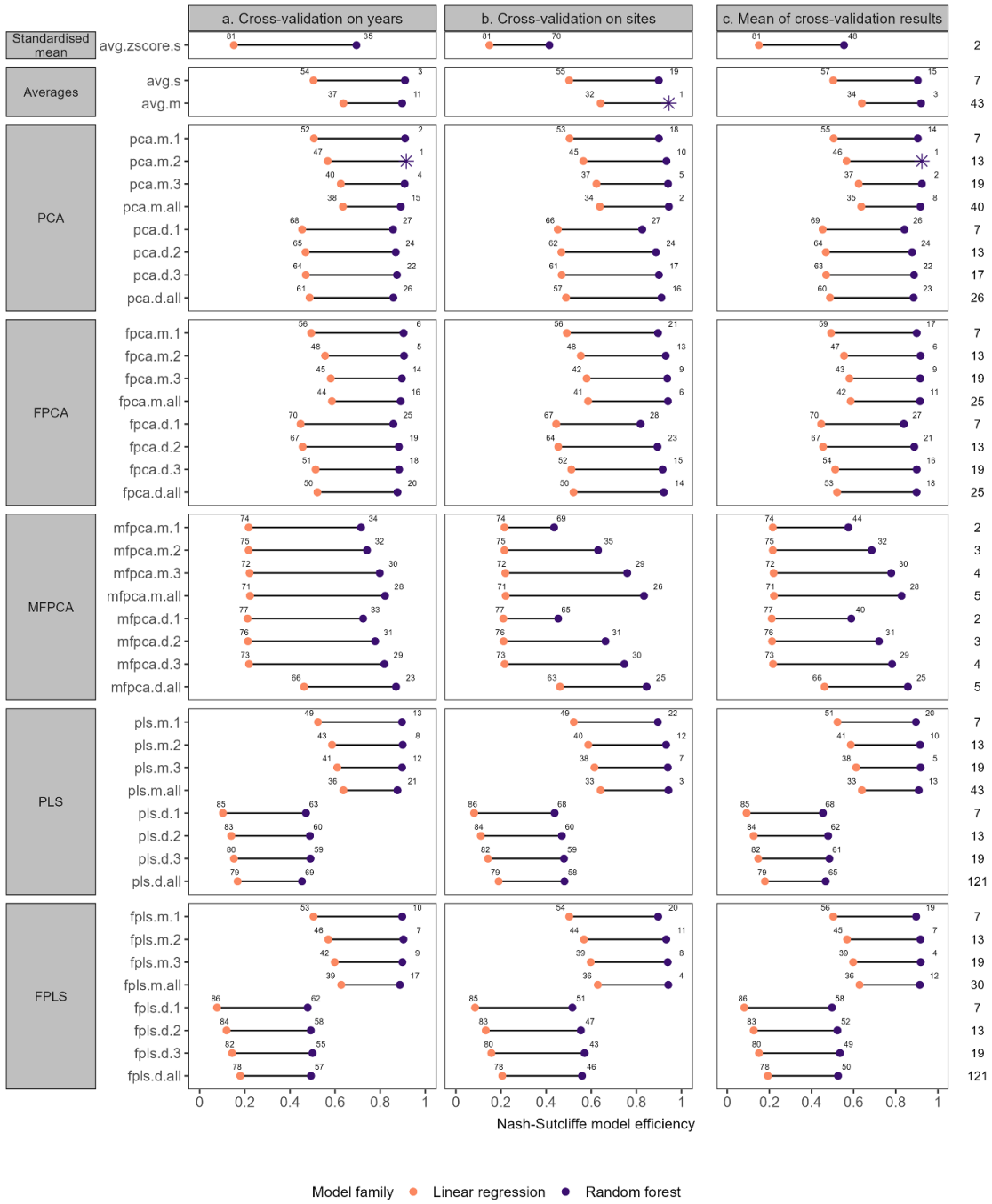

Figure 5. Models' performance estimated by cross-validation (a) on years or (b) on sites, and (c) averaged. Higher value of Nash-Sutcliffe efficiency indicate better performance. For each column, model's ranking is indicated above corresponding dots and the best model is highlighted by a *. Number of predictors is indicated on the right of the figure. See table 1 for models' name abbreviations.

principal components (hereafter referred to as 'score 1') derived from monthly precipitation, minimum temperature, and maximum temperature (supplementary figure 4). In terms of importance, irrigated soybean area ranked after all scores 1 but before all scores associated with the second principal components (hereafter referred to as 'scores 2').

Figure 7(a) shows the correlations between monthly climate averages and scores of the first principal component (score 1) for each type of climate variable (precipitation, minimum temperature, maximum temperature etc.) included in the *pca.m.2* model (i.e. the model with highest predictive accuracy in your study). Results show that score 1 is strongly positively correlated with all monthly climate variables but net solar radiation. A positive (negative) correlation indicates that an increase in the corresponding monthly climate variable would increase (decrease) the value associated to the first principal component (score 1). For example, an increase of precipitation on month 4 would increase the value of score 1 but, on the contrary, an increase of net solar

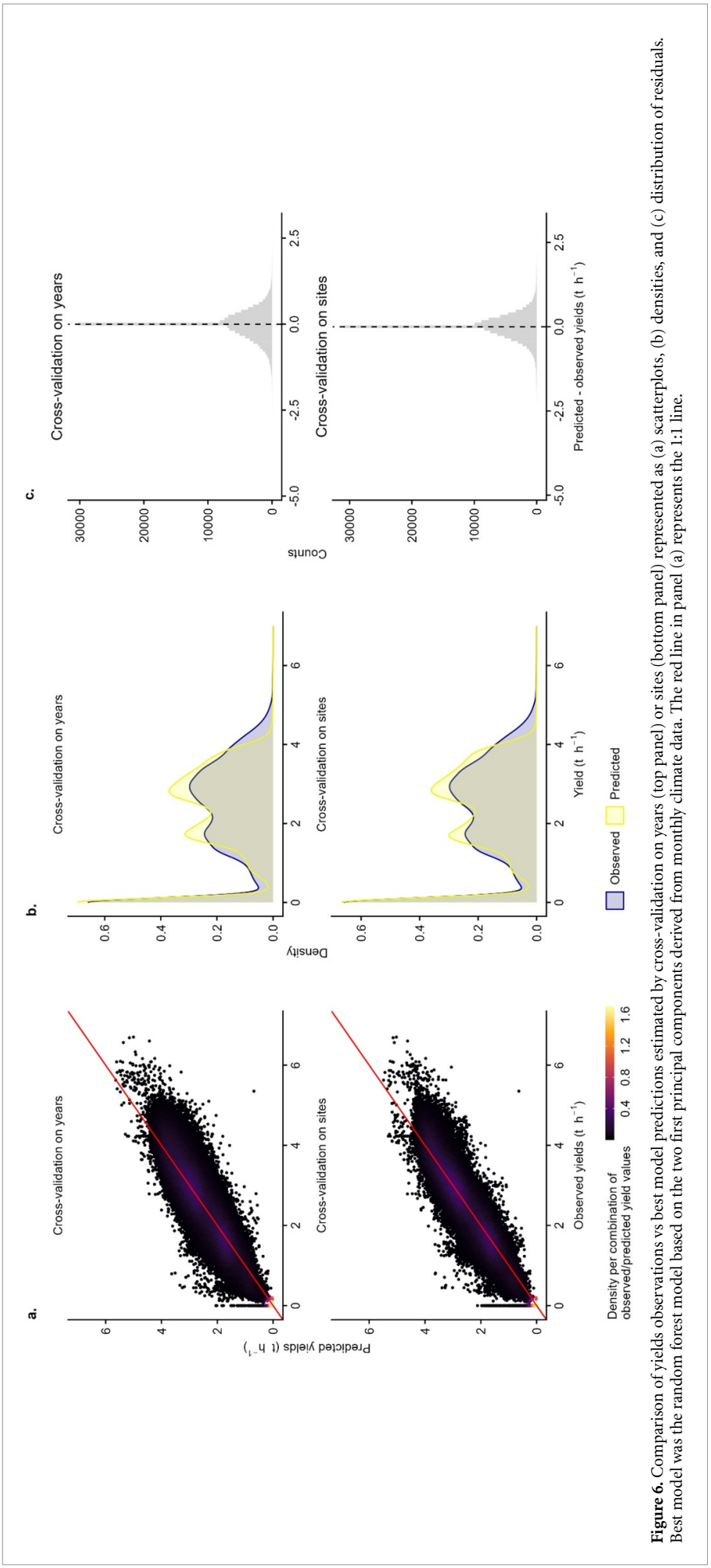

Figure 6. Comparison of yields observations vs best model predictions estimated by cross-validation on years (top panel) or sites (bottom panel) represented as (a) scatterplots, (b) densities, and (c) distribution of residuals. Best model was the random forest model based on the two first principal components derived from monthly climate data. The red line in panel (a) represents the 1:1 line.

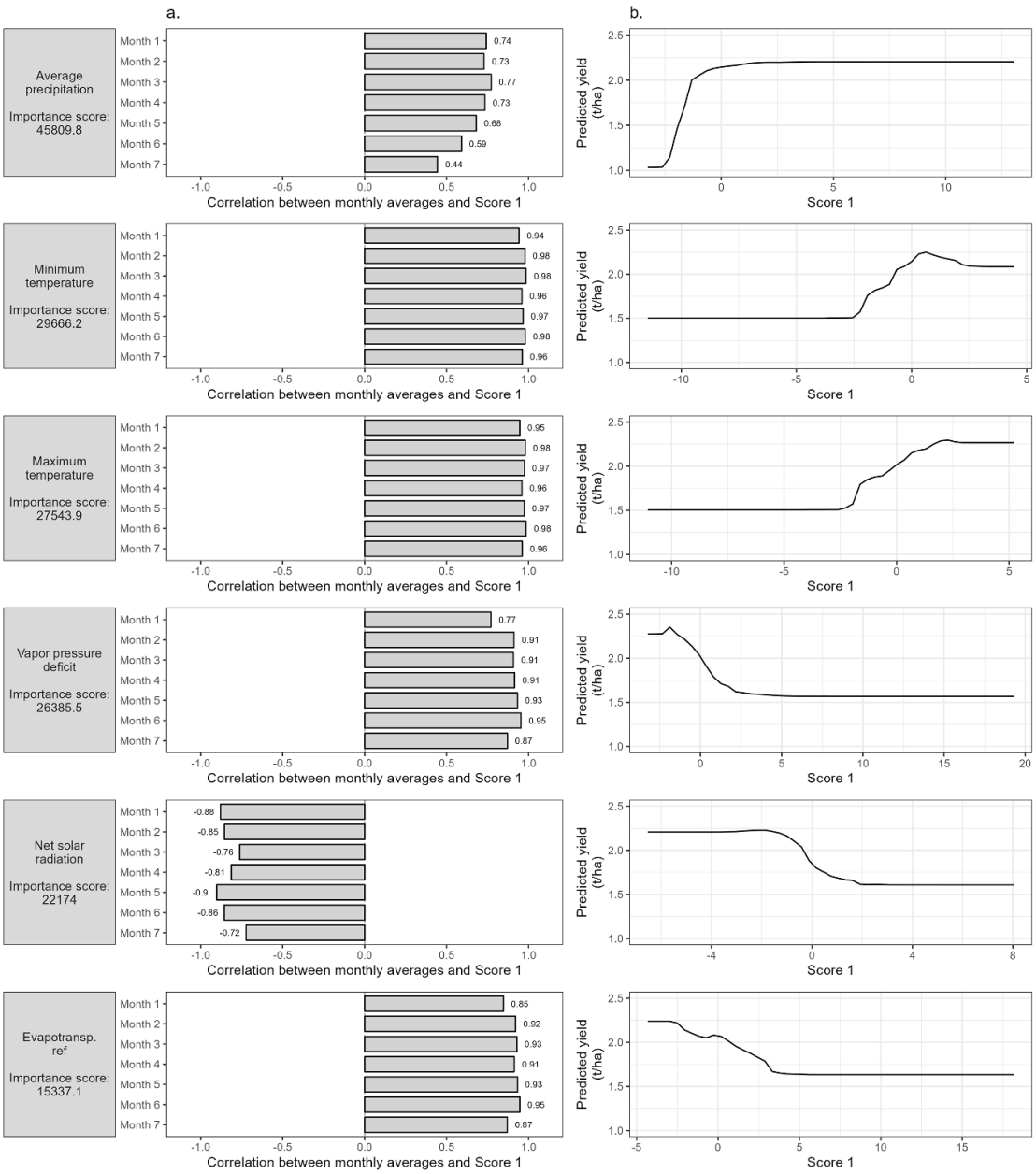

Figure 7. Contribution of climate data to the most important predictors in the best model (a) and partial dependency plot associated with each predictor (b). Best model was the random forest model including the scores associated with the first and second components derived from principal component analysis applied on monthly data. Panel a: for each climate variable, height, direction of the bars, and the number at the top of the bars indicate Pearson correlation coefficient between monthly averages and the score associated with the first principal component (i.e. 'score 1'). Panel b: the expected value of model yield prediction is plotted as a function of each predictor.

radiation will decrease score 1 because the correlations are negative in all months for this variable.

According to partial dependency plots presented in the figure 7(b), predicted yield increased with higher values of scores 1 for temperatures and precipitation; this suggests that warmer and wetter environments are more favorable to soybean yield, considering the positive correlations between these scores and monthly climate averages. Regarding solar radiation, reference evapotranspiration, and vapor pressure deficit, higher score values are associated with lower soybean yields. Negative correlations between the scores 1 and monthly radiation averages suggest that soybean yield would benefit of higher radiation. Conversely, lower reference evapotranspiration and vapor pressure deficit would be detrimental for soybean production. Note that irrigation is associated with higher values of yield soybean compared to absence of irrigation (supplementary figure 6).

By comparison with scores 1, scores 2 correlated less importantly with climate through the season (absolute value of Pearson coefficient ranging from 0 to 0.65), and correlation values varied between months. The interpretation of the effects of climate variables is more complex for scores 2 because the sign of the correlations

depends on the month, as shown in supplementary figure 6(a). However, as all scores 2 had a much lower importance in the model, their impact on model prediction is very small (supplementary figure 7(b)).

3.3. Sensitivity analyses

Sensitivity analyses conducted at US and Brazil scales included 1004 and 527 sites (supplementary figure 8), corresponding to 37 147 and 18 429 sites-year, respectively. Higher yields were observed in the US compared to Brazil. Higher temperatures, precipitations, and amounts of solar radiation were observed in Brazil (supplementary table 5). For both countries, PCA, FPCA, MFPCA, and PLSR were able to capture most of the variability in climate data, with lower performance for monthly precipitation (supplementary figures 9 and 10).

Similar to analysis at the global scale, RF models based on scores derived from month-based PCA performed better on average at national scale. The *pca.m.3* and *pca.m.2* models ranked first and second-best models to predict soybean yield in the US (mean NSE values for both models: 0.94; supplementary figure 11(c)) and third and first best models in Brazil (NSE values: 0.9315 and 0.9305 respectively, supplementary figure 12(c)). Similar ranking was obtained when using RMSE as an indicator of predictive performance (supplementary figures 13 and 14). The RF model based on monthly averages (*avg.m*) performed poorly in both US and Brazil analyses (ranked at the 8th and 12th position, respectively; supplementary figures 13(c) and 14(c)), compared to the results obtained at the global scale (figure 5(c) and supplementary figure 2(c)).

The scores associated with component derived from month-based PCA had similar interpretation as in global analysis, i.e. score 1 being strongly correlated to the overall trend in climate data and subsequent scores highlighting specific months of the growing season (supplementary figures 15 and 16). The scores 1 of precipitation, minimum temperature, and maximum temperature were among the most important climate predictors in the models fitted on US and Brazil (supplementary figure 17). Other important climate predictors were reference evapotranspiration in the US and vapor pressure deficit in Brazil. In the US model, the fractional area of irrigated soybean had the highest importance, followed by all first scores derived from month-based PCA, while in Brazil it was less important.

3.3.1. Impact of climate data aggregation method on model predictions under global warming

In both US and Brazil, the RF *pca.m.3* and *pca.m.2* models showed higher performance, respectively (supplementary figures 11–14). The model variants considered to assess the sensitivity of yield projections according to the chosen modeling techniques are presented in supplementary material 4. In addition, the models based on monthly averages of climate data (*avg.m*) was included because this approach is widely applied in many predictive models of crop yield. For each model variant and each climate scenario, the difference between the median of predicted yields under increased temperature and the median of yields predicted in 1981–2016 climate (+0 °C) was computed for each site located in the US (figure 8) and in Brazil (figure 9).

Major differences in yield predictions were obtained between the different types of model and temporal resolution of climate data. While month-based LR predictions generally suggest that soybean yields could uniformly increase over the US or over Brazil with temperature increase, month-based RF identified areas in both countries where decreases in soybean yield would be expected, especially in the case of extreme temperature increase (+4 °C). Areas with higher yield losses would be mainly located in the South-West coast and in the South. Contradictory projections were obtained for the North-Center of the US. Projected yields increased according to the month-based RF models, while the reverse was predicted by the daily-based RF models (figure 8). For Brazil, areas with increased or decreased yields were consistent when considering projections of RF models, although daily-based models tended to attenuate the detrimental effect of temperature increase on yield. Contrasted conclusions were obtained by LR month-based and daily-based models (figure 9).

4. Discussion

This study examining the impact of climate data aggregation methods for predicting soybean yield presents four key findings. First, RF models outperformed LR in predictability and was less sensitive to the temporal resolution and aggregation method. Second, there was no evidence that using daily climate data combined with a dimension reduction technique improved predictive performances compared to using monthly climate predictors. Three, more sophisticated methods based on functional data analyses (i.e. FPCA, MFPCA, or FPLSR) to aggregate climate data did not improve models' predictive performances compared to simpler aggregation techniques (i.e. mean, PCA, or PLSR). Finally, the RF model with climate predictors derived from PCA applied on monthly climate data showed superior predictive performances at the global scale as well as at the national scale.

Climate variables such as temperature and precipitation are commonly used in crop yield forecasting [36], yet evidence about the impact of climate data aggregation methods on predictive performances

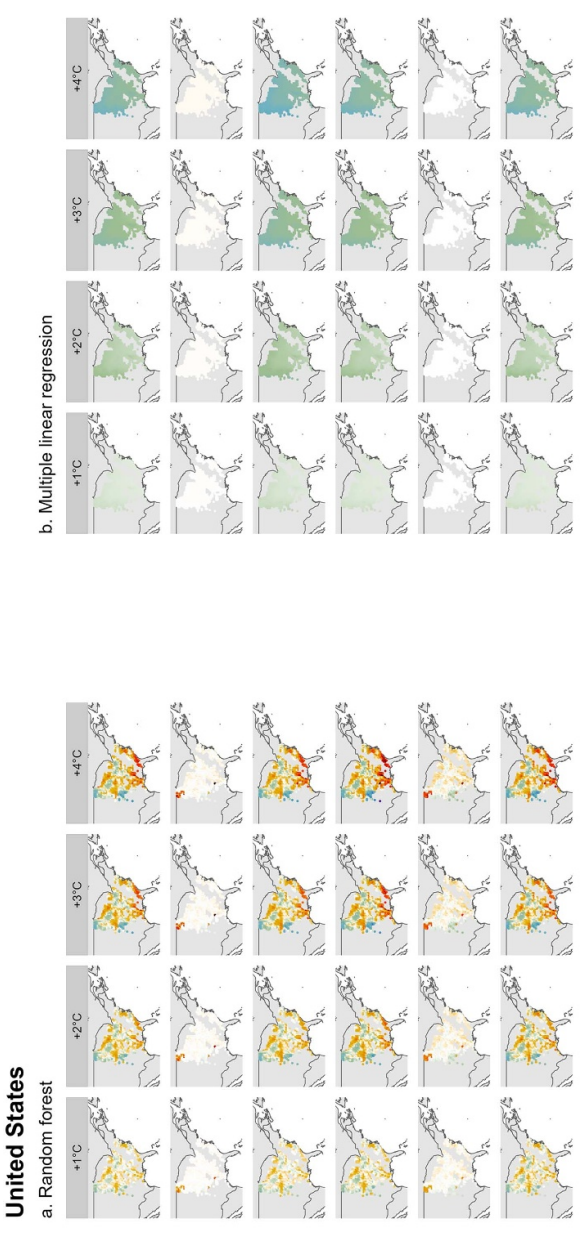

Figure 8. Relative difference (%) in soybean yield projections under different scenarios of temperature increase compared to historical climate in the US. Medians were computed by site, over 1981–2016 period, for each scenario of temperature increase compared to historical (i.e. 1981–2016) climate in this region. Abbreviations: *pca.m.3*: models using three scores derived from month-based principal component analysis; *pca.d.3*: models using three scores derived from daily-based principal component analysis; *fpca.m.3*: models using three scores derived from month-based functional principal component analysis; *pls.m.2*: models using two scores derived from month-based partial least-square regression; *pls.d.3*: models using three scores derived from daily-based partial least-square regression; *avg.m*: model based on monthly averages of climate data.

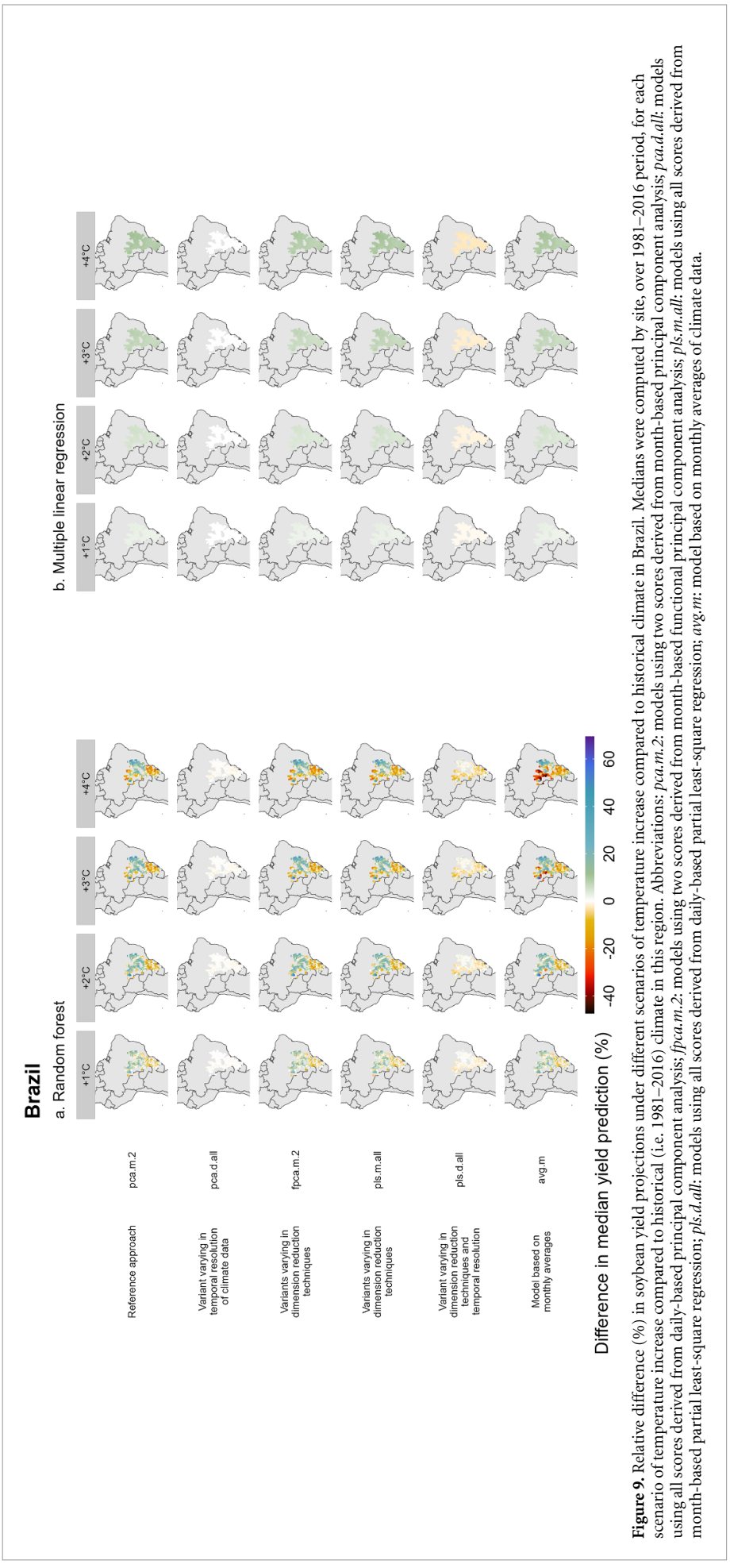

Figure 9. Relative difference (%) in soybean yield projections under different scenarios of temperature increase compared to historical climate in Brazil. Medians were computed by site, over 1981–2016 period, for each scenario of temperature increase compared to historical (i.e. 1981–2016) climate in this region. Abbreviations: *pca.m.2*: models using two scores derived from month-based principal component analysis; *pca.d.all*: models using all scores derived from daily-based principal component analysis; *fpca.m.2*: models using two scores derived from month-based functional principal component analysis; *pls.m.all*: models using all scores derived from month-based partial least-square regression; *pls.d.all*: models using all scores derived from daily-based partial least-square regression; *avg.m*: model based on monthly averages of climate data.

comes from few studies [10, 20–23] focusing on limited dimension reduction techniques and/or temporal resolution. The study extends previous literature by comparing a wide range of aggregation techniques to efficiently incorporate climate data in yield forecast models. The good performances obtained with aggregated climate data align with prior findings indicating equivalent [20] or lower [10] prediction performance of models with finer time resolution compared to their monthly counterparts. Although applying PCA to monthly climate data slightly improved predictive performance, concerns about reduced interpretability of climate variable effects arise. However, PCA-based models can be easily interpreted, for two reasons. First, the low number of selected principal components (here, two) in the yield forecasting models ensures parsimony. Moreover, the orthogonality of principal components, in contrast to the often highly correlated monthly averages, enhances interpretability (supplementary figure 5). Secondly, standard graphical representations can be used to interpret the effects of climate variables on yields when the forecasting models rely on principal components. Specifically, correlations between the original climate variables and the principal components can be conveniently displayed graphically, while the relationship between scores of principal components and yields can be visualized using partial dependence plots. By combining these two types of graphic, it is possible to analyze the effects of the climate variables on yield.

Thanks to their greater flexibility, machine learning algorithms such as RF [29], (extreme) gradient boosting [37, 38], and deep learning [39] are widely employed to predict yields of crops from climatic predictors [36, 40] and often outperform traditional methods such as LR or process-based models, particularly in soybean yields predictions [8, 41, 42]. Among these techniques, RF was found to be one of the best algorithm in soybean yield prediction [8, 41, 42] as well as in other major crops including wheat [43] or maize [44]. Previous articles report that other algorithms which have not been considered in our study such as neural networks [12] also perform well to forecast yield of soybean. Guilpart and colleagues [8], who used similar data as in the present study, demonstrated the best prediction performances of RF in soybean prediction over other machine learning and deep learning techniques (i.e. neural networks, generalized additive models, and gradient boosting). Additionally, differences between deep learning and machine learning techniques were proved to be negligible in crop yield prediction [45].

In this study, RF models showed highest predictive performances compared to LR models. These differences can be explained by the inherent inability of LR to extrapolate to data dimensions where no training has been done, which is not the case of RF models. In other words, RF models would not make predictions outside the range of the data used to train the model, while LR can predict for the conditions beyond this range. This prevents overfitting issues, but can lead to unrealistic estimations (supplementary figure 3). In addition, situations where multiple predictors correlated with (and within) each other and drive the response similarly affect LR, which is not the case of the RF.

Several types of data exist to develop crop yield predictive models, such as survey data on yields, which are available in many countries for long time periods. However, these data are generally aggregated at large administrative scales, such counties or countries. Experimental or on-farm data are also a valuable source of information [46], because they are more precisely located than the gridded-cell data and provide information regarding farm management However, these data cover restricted spatial area and time period, and are thus not suitable for modeling yields at the continental or global scale. The reanalysis data used in this study combines the advantages to span large scale in space and time while covering a large diversity of yield and climate conditions, as shown in table 1. The climate data used in the present study were able to explain a large share of the spatial variability of yields, as showed by the good predictive performance of our models. However, such datasets can suffer from uncertainties [47, 48]. Examining the impact of these uncertainties on the predictive accuracy of models deserves a full investigation on itself and is beyond the scope of this study, which aims to compare different modeling approaches and techniques for aggregating climate data to predict crop yield.

Sensitivity analyses were conducted to assess the impact of increasing temperature scenarios on yields. The results align with studies simulating soybean yield under climate change in the US [49] and in some regions of Brazil [50]. Although highly unrealistic because temperature will not be the only climate feature that would change in a context of global change, this analysis contributed to evaluate the compared models and approach, as done in previous studies [51]. To precisely project the impact of climate change on soybean productivity in these countries, our findings need to be consolidated using more elaborated climate change scenarios [8].

5. Conclusion

This study simultaneously evaluates the impact of temporal resolutions and aggregations of climate predictors on the performance of machine learning and LR models predicting crop yields. Key results indicate that (i) RF outperformed and was less sensitive to climate aggregation than LR; (ii) there was no evidence that using daily climate data improves predictive performance over using monthly data in our models;

(iii) employing PCA on monthly data, coupled with a RF algorithm, yields the most accurate predictions for crop yields. These results highlight the significance of temporal resolution and climate data aggregation in projecting climate impacts on crop yields. Thus, careful consideration of optimal time resolution and aggregation techniques is imperative when developing models for crop yield prediction, particularly for future climate projections.

Data availability statements

Datasets are fully available online:

- Soybean 0.5° grid-wise yield data covering the 1981-2016 period: https://doi.pangaea.de/10.1594/PANGAEA.909132
- Proportion of soybean irrigated area in each grid-cell was retrieved from the SPAM2010 v2.0 dataset: http://mapspam.info/
- Climate variables at a resolution of 0.1° covering the period from January 1950 to present: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview.
- Crop calendars provided by the Agricultural Market Information System: www.amis-outlook.org/amis-about/calendars/soybeancal/en/
- No new data were created or analysed in this study.

Acknowledgments

This work was supported by the ANR under the *Investissements d'avenir* program with the reference ANR-16-CONV-0003 (CLAND).

Authors contribution

Mathilde CHEN: Conceptualization, Investigation, Methodology, Formal analysis, Visualization. Writing—original manuscript. **Nicolas GUILPART**: Conceptualization, Supervision, Writing—review & editing. **David MAKOWSKI:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing—review & editing.

Conflict of interest

We declare no competing interests.

Code availability statement

All relevant R (version 4.2.2 www.r-project.org) scripts and documentation will be available online after acceptance via the project repository (https://github.com/MathildeChen/SOYBEAN_PRED_COMP) or under the request of the corresponding author. The python script used to run and set parameters of the CDS API will also be provided in the dedicated repository.

ORCID iDs

Mathilde Chen https://orcid.org/0000-0002-5982-2143
Nicolas Guilpart https://orcid.org/0000-0003-3804-0211
David Makowski https://orcid.org/0000-0001-6385-3703

References

[1] Duveiller G, Kayitakire F, Hoefsloot P, Hansen J, Ines A and Dam J 2012 Combining crop models and remote sensing for yield prediction—concepts, applications and challenges for heterogeneous, smallholder environment *Report of CCFAS-JRC Workshop at Joint Research Centre, Ispra, Italy* (Publications Office) (Accessed 13–14 June 2012)
[2] Silva J V and Giller K E 2020 Grand challenges for the 21st century: what crop models can and can't (yet) do *J. Agric. Sci.* **158** 794–805
[3] Zelingher R and Makowski D 2022 Forecasting global maize prices from regional productions *Front. Sustain. Food Syst.* **6** 836437
[4] FAO 2022 *World Food and Agriculture—Statistical Yearbook 2022* (FAO)
[5] FAOSTATS 2023 *Trade/Crops and Livestock Products* FAOSTATS (available at: www.fao.org/faostat/fr/#data/TCL/metadata)
[6] Gale F, Valdes C and Mark A 2019 Interdependence of China, United States, and Brazil in soybean trade *USDA, Economic Research Service, Service ER; 2019* Contract No: OCS-19F-01 (https://doi.org/10.2196/14279)
[7] 2019 United States is Europe's main soya beans supplier with imports up by 112% [press release] (European Commission) (available at: https://ec.europa.eu/commission/presscorner/detail/en/IP_19_161) (Accessed 7 January 2019)
[8] Guilpart N, Iizumi T and Makowski D 2022 Data-driven projections suggest large opportunities to improve Europe's soybean self-sufficiency under climate change *Nat. Food* **3** 255–65
[9] Muñoz-Sabater J *et al* 2021 ERA5-Land: a state-of-the-art global reanalysis dataset for land applications *Earth Syst. Sci. Data* **13** 4349–83
[10] Sharif B, Makowski D, Plauborg F and Olesen J E 2017 Comparison of regression techniques to predict response of oilseed rape yield to variation in climatic conditions in Denmark *Eur. J. Agron.* **82** 11–20
[11] Zhu P, Burney J, Chang J, Jin Z, Mueller N D, Xin Q, Xu J, Yu L, Makowski D and Ciais P 2022 Warming reduces global agricultural production by decreasing cropping frequency and yields *Nat. Clim. Change* **12** 1016–23
[12] von Bloh M, RdS N J, Wangerpohl X, Saltık A O, Haller V, Kaiser L and Asseng S 2023 Machine learning for soybean yield forecasting in Brazil *Agric. For. Meteorol.* **341** 109670
[13] Jolliffe I 2011 Principal component analysis *International Encyclopedia of Statistical Science* ed M Lovric (Springer Berlin Heidelberg) pp 1094–6
[14] Wold S, Sjöström M and Eriksson L 2001 PLS-regression: a basic tool of chemometrics *Chemometr. Intell. Lab. Syst.* **58** 109–30
[15] Yu Q, Li L, Luo Q, Eamus D, Xu S, Chen C, Wang E, Liu J and Nielsen D C 2014 Year patterns of climate impact on wheat yields *Int. J. Climatol.* **34** 518–28
[16] Ramsay J O and Silverman B W 2005 *Functional Data Analysis* (Springer)
[17] Happ C and Greven S 2018 Multivariate functional principal component analysis for data observed on different (Dimensional) domains *J. Am. Stat. Assoc.* **113** 649–59
[18] Krämer N, Boulesteix A-L and Tutz G 2008 Penalized partial least squares with applications to B-spline transformations and functional data *Chemometr. Intell. Lab. Syst.* **94** 60–69

[19] Ullah S and Finch C F 2013 Applications of functional data analysis: a systematic review *BMC Med. Res. Methodol.* **13** 43
[20] Kang Y, Ozdogan M, Zhu X, Ye Z, Hain C and Anderson M 2020 Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US midwest *Environ. Res. Lett.* **15** 064005
[21] Pham H T, Awange J, Kuhn M, Nguyen B V and Bui L K 2022 Enhancing crop yield prediction utilizing machine learning on satellite-based vegetation health indices *Sensors* **22** 719
[22] Bonneu F, Makowski D, Joly J and Allard D 2022 Machine learning based on functional principal component analysis to identify major influential factors of wheat yield *SSRN Electron. J.* (available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4760642)
[23] Sartore L, Rosales A N, Johnson D M and Spiegelman C H 2022 Assessing machine leaning algorithms on crop yield forecasts using functional covariates derived from remotely sensed data *Comput. Electron. Agric.* **194** 106704
[24] King M, Altdorff D, Li P, Galagedara L, Holden J and Unc A 2018 Northward shift of the agricultural climate zone under 21st-century global climate change *Sci. Rep.* **8** 7904
[25] Iizumi T and Sakai T 2020 The global dataset of historical yields for major crops 1981–2016 *Sci. Data* **7** 97
[26] Dupin M, Reynaud P, Jarošík V, Baker R, Brunel S, Eyre D, Pergl J and Makowski D 2011 Effects of the training dataset characteristics on the performance of nine species distribution models: application to diabrotica virgifera virgifera *PLoS One* **6** e20957
[27] Yu Q, You L, Wood-Sichra U, Ru Y, Joglekar A K B, Fritz S, Xiong W, Lu M, Wu W and Yang P 2020 A cultivated planet in 2010—Part 2: the global gridded agricultural-production maps *Earth Syst. Sci. Data* **12** 3545–72
[28] AMIS 2022 SOYBEANS: planting and harvesting calendar (available at: www.amis-outlook.org/amis-about/calendars/soybeancal/en/)
[29] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
[30] Morales A and Villalobos F J 2023 Using machine learning for crop yield prediction in the past or the future *Front. Plant Sci.* **14** 1128388
[31] Kim S-H, Yang Y, Timlin D J, Fleisher D H, Dathe A, Reddy V R and Staver K 2012 Modeling temperature responses of leaf growth, development, and biomass in maize with MAIZSIM *Agron. J.* **104** 1523–37
[32] Wallach D, Makowski D, Jones J W and Brun F 2018 Working with dynamic crop models *Methods, Tools and Examples for Agriculture and Environment* 3rd edn (Elsevier) p 613
[33] Roberts D R *et al* 2017 Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure *Ecography* **40** 913–29
[34] Genuer R and Poggi J-M 2020 Variable importance *Random Forests with R* ed R Genuer and J-M Poggi (Springer International Publishing) pp 57–76
[35] Biecek P and Burzykowski T 2021 *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models* (CRC Press)
[36] van Klompenburg T, Kassahun A and Catal C 2020 Crop yield prediction using machine learning: a systematic literature review *Comput. Electron. Agric.* **177** 105709
[37] Chen T and Guestrin C ed 2016 Xgboost: a scalable tree boosting system *Proc. 22nd Acm Sigkdd Int. Conf. on Knowledge Discovery and Data Mining*
[38] Friedman J H 2002 Stochastic gradient boosting *Comput. Stat. Data Anal.* **38** 367–78
[39] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *nature* **521** 436–44
[40] Bali N and Singla A 2022 Emerging trends in machine learning to predict crop yield and study its influential factors: a survey *Arch. Comput. Methods Eng.* **29** 95–112
[41] Barbosa Dos Santos V, Moreno Ferreira Dos Santos A, da Silva Cabral de Moraes J R, de Oliveira Vieira I C and de Souza Rolim G 2022 Machine learning algorithms for soybean yield forecasting in the Brazilian Cerrado *J. Sci. Food Agric.* **102** 3665–72
[42] Kaul M, Hill R L and Walthall C 2005 Artificial neural networks for corn and soybean yield prediction *Agric. Syst.* **85** 1–18
[43] Jeong J H *et al* 2016 Random forests for global and regional crop yield predictions *PLoS One* **11** e0156571
[44] Leng G and Hall J W 2020 Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models *Environ. Res. Lett.* **15** 044027
[45] Richetti J, Diakogianis F I, Bender A, Colaço A F and Lawes R A 2023 A methods guideline for deep learning for tabular data in agriculture with a case study to forecast cereal yield *Comput. Electron. Agric.* **205** 107642
[46] Silva J V, Heerwaarden J, Reidsma P, Laborte A G, Tesfaye K and Ittersum M 2023 Big data, small explanatory and predictive power: lessons from random forest modeling of on-farm yield variability and implications for data-driven agronomy *Field Crops Res.* **302** 109063
[47] Iizumi T, Kotoku M, Kim W, West P C, Gerber J S and Brown M E 2018 Uncertainties of potentials and recent changes in global yields of major crops resulting from census- and satellite-based yield datasets at multiple resolutions *PLoS One* **13** e0203809
[48] Ruane A C *et al* 2021 Strong regional influence of climatic forcing datasets on global crop model ensembles *Agric. For. Meteorol.* **300** 108313
[49] Petersen L K 2019 Impact of climate change on twenty-first century crop yields in the U.S *Climate* **7** 40
[50] Zilli M, Scarabello M, Soterroni A C, Valin H, Mosnier A, Leclère D, Havlík P, Kraxner F, Lopes M A and Ramos F M 2020 The impact of climate change on Brazil's agriculture *Sci. Total Environ.* **740** 139384
[51] Lobell D B and Burke M B 2010 On the use of statistical models to predict crop yield responses to climate change *Agric. For. Meteorol.* **150** 1443–52