Comparison of Methods To Aggregate Climate Data To Predict Crop Yield: An Application to Soybean

Mathilde Chen
mathilde.chen@inrae.fr

Inrae  https://orcid.org/0000-0002-5982-2143

Nicolas Guilpart
AgroParisTech

David Makowski
Inrae

Research Article

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

Posted Date: January 26th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3897405/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Version of Record: A version of this preprint was published at Environmental Research Letters on April 24th, 2024. See the published version at https://doi.org/10.1088/1748-9326/ad42b5.
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

KEYWORDS: Agriculture, soybean yield prediction, climatic predictors, dimensions reduction, model comparison
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 3,447 sites worldwide, this study compares different
temporal resolutions (daily, monthly, or seasonal) and dimension reduction techniques (principal
component analysis, partial least square regression, and their functional counterparts) to
aggregate climate data used as inputs of machine learning and linear regression models
predicting yields. Results showed that random forest models outperformed and were less
sensitive to climate aggregation methods than linear regressions 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 principal component analysis 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.
INTRODUCTION

Crop yield predictions at large scales play a significant role for commodity trading and implementation of food security policies [1]. Predictions of yields at a national level are essential to prevent food shortages in case of harvest losses or failures, while continental and global yield forecasts are frequently used to make projections about the impact of climate change on crops [2]. Accurate yield prediction is particularly strategic for commodities traded intensively on international markets [3]. For these commodities, variations in yield can have a major impact on price, and it is therefore essential to anticipate them [4]. The case of soybean is particularly interesting because Brazil and the United States (US) produce the majority of global production, accounting for 69% of the total production in 2018 [5] and more than 85% of global exports in 2021 [6]. Several countries depend heavily on soybean imports such as China (which accounted for ca. 60% of global soybean imports in 2017 [7]) and the European Union [8], for which forecasting soybean yields in Brazil and the US is crucial for anticipating disruption in their supply chains. Understanding whether these countries will be able to meet their soybean demand in future climatic conditions is of utmost importance and could help them to identify new areas where soybean production could be relocated [9].

Historically, crop yield predictions were obtained with process-based models which intend to integrate biophysical mechanisms underlying plant growth and development [10]. However, these models are not always reliable due to major issues related to the estimation of their numerous parameters [11]. Previous research showed that these models may provide inaccurate forecasts [12, 13] which can lead to contradictory conclusions depending on the type of model used.[14] In parallel, statistical linear regression models are simpler and less costly to implement compared to process-based models [10]. For these reasons, they were frequently used in crop yield prediction as well,[15, 16] but sometimes showed poor predictive performance [15]. Standard regression models do not perform well when their inputs are highly correlated and do not always capture all the possible interactions between predictors [15]. Because of their greater flexibility, machine learning algorithms such as random forest (RF) [17], (extreme) gradient boosting [18, 19], and deep learning [20] are now commonly used to predict yields of crops from climatic predictors [10, 21] and often show better performances compared to traditional statistical methods such as linear regression, in particular to predict soybean yields [9, 22, 23].
The use of statistical and machine learning methods raises a specific problem concerning the
time step of the climate variables used as yield predictors. Several climate datasets provide
gridded daily standard variables such as temperature, precipitation, or solar radiation at the
global scale, like the ERA5-land dataset [24]. However, while only one crop yield observation is
usually available every year, several hundred of daily climate inputs can potentially be used as
predictors for yield forecast. 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 usually averaged over months [9, 25] or seasonally [26] or summarized as index relying on a
priori knowledge on soybean physiology [27]. This approach considerably reduces the number of
climate predictors, making parameters estimation tractable and reducing the risk of overfitting.
However, these approaches are somewhat arbitrary as there is no obvious optimal rule to decide
which temporal resolution should be considered to derive 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 example, principal component analysis (PCA) produces linear combinations of original
predictors that summarize the data without losing too much information [28]. Another dimension
reduction technique is partial least square regression (PLSR), which replaces both the initial
predictors and predicted set of variables by a reduced number of latent variables which are
included one by one in an iterative process [29].

Generally, climate data are treated as finite discrete and independent observations, ignoring the
temporal pattern underpinning climate data which can also influence crop yield [30]. Compared
to usual PCA and PLSR, techniques based on functional data analysis such as functional PCA
(FPCA) [31], multivariate FPCA (MFPCA) [32], and functional PLSR (FPLSR) [33] take into
account the order of the observations in time. By assuming that observed discrete time series
arise from smooth functions of time, these functional techniques can be seen as continuous
versions of PCA and PLSR [34].

4
Previous studies [25, 35, 36] examining the impact of the method of aggregation of daily climate
time series on crop yield predictions 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 [37] or in the US [38]
at the region or county level. In addition, none of these studies was conducted at a larger scale,
which limit our understanding of future challenges that agricultural production will face, since
production zones are expected to shift under climate change [39]. 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) modelling
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 [40, 41] and
climate [24] on 3,447 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.

MATERIAL AND METHODS

Data

Soybean yield

Soybean yield data were taken from the global dataset of historical yields, which provides
worldwide 0.5° (~55 km²) grid-wise data covering the 1981-2016 period (updated 1.2-1.3
version, available at: https://doi.pangaea.de/10.1594/PANGAEA.909132) [40, 41]. Yield values
reported in this dataset result from the combination of several sources of information, including
national scale yield statistics from the Food and Agriculture Organization, global crop calendars
and harvested areas, and satellite-derived net primary production values. A set of locations
representative of global soybean production was constituted with grid-cells located in major
soybean producers (Argentina, Brazil, Canada, China, India, Italy, and US) with substantial
soybean area (i.e. harvested area of soybean of at least 1% based on the M3 [42] crop mask). To
avoid any confusion with technological progress due to improved cultivars and technological
progress, yield data were detrended following the procedure presented in Supplementary Material 1.

Previous work showed that increasing the range of environmental conditions represented in the training dataset improves the performance of predictive models [43]. Following the procedure of previous work [9], several grid-cells located in areas with climate preventing soybean cultivation and leading to zero yields (such as deserts and arctic areas) were also included. These grid-cells were randomly drawn from six zones known to be environmentally improper for any crop production based on the last version of the Köppen-Geiger climate classification (available at: http://koeppen-geiger.vu-wien.ac.at/present.htm). The number of selected grid-cells was such that the total amount of zero yield values represented 20% of the global yield dataset. The final dataset covered 3,447 sites, among which 663 were located in improper climate conditions for soybean production (Figure 1).
Figure 1. Sites included for analyses.

- Selected sites in major soybean producing areas (N=2784)
- Selected sites in geographical areas unsuitable for soybean production (yield = 0 tons/hectares) (N=663)
Irrigation fraction

Using the SPAM2010 dataset [44] (version v3.2, available at http://mapspam.info/), the fractional area of irrigated soybean (i.e., the proportion of irrigated soybean area) was retrieved for each grid-cell. A fractional area of 0 means that 100% of soybean grown in the considered grid-cell was rainfed. SPAM2010 data are available for around the year 2005, at a spatial resolution of ~0.08°. To be consistent with yield data, irrigation data were resampled to the spatial resolution of 0.5° using the aggregate() function of the terra R package (version 1.7-46, https://CRAN.R-project.org/package=terra).

Climate data

Climate predictors were derived from ERA5-land data of the European Centre for Medium-Range Weather Forecasts atmospheric reanalysis of the global climate [24]. This dataset provides hourly and monthly estimates of a large number of climate variables at a resolution of 0.1° covering the period from January 1950 to present. The dataset is hosted by the Climate Data Store (CDS) at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview. Using the application programming interface (API) of CDS (https://cds.climate.copernicus.eu/cdsapp#!/software/app-c3s-daily-era5-statistics?tab=overview), we directly obtained the data aggregated at daily frequency, from 1981 to 2016 and aggregated at a 0.5° scale, to be consistent with the soybean yield dataset. The 10 daily climate variables obtained were minimum and maximum temperatures at 2 meters, minimum and maximum dewpoint temperatures at 2 meters (all in K), average precipitation (in mm), surface net solar radiation (in J.m⁻²), U and V components of wind speed at 10 meters (m.s⁻¹), and surface pressure (in Pa). From these 10 initial variables, six new variables were derived: maximum and minimum temperatures (both in °C), net surface solar radiation (MJ.m⁻²), total precipitation (mm), reference evapotranspiration (mm.day⁻¹), and vapor pressure deficit (kPa). See Supplementary Table 1 for details on climate variable computation. Each variable was computed for each day of soybean growing season, which was defined country-by-country according to the crop calendars provided by the Agricultural Market Information System (available at: http://www.amis-outlook.org/amis-about/calendars/soybeancal/en/) [3]. Soybean growing season ranged from November 1st to May 31st in Argentina and Brazil, from May 1st to November 30th in Canada, from June 1st to December 31st in India, and from April 1st to October 31st in China, US, Italy, and for sites located in unproper climatic conditions for soybean.
production. In total, 1,272 to 1,284 climatic predictors were derived for each combination of site and year (hereafter referred to as “site-year”) depending on the country (i.e., Brazil/Argentina vs. other countries) and the type of year (i.e., bissextile vs non-bissextile years).

The range of yield and climate variables covered by the dataset is showed in Table 1. Rainfed soybean was exclusively grown in 856 sites (i.e. fractional area of irrigated soybean in 2010 is equal to 0% in these sites). In sites where fractional area of irrigated soybean was superior to 0%, irrigated soybean production covered 54.4% of soybean production in average.
Table 1. Mean (standard deviation) and range [minimum - maximum] of soybean yield and seasonal climate data over the 1981-2016 period.

<table>
<thead>
<tr>
<th></th>
<th>Sites located in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>soybean producing&lt;sup&gt;1&lt;/sup&gt; areas only (98,361 sites-year)</td>
</tr>
<tr>
<td>Soybean production</td>
<td></td>
</tr>
<tr>
<td>Yield, t.ha&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>2.55 (1.02) [0.07 - 6.7]</td>
</tr>
<tr>
<td>Seasonal&lt;sup&gt;3&lt;/sup&gt; climate data</td>
<td></td>
</tr>
<tr>
<td>Maximum temperature, °C</td>
<td>25.37 (3.64) [14.26 - 34.85]</td>
</tr>
<tr>
<td>Minimum temperature, °C</td>
<td>15.69 (4.21) [3.34 - 25.19]</td>
</tr>
<tr>
<td>Average precipitation&lt;sup&gt;4&lt;/sup&gt;, mm</td>
<td>0.17 (0.07) [0.02 - 0.57]</td>
</tr>
<tr>
<td>Solar radiation, MJ/m&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.63 (0.06) [0.41 - 0.85]</td>
</tr>
<tr>
<td>Reference evapotranspiration, mm day&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>1.73 (0.44) [0.69 - 5.07]</td>
</tr>
<tr>
<td>Vapor pressure deficit, kPA</td>
<td>0.74 (0.22) [0.24 - 2.34]</td>
</tr>
</tbody>
</table>

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

<sup>1</sup> Sites displayed as green dots in Figure 1.

<sup>2</sup> Sites displayed as black dots in Figure 1.

<sup>3</sup> Mean over soybean growing season.

<sup>4</sup> Not cumulated.
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 previous step; 3) assessment of models’ predictive performances.

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.

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 [37]. Daily (i.e. not cumulative) climate data were averaged by (ii) month or (iii) over the soybean growing season. Finally, a variant approach was to rescale climatic data, so that each climate variable had a mean of zero and a standard deviation of one, and to use (iv) the mean of all rescaled data. The four temporal resolution were hereafter referred to 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 resolution and dimension reduction techniques examined in this study are shown in Table 2.

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

<table>
<thead>
<tr>
<th>Dimension reduction technique</th>
<th>Temporal resolution of climate data</th>
</tr>
</thead>
<tbody>
<tr>
<td>No dimension reduction</td>
<td>standardized seasonal</td>
</tr>
<tr>
<td></td>
<td>seasonal</td>
</tr>
<tr>
<td></td>
<td>monthly</td>
</tr>
<tr>
<td></td>
<td>cumulative daily</td>
</tr>
<tr>
<td>Principal component analysis (PCA)</td>
<td>avg.zscore.s</td>
</tr>
<tr>
<td></td>
<td>avg.s</td>
</tr>
<tr>
<td></td>
<td>avg.m</td>
</tr>
<tr>
<td></td>
<td>pca.m</td>
</tr>
<tr>
<td></td>
<td>pca.d</td>
</tr>
<tr>
<td>Functional PCA</td>
<td>fpcam</td>
</tr>
<tr>
<td></td>
<td>fpcad</td>
</tr>
<tr>
<td>Multivariate functional PCA</td>
<td>mfpca.m</td>
</tr>
<tr>
<td></td>
<td>mfpca.d</td>
</tr>
<tr>
<td>Partial least square regression (PLSR)</td>
<td>plsr.m</td>
</tr>
<tr>
<td></td>
<td>plsr.d</td>
</tr>
<tr>
<td>Functional PLSR</td>
<td>fplsr.m</td>
</tr>
<tr>
<td></td>
<td>fplsr.d</td>
</tr>
</tbody>
</table>

The cumulative variance explained by the components derived from each dimension reduction technique is presented in Figure 3. For PCA, FPCA, MFPCA, or PLSR, the two first components were generally able to capture 90% of the variability in daily and monthly climate data, with the exception of monthly precipitations. Data transformation by FPLSR was less effective than other techniques of dimension reduction (Figure 3).
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.

<table>
<thead>
<tr>
<th>Maximum temperature (°C)</th>
<th>PCA</th>
<th>FPCA</th>
<th>MFPCA</th>
<th>PLSR</th>
<th>FPLSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum temperature (°C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average precipitation (mm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net solar radiation (MJ/m²)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evapotranspiration (mm/day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vapor pressure deficit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All variables (only for MFPCA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cumulative explained variance (%)

Scores

Temporal resolution: - Daily averages, ⭐ Monthly averages
Models for soybean yield prediction

Two modelling 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. On the contrary, RF [17] is a tree-based machine-learning method that makes no assumption regarding the distribution and relationship between predictors and yield. Briefly, RF’s 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 [45], especially soybean [9].

For each approach, predictive models based on aggregation methods presented in Table 1 were fitted. For each dimension-reduction method (i.e. PCA, FPCA, MFPCA, PLSR, and FPLSR) applied on daily climatic data or on monthly averages, models including one, two, three, or all scores as predictors were considered. In total 43 RF models and 43 LR models were compared.

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. These are measures commonly used for agricultural systems and crop models [9, 46, 47]. 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 highlighted 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 [45, 48]. For each model, NSE and RMSE were computed following two separate cross-validation procedures. First, a year-by-year cross-validation was performed: for each year, yield was predicted by a model trained on the full dataset excluding the data of the selected year. This type of cross-validation aims at assessing the model capability in predicting yields in a new year, not included in the training dataset (temporal extrapolation). Secondly, a
group-wise cross-validation based on 10 groups of sites (randomly selected) was performed. Yield time series corresponding to each group were subsequently used as a test dataset. This type of cross-validation aims at assessing the model capability in predicting yields in new areas, not included in the training dataset (spatial extrapolation). See Supplementary Figure 1 for a schematic representation of each cross-validation procedure.

**Interpretation of the best model**

To examine whether the best data-driven model fits with the knowledge we have on climate’s effect 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 [49]. Partial dependence plot is another tool used to summarize the effect of a particular predictor on the predicted variable. It shows how the predicted value of a model behaves as a function of a given predictor [50].

**Sensitivity analyses**

First, the full procedure presented in Figure 2 was repeated separately at the scale of the US and Brazil to assess the robustness of the ranking of the different approaches to the geographical region considered. Based on results of the first sensitivity analysis, we identified the models that performed the best in each country, and selected several variants of them (e.g. same dimension reduction technique but different temporal resolution of climate data). 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 increased by +1°C, 2°C, 3°C or 4°C, successively, while other climate parameters were kept unchanged. For each site, the relative difference between the median of predictions in each climate scenario and the median of estimations in historical climate conditions (i.e. climate from 1981 to 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 [9]. Although not realistic, this second sensitivity analysis was conducted in the purpose of to assess the sensitivity of yield projections under global warming to the chosen modelling techniques.
All relevant R (version 4.2.2) scripts, python script used to run and set parameters of the CDS API, and documentation are available via the project repository (https://github.com/MathildeChen/SOYBEAN_PRED_COMP).

RESULTS

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, as well as 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] efficiency: 0.44 [0.17]) regardless of temporal resolution of climate data and dimension reduction technique (Figure 4a). Models in which climate data were incorporated as monthly or seasonal averages or as scores derived from PCA, FPCA, or PLSR showed higher NSE compared to those based on standardized means, MFPCA, or FPLSR scores (Figure 4b). Generally, models based on monthly averages performed better compared to those based on daily cumulative climate data or seasonal averages (Figure 4c), but the reverse was observed among models based on MFPCA (Figure 5a,b). FPCA and MFPCA methods showed equivalent or lower performance compared to PCA (Figure 4b and Figure 5). Similarly, FPLSR exhibited lower and highly variable predictive accuracy compared to PLSR (Figures 4b and Figure 5). Among RF models based on 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 performed better according to the year-by-year cross-validation (Figure 5a), while the contrary was observed from cross-validation on groups of sites (Figure 5b).
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 number of models included in each boxplot is reported on the right. 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.
Characteristics of the best predictive model

At the global scale, three RF models showed equivalent predictive performance (mean NSE: 0.92 for the three models; Figure 5c): the RF incorporating two or three scores derived from PCA applied on monthly data as well as the one based on monthly averages of climate variables (i.e. pca.m.2, pca.m.3, and avg.m, respectively). 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).

For the pca.m.2 model, the scatter plot of observed and predicted yields, as well as their respective density is presented for each cross-validation strategy in Figure 6. Although the model tended to underestimate higher yields while overestimating lower values (Figures 6a and 6b), model residuals were symmetrically distributed (Figure 6c).
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.
The two first scores for each climate variable, as well as fractional area of irrigated soybean were ranked based on their importance in the \textit{pca.m.2} model (Supplementary Figure 3). Predictors showing highest importance in \textit{pca.m.2} model were the first PCA scores, derived from monthly precipitation, minimum temperature, and maximum temperature. In terms of importance, irrigated soybean area ranked after all scores 1 derived from PCA but before all scores 2. Compared to other scores derived from the PCA, scores 1 explained the greatest variation in climate data (Figure 2) and correlated with climate variables in all months of the growing season (Supplementary Figure 3). Second scores correlated with climate data in some specific months only. For example, scores 2 correlated with solar radiation in the middle (month 2 to 4) and the end (months 6 to 7) of the growing season (Supplementary Figure 4).

According to partial dependency plots presented in Supplementary Figure 5, predicted yield increased with higher values of PCA scores 1 for temperatures and precipitation; this suggests that warmer and wetter environments are more favorable to soybean yield, considering the correlation between these scores and monthly averages of temperatures and precipitation (Supplementary Figure 4). Regarding solar radiation, reference evapotranspiration, and vapor pressure deficit, higher scores values were associated with lower soybean yields. Correlations between the scores 1 and corresponding monthly averages suggest that soybean yield would benefit of higher radiation. Conversely, lower reference evapotranspiration and vapor pressure deficit would be detrimental for soybean production. Irrigation is associated with higher values of soybean compared to absence of irrigation (Supplementary Figure 5).

**Sensitivity analyses**

Sensitivity analyses conducted at US and Brazil scales included 1,004 and 527 sites (Supplementary Figure 6), 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 3). For both countries, PCA, FPCA, MFPCA, and PLSR were able to capture most of the variability in climate data, with lower performance for monthly total precipitation (Supplementary Figures 7 and 8).

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 \textit{pca.m.3} and \textit{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 9c) and third and first best models in Brazil (NSE values: 0.9315 and 0.9305 respectively, Supplementary Figure 10c). Similar ranking was obtained when using RMSE as an indicator of predictive performance (Supplementary Figures 11 and 12). The RF model based on monthly averages \( \text{avg.m} \) performed poorly in both US and Brazil analyses (ranked at the 8\(^{th}\) and 12\(^{th}\) position, respectively; Supplementary Figure 11c and 12c), compared to the results obtained at the global scale (Figure 5c and Supplementary Figure 2c).

The scores derived from month-based PCA had similar interpretation as in global analysis, i.e. the first one being strongly correlated to the overall trend in climate data and subsequent scores highlighting specific months of the growing season (Supplementary Figures 13 and 14). The first scores of monthly precipitations, minimum temperature, and maximum temperature were among the most important climate predictors in the models fitted on US and Brazil. 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 (Supplementary Figure 15).

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

In both US and Brazil, the RF models using two or three scores derived from month-based PCA (i.e. \( \text{pca.m.3} \) and \( \text{pca.m.2} \) for US and Brazil, respectively) showed higher performance (Supplementary Figures 9-12). The considered variants of these models were the RF and LR models using scores derived from daily-based PCA (\( \text{pca.d.3} \) and \( \text{pca.d.all} \)), monthly-based FPCA (\( \text{fpca.m.3} \) and \( \text{fpca.m.2} \)), month-based PLSR (\( \text{pls.m.2} \) and \( \text{pls.m.all} \)), and daily-based PLSR (\( \text{pls.d.3} \) and \( \text{pls.d.all} \)) and. In addition, the models based on monthly averages of climate data (\( \text{avg.m} \)) was included because this approach is widely applied in many yield predictive models. Projections of soybean yields in the US and Brazil under different climate change scenarios (+1, +2, +3, and +4°C) were computed using each of the variants mentioned above. 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 7) and in Brazil (Figure 8).
For both the US and Brazil, major differences in yield predictions were obtained between RF and LR models and between models based on monthly or daily climate data. While month-based LR predictions generally suggest that soybean yields could uniformly increase over US with temperature increase, month-based RF identified areas in the US and Brazil 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 7). 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 8).
Figure 7. 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.3: 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 8. 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.3: 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.
DISCUSSION

This study examining the impact of climate data aggregation methods to predict soybean yield presents four key findings. First, RF models showed higher predictive ability and lower sensitivity to the temporal resolution and aggregation method compared to LR. 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 the best predictive performances at the global scale as well as at the national scale in the US and Brazil.

Although climate variables such as temperature and precipitation are commonly used in crop yield forecasting [21], the impact of climate data aggregation methods on predictive performances has been evaluated in few studies only [25, 35-38], and these studies focused on a very small number of dimension reduction techniques and/or temporal resolution. The present study extends previous literature by comparing a wide range of aggregation techniques to efficiently incorporate climate data when developing a yield forecast model at the global and national scales.

This study showed that machine learning models using monthly climate averages were more accurate than linear regression to predict soybean yield, which consolidates evidence from numerous studies [9, 22, 23, 27] in this area of research. In addition, RF models were less sensitive to the method chosen to aggregate climate data compared to LR models, whose performances strongly varied among the tested approaches. LR models assume that yield response to climate inputs is linear, while RF algorithm does not make any assumption on the relationship between yield and climate predictors, allowing a greater flexibility in the way climate predictors are introduced in the model. Among the tested approaches to aggregate monthly climate data, PCA or PLSR were the two best dimension reduction approaches in our applications. In PCA, the derivation of principal components is independent from the predicted variable and low variance components are ignored. On the contrary, PLSR works by extracting successive linear combinations of the predictors or independent variables which explain both
variation of the dependent and independent variables. These techniques showed equivalent performances in our study on soybean, which is consistent with one previous study on rapeseed [25]. Applying no dimension reduction (i.e. incorporating the monthly averages directly) also led to accurate predictions at the global scale, while the use of models based on functional approaches (i.e. FPCA, MFPCA, or FPLSR) showed similar or lower performance than the other approaches in models based on monthly climate data. In this study, most of the variability of monthly climate explaining yield was captured by the two or three first scores derived from PCA, FCPA, PLSR, and FPLSR, explaining why models based on all scores did not improve predictive accuracy.

One key result of this study is that the models relying on daily climate data did not show higher prediction accuracy compared to those using monthly or seasonal data resolution. While frequent climate observations contain more information during periods of crop growth, aggregation over a longer time period remove noise and reduce risk of model overfitting. The inclusion of a higher number of variables (e.g., by using a finer temporal resolution) could increase the variance of the parameter estimates and thus actually increase prediction errors. This result is consistent with previous studies which found equivalent [35] or lower [25] prediction performance of models with finer time resolution compared to their monthly equivalents. Applying dimension reduction techniques on daily climate data was found to efficiently resume the information contained in these data (Figure 3) but does maybe not make the best use of all the information they can provide. Other studies showed good performance of predictor selection procedures (e.g. Lasso, ElasticNet, or sparse PLSR) on monthly [25] or weekly [36] climatic data compared to dimension reduction methods.

Similar to month-based models, RF was more appropriate to predict soybean yield from daily climate predictors than LR. PCA and PLSR led to equivalent predictive accuracy, but FPCA performed slightly better with daily predictors. Another study [37] also found that a model based on daily-based FPCA scores was appropriate to predict maize yields in France. However, authors only compared this approach with a model based on monthly averages (with no dimension reduction), which make the comparison between the two approaches (day vs month and FPCA vs no dimension reduction) more difficult. By contrast to analyses conducted on monthly averages,
including all scores derived from PCA, FPCA, MFPCA, PLSR, or FPSR based on daily data led to slightly better predictive performance.

In this study, we essentially focus on the predictive performance (i.e. NSE and RMSE) of the compared approaches to drive model selection. However, among models with similar performance, other criteria might be important to take into account when choosing the “best” model. Here, applying PCA on monthly climate data yielded to higher predictive accuracy, but one can argue that this method would not be the optimal way to incorporate climate data in a predictive model, because the scores derived from a PCA are less easy to interpret compared to monthly averages, for example, which performed equivalently in our study. Although many predictive models rely on monthly climate data, such models are less parsimonious. In addition, PCA scores for a same climate variable present the advantage to be orthogonal, on the contrary of monthly averages which are highly correlated (see Supplementary Figure 4).

Among ML techniques, RF was found to be one of the best algorithm in soybean yield prediction [9, 22, 23] as well as in other major crops including wheat [51] or maize [11]. Other articles report that other algorithms which have not been considered in our study such as neural networks [27] also perform well to forecast yield of soybean. Guilpart and colleagues [9], 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). Differences between deep learning and machine learning techniques were proved to be negligible in crop yield prediction [52].

In addition to its predictive capability, RF can also provide useful information to interpret model’s behavior. Visualization tools are more and more often applied to identify the main factors of crop yield variations using RF or similar ML algorithms (e.g. in [37, 53, 54]). Here, we used importance ranking and partial dependence plots to identify the most influential predictors and found that precipitations and temperatures were the most important (Supplementary Figures 3 and 15), which is consistent with previous studies[55] and with our knowledge on soybean physiology [56]. This is particularly interesting because it means that no information on this aspect was a priori needed, allowing us to overcome any assumption that could make the model setting more suggestive (e.g. determining sensitivity period or threshold). Sensitivity analyses
conducted in different increasing temperature scenarios were consistent with studies simulating soybean yield under climate change in the US [57] and in some regions of Brazil [58]. Although highly unrealistic because temperature won’t 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 [12]. 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 [9].

CONCLUSION

This study simultaneously evaluates the impact of temporal resolutions and aggregations of climate predictors on the performance of machine learning and linear regression models predicting crop yields. Key results of this study indicate that (i) random forest outperformed and was less sensitive to climate aggregation than linear regression; (ii) there was no evidence that using daily climate data improves predictive performance over using monthly data in our models; (iii) principal component analysis applied on monthly data combined with the use of a random forest algorithm to predict crop yields resulted in the most accurate predictions. These findings show that the projected impact of climate on crop yields depends on the temporal resolution and aggregation of climate input data. This suggests that the choice of the most appropriate time resolution and aggregation techniques should be carefully conducted when developing models to predict crop yield, especially to make projections in future climate conditions.

DECLARATIONS

Acknowledgement

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.

Declaration of interest

We declare no competing interests.

Data availability statement

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 v3.2 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.

Code availability statement

All relevant R (version 4.2.2 http://www.r-project.org) scripts and documentation will be available online after acceptance via the project repository (https://github/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.
REFERENCES


Figures

Figure 1

Sites included for analyses.

Figure 2

Step 1: Aggregation of climatic predictors

- Daily climatic data for soybean growing season (maximum and minimum temperatures, precipitation, solar radiation, vapor pressure deficit, reference evapotranspiration)

- Dimension reduction techniques: PCA, FPCA, MFPCA, PLSR, FPLSR

- Standardized average over growing season
- Averages over growing season
- Monthly averages

Step 2: Models for yield prediction

- Scores derived from monthly data
- Scores derived from daily cumulated data

Step 3: Predictive performance

- Random Forest (5 types of models)
- Linear Regression (5 types of models)

i) Model efficiency
ii) Root mean square error of prediction
computed using 2 cross-validation procedures
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.

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.

**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 number of models included in each boxplot is reported on the right. 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.
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

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.3: 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 8

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.3: 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.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.pdf