An agroclimatic based mathematical framework for early prediction of durum wheat yield in Spain

H. Rocha[a,b,*], J.M. Dias[a,b]

[a] *CeBER & FEUC, Av. Dias da Silva 165, 3004–512 Coimbra, Portugal.*
[b] *INESCC, Rua Silvio Lima, Polo II, 3030–290 Coimbra, Portugal.*

Abstract

Durum wheat is an important crop in semi-arid Mediterranean regions as Andalusia, an autonomous community in the southern part of Spain. Accurate early predictions of durum wheat yield can provide precious information for within-season adjustment of crop managing as well as for economical and political stakeholders. In this study, an alternative methodology to mechanistic crop models is proposed for within-season early prediction of durum wheat yield in Spain based on estimates for its larger producer community, Andalusia. The proposed mathematical framework embeds the construction of Radial Basis Functions (RBF) interpolation models based on the sown area and a large number of climatic variables. Global warming and increasing occurrence of extreme weather events are only two of the factors that make crop yield forecast extremely difficult as they can lead to an increased interannual yield variability. Nevertheless, the RBF models proposed presented good quality yield predictions clearly outperforming multivariate linear models used as benchmark. Moreover, RBF models' predictions made four months prior to harvest are able to capture the trend of the yield series as well as near-harvest predictions.

Keywords: Durum wheat yield, Mediterranean climate, Radial Basis Functions, Cross-validation, Variable Screening

1. Introduction

Wheat is the most popular agricultural crop in the northern hemisphere occupying the main growing area of its temperate zone. It is classified into two types according to the texture of the grain: durum wheat, mainly used for manufacturing of pasta, and soft wheat, also known as bread wheat as baking is its main use. Durum wheat is more resistant to extreme climatic conditions as compared to soft wheat [7]. For that reason, durum wheat is an important

*Corresponding author. Tel.: +351 239 851040; fax: +351 239 824692.
Email addresses: hrocha@mat.uc.pt (H. Rocha), joana@fe.uc.pt (J.M. Dias)

crop in semi-arid Mediterranean regions [12], where extreme climatic conditions as drought and high temperatures are the typical production constraining variables. One of such regions is Andalusia, an autonomous community in the southern part of Spain. Andalusia is an important producer of wheat at an international level and represents 8% of the surface area and 9% of the production of the European Union. Furthermore, Andalusia is the largest durum wheat producing community in Spain (∼70%) [3].

Accurate early predictions of durum wheat yield can provide precious information for within-season adjustment of crop managing, e.g. optimization of fertilization as nitrogen effect is well known [1], for economical players, e.g. commodity trading, or for political stakeholders, e.g. assessment of climate change impact [10]. Predicting durum wheat yield has received attention from researchers in recent years (see, e.g. [12, 13, 14, 29]). Typically, mechanistic crop simulation models that attempt to mimic interactions between different factors that determine crop development and growth are used to estimate crop yield. E.g., the CERES-wheat model was used for early prediction of durum wheat yield in Central Italy [12]. CERES-wheat model simulate the development, growth and yield of wheat considering the interactions between soil, plant genetics, climate and nitrogen supply [22]. Other well-known generic mechanistic crop simulation models used to predict wheat yield include SIRIUS [11] and EPIC [34]. Regardless of the models' structure, their goal is to simulate crop development and growth in response to a number of different input variables fed to the model in a daily time basis. Yield prediction based on mechanical crop models shows very good results for short-term estimations and for similar scenarios, i.e. fields with similar soil, local weather and nitrogen supply [15]. However, prediction accuracy deteriorates for longer-term estimates. One of the difficulties for obtaining accurate long-term estimations is related to unknown future weather conditions that need to be fed to the models and whose forecast accuracy diminishes with long-term predictions. Another difficulty is to be able to provide national estimates as aggregated yields depend on yields obtained for a number of fields with different soil, local weather and nitrogen supply characteristic [4].

The goal of this paper is to propose an alternative methodological approach to provide early predictions of durum wheat yield in Spain based on estimates for its larger producer community, Andalusia. As illustrated in Fig. 1, the yield trends are identical for Spain and Andalusia, as it would be expected knowing that about 70% of the Spanish durum wheat yield is granted by Andalusia. In Fig. 1 it is also possible to observe that both the Spanish and the Andalusian area of durum wheat has clearly decreased since the beginning of the century. That decreasing trend has occurred worldwide due to several factors including global warming, increasing occurrence of extreme weather events, and new agricultural policy guidelines that have steered farmers towards more profitable crops [2]. In addition to the sown area, Mediterranean climatic conditions are the most important variables to explain the annual variation of durum wheat yield [8]. Weather has a major impact on plants as well as pests and diseases and it has been reported that as much as 80% of the variability of agricultural

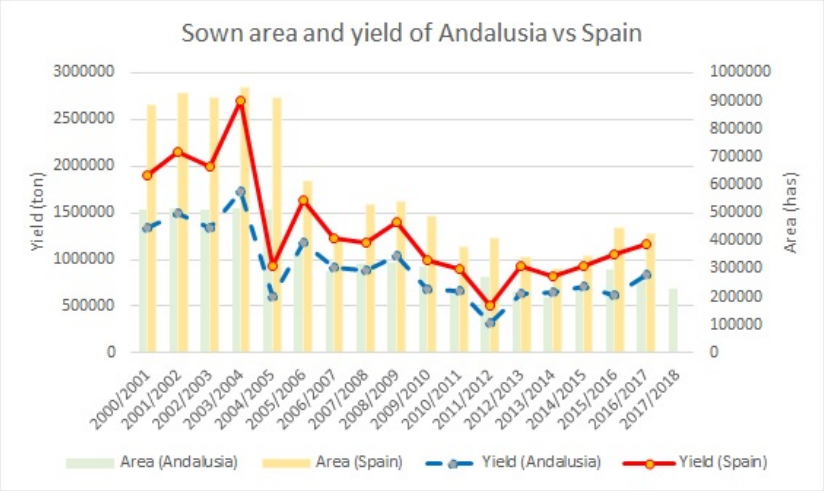


Figure 1: Sown area (has) and durum wheat yield (ton) of Andalusia and Spain for the time frame in study – season 2000/2001 up to the season 2016/2017. Andalusia sown area for 2017/2018 season is provisional but is also displayed.

production is due to the variability in weather conditions [19]. A mathematical framework based on the sown area and a large number of climatic variables is proposed to answer the question: "at a given month of the season, what is the durum wheat yield expected for Spain?". This framework embeds the construction of Radial Basis Functions interpolation models to estimate durum wheat yield considering the sown area and selected subsets of climatic variables calculated through a tailored variable screening strategy that aims at selecting weather variables based on their prediction features.

2. Materials and methods

2.1. Durum wheat agroclimatic data in Andalusia

Andalusia is located in the southern part of Spain, east of the southern part of Portugal and north of the Mediterranean Sea, as illustrated in Fig. 2. It is the second largest Spanish autonomous community in area being divided into eight provinces: Huelva, Seville, Cádiz, Córdoba, Málaga, Jaén, Granada and Almería. These provinces have distinct weather conditions that are monitored by a set of agroclimatic stations since the year 2000. These stations can perform, and store daily, several meteorological measurements in an automated way [16]. Sown areas and durum wheat yield historical series are available for each of these provinces. Thus, predictions for these provinces can be calculated considering the local climatic conditions as an attempt to minimize the weather bias. In Fig. 2, the provinces highlighted in bold, Cádiz, Córdoba, Huelva, Málaga and Seville, are the largest durum wheat producers corresponding to an aggregated

Figure 2: Andalusia geographic location and its corresponding eight provinces. Highlighted in bold, Cádiz, Córdoba, Huelva, Málaga and Seville, are the largest durum wheat province producers.

yield of approximately 95% of the Andalusia annual durum wheat yield. For that reason, predictions are only calculated for these five provinces starting in season 2000/2001 up to the current season 2017/2018. Historical data of durum wheat yields and sown area for each of these five provinces are displayed in Table 1.

The recommended sowing date in Andalusia is from middle November to middle December. The harvest will be made when the grain has reached physiological maturity which typically occurs in June. Mechanistic crop simulation models are mainly focused on the months of March, April and May, since during that time frame the most important development stages occur and can thus be simulated: in March the tillers grow, in April stem elongation and inflorescence occur while in May occurs the grain-filling [12]. The advantage of complex mechanical crop models is to capture soil–weather–crop interactions, for a specific location, allowing the identification of the most important variables for the different crop stages. However, all the climatic information prior to these important crop development stages is ignored and that information is key for two reasons: it is important for within-season crop development as well and may

Table 1: Tonnes of durum wheat yields and sown area for the five largest durum wheat producer provinces of Andalusia. Sown areas for season 2017/2018 are provisional.

	Cádiz		Córdoba		Huelva		Málaga		Seville	
Season	Area (Has)	Yield (Ton)	Area (Has)	Yield (Ton)	Area (Has)	Yield (Ton)	Area (Has)	Yield (Ton)	Area (Has)	Yield (Ton)
2000/2001	89960	240650	130959	379485	18064	43571	35084	75059	192142	505028
2000/2002	88967	238856	133493	421615	18067	50207	34603	68800	191594	604808
2000/2003	87662	231238	135725	357031	18351	43677	36446	106985	182291	509863
2000/2004	87330	307617	136545	491174	17819	48959	34598	110223	190440	647415
2000/2005	85278	201053	136024	127912	17980	21972	35365	40036	185686	192757
2000/2006	59692	182634	95310	326971	13517	40500	29279	93455	128525	455153
2000/2007	57283	182634	78853	214098	9791	33500	26969	89229	100325	346110
2000/2008	66589	182634	84702	242363	9517	26500	27853	88747	106440	299120
2000/2009	64422	181681	84600	267296	11413	43368	25097	80971	118875	426739
2000/2010	61204	98384	71169	146808	12375	23205	20715	52561	127084	328050
2000/2011	53886	163706	53471	142022	8925	31238	20163	48877	82325	248490
2000/2012	60719	101704	60803	57606	11807	10036	20371	40704	107395	102360
2000/2013	50802	122383	50274	146297	9810	41889	15674	42897	86513	261882
2000/2014	42392	137771	47337	165630	9362	28086	15657	33595	76020	264510
2000/2015	49623	159750	53033	135384	10708	41048	15315	43570	95068	302399
2000/2016	65174	124616	59879	157230	13045	37273	19587	35620	125114	241924
2000/2017	65174	177139	51050	164891	11114	51124	15200	37225	115892	382056
2000/2018	58614	–	51045	–	11114	–	15250	–	92461	–

have buried the climatic trends for the remaining of the season. For these reasons, all weather data ranging from the month following the end of the previous season (July) to the end of the current season (June) is considered aiming at longer term yield forecasts.

The daily measurements stored daily in the agroclimatic stations covering the different provinces of Andalusia include minimum, mean and maximum temperature (^{o}C), solar radiation (MJ/m^2), rainfall (mm), evapotranspiration (mm) and relative humidity (%). The most important weather variables affecting durum wheat yield are rainfall and temperature [20]. Low precipitation and its temporal distribution in a Mediterranean environment can explain as much as 75% of the wheat yield variability [6], while quality of durum wheat under Mediterranean conditions is mainly affected by precipitation and air temperature [9]. Based on daily weather data provided by the agroclimatic stations, monthly variables were calculated for each province including monthly minimum, mean and maximum temperature and monthly accumulated rainfall. The average values of these monthly variables for season 2000/2001 up to season 2016/2017 and months from July to February in current season 2017/2018 are displayed in Fig. 3. These monthly variables can provide information for thermal or water stress conditions occurring before and during the crop cycle. However, the effect of water deficit or high temperatures is not univocal [12], and the combined effect of meteorological variables can lead to different results compared to their single impact [35]. Therefore, adding to these four monthly variables, additional monthly weather variables were considered, including the number of days with no rainfall in the month, and the monthly mean of following variables: maximum and minimum temperatures, solar radiation, evapotranspiration and relative humidity. The total number of monthly weather variables considered exceeded one hundred (10×12 months $= 120$ monthly weather variables).

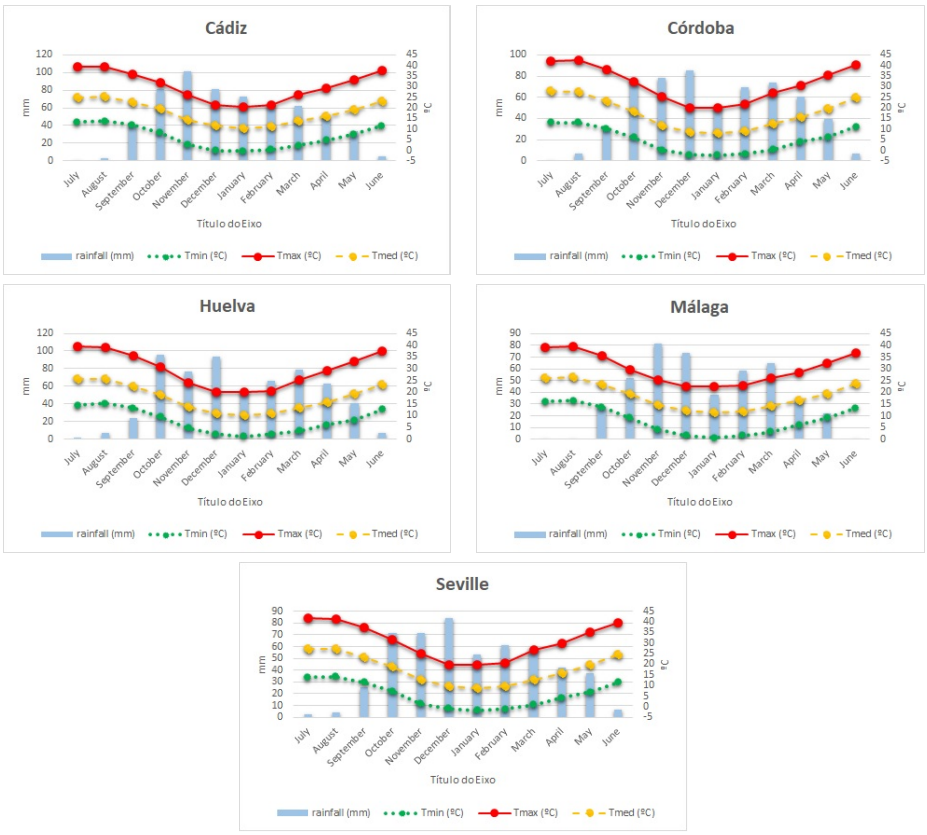


Figure 3: Rainfall and temperatures for the five largest durum wheat producer provinces of Andalusia considering the average values for the time frame in study, season 2000/2001 to season 2016/2017 and months from July to February in current season 2017/2018.

2.2. Radial Basis Functions Interpolation Models

There are many examples in the literature of successful applications of Radial Basis Functions (RBF) interpolation models, including in agriculture [23, 30], radiotherapy [24, 25] or aeronautics [26, 27]. RBF interpolation models have been shown to provide excellent surrogates for modeling sparsely known responses that, in practice, can be used for optimization or forecast purposes [28]. The response surfaces obtained by RBF interpolation models are able to capture the relationships between predictors (explanatory variables) and outcome(s) (response variable(s)) which encourages its use for prediction of unknown outcomes given the predictors' values. Moreover, it has been shown that stochastic models in the presence of noise coincide with the corresponding radial basis algorithm [36]. Computation of RBF response surfaces can be obtained even for a small set of poorly distributed data points in a high dimensional space. Despite this important feature, RBF response surfaces' shape between data points is strongly correlated to the basis functions used. Thus, for a given set of data

points some basis functions can capture the correct trends while other basis functions may fail to extract the main relationships between predictors and outcome displaying misleading trends. Therefore, it is fundamental to choose the most appropriate radial basis function, for a given set of data points, using quantitative measures instead of a qualitative choice made *a priori* most of the times based on authors' intuition [28]. RBF interpolation is briefly described in the following section.

2.2.1. RBF interpolation

Let $\mathbf{x}^1, \ldots, \mathbf{x}^N$ be a set of data points with $\mathbf{x}^k = (x_1^k, \ldots, x_n^k) \in \mathbb{R}^n$, $k = 1, \ldots, N$ and let's assume that the response $y(\mathbf{x})$ is only known at these N data points. If $\phi(x)$ represents a given basis function, a RBF model $g(\mathbf{x})$ can be mathematically expressed as:

$$g(\mathbf{x}) = \sum_{k=1}^{N} \alpha_k \phi(\|\mathbf{x} - \mathbf{x}^k\|), \tag{1}$$

where $\|\mathbf{x} - \mathbf{x}^k\|$ corresponds to the Euclidean distance between $\mathbf{x}$ and $\mathbf{x}^k$,

$$||\mathbf{x} - \mathbf{x}^k|| = \sqrt{\sum_{i=1}^{n} |\theta_i| \left(x_i - x_i^k\right)^2},$$

parameterized by scalars $\theta_1, \ldots, \theta_n$ [28]. The interpolation conditions given by the following system of equations

$$\sum_{k=1}^{N} \alpha_k \phi(||\mathbf{x}^j - \mathbf{x}^k||) = y(\mathbf{x}^j), \quad \text{for } j = 1, \ldots, N,$$

allow a straightforward calculation of the α coefficients in Eq. 1 for each set of fixed θ parameters. The most prominent examples of basis functions that are typically used in practice are the multiquadric RBF, $\phi(x) = \sqrt{1 + x^2}$, the thin plate RBF, $\phi(x) = x^2 \ln(x)$, the cubic RBF, $\phi(x) = x^3$, and the Gaussian RBF, $\phi(x) = \exp(-\mathrm{x}^2)$, graphically illustrated in Fig. 4. The first three RBFs can be used to capture possible growth rates of the response – linear, almost quadratic, and cubic, respectively. The Gaussian RBF can be used to capture a possible exponential decay trend of the response [21]. As most of the times it is not possible to know *a priori* the response trends, cross-validation can be used to determine the most suitable basis function for the data set at hand and also to compute the θ parameters.

2.2.2. Cross-validation

Selection of the most suitable basis function for the set of data points at hand can simply be done by testing different possible RBFs. However, the same basis function $\phi(x)$ gives origin to RBF interpolation models that behave differently between data points for distinct sets of model parameters $\theta_1, \ldots, \theta_n$. The

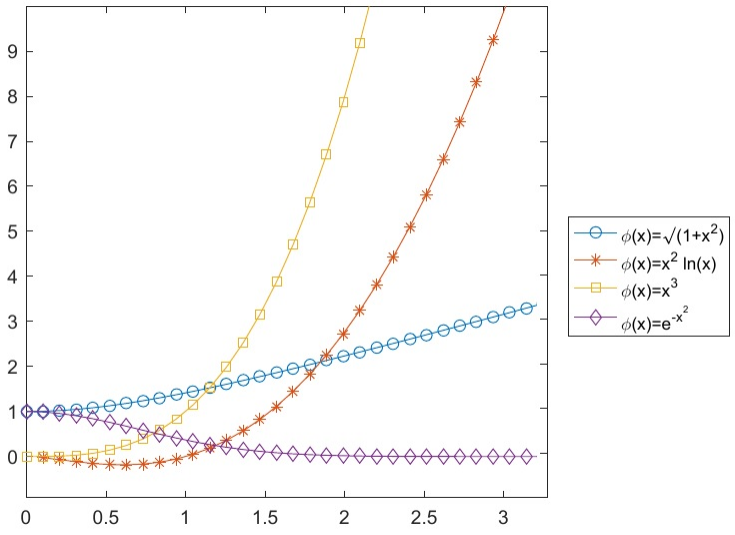


Figure 4: Graphical illustration of multiquadric RBF, $\phi(x) = \sqrt{1+x^2}$, thin plate RBF, $\phi(x) = x^2\ln(x)$, cubic RBF, $\phi(x) = x^3$, and Gaussian RBF, $\phi(x) = \exp(-\mathrm{x}^2)$.

model parameter tuning should be focused on finding the RBF model with best prediction ability. Cross-validation (CV) can be used as proxy of the model's ability to predict out-of-sample responses [32]. Thus, for each RBF $\phi(x)$, the θ model parameters are calculated in order to minimize the CV error, i.e. to improve the prediction features of the model. Algorithm 1 depicts the leave-one-out CV routine used for calculating the θ model parameters that leads to the RBF model with best predicting ability [28].

It is important to highlight that the minimization of the CV error, $E^{CV}(\theta_1, \ldots, \theta_n)$, leads to a demanding highly non-convex global optimization problem in a possibly high dimension n. This optimization problem could be simplified if all θ parameters were considered equal. However, this simplification would erase the benefits of using different θ model parameters which enable the scaling of each variable x_i based on its importance on explaining the response' variance. Thus, the use of different θ enables an implicit variable screening embedded in the RBF parameter tunning which makes the optimization effort worth it. However, this effort might become unbearable for very large n which makes the selection of appropriate subsets of variables an important step in this framework.

2.3. Variable Screening

Generically, a regression model with a large number of explanatory variables may present several issues including data over-fitting. In this context, a large number of explanatory variables brings an extra issue: increased difficulty on

Algorithm 1 Leave-one-out CV for RBF model parameter tuning

Input:

- Set of data points, $\mathbf{x}^1, \ldots, \mathbf{x}^N$.
- Known response of the N data points, $y(\mathbf{x}^1), \ldots, y(\mathbf{x}^N)$.

Iteration:

1. Consider $\theta_1, \ldots, \theta_n$, a fixed set of RBF model parameters.
2. For $k = 1, \ldots, N$, compute the N RBF interpolation models $g_{-k}(\mathbf{x})$ considering the subset of data points $(\mathbf{x}^j, y(\mathbf{x}^j))$ for $1 \leq j \leq N, j \neq k$.
3. Compute the leave-one-out CV error:

$$E^{CV}(\theta_1, \ldots, \theta_n) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left(g_{-k}(\mathbf{x}^k) - y(\mathbf{x}^k)\right)^2}. \tag{2}$$

obtaining the optimal value of the CV error in Eq. 2. As previously described, the potential number of predictors of durum wheat yield is over one hundred. The purpose of variable screening is to select a subset of explanatory variables that most influence response variation. Thus, if the change of a given variable leads to an insignificant change in response then that variable should not be selected or should be removed.

The traditional procedure for selecting a subset of explanatory variables is to consider the input variables that have a statistically significant correlation with the outcome. In this study, this common procedure would select a very large number of explanatory variables which makes the optimization of the CV error more difficult and may lead to over-fitted RBF models that behave extremely well in-sample but perform poorly out-of-sample. Selecting procedures that require the outcome for certain data points, e.g. ANOVA, are not feasible as well for the problem at hand. Other variable screening procedures are only valid for data points that follow a specific distribution which is not the case. Such an example is the main effects estimate method [33], that requires the data points to be uniformly distributed in a rectangular domain.

More flexible variable screening procedures include the forward and the backward selection methods that are commonly used to calculate the explanatory power of linear model's predictors. In general, we can assume that these procedures are suitable for variable screening when data is fitted by nonlinear models as well. A combination of forward and backward selection methods is proposed focusing in the predicting features of the models instead of the coefficient of determination (R^2) commonly used. Algorithm 2 depicts the variable screening strategy that starts with forward selection and ends with backward screening.

Algorithm 2 Forward-backward procedure for variable selection

Input:

- Set of data points, $\mathbf{x}^1, \ldots, \mathbf{x}^N$.
- Known response of the N data points, $y(\mathbf{x}^1), \ldots, y(\mathbf{x}^N)$.
- Auxiliary data points, $\hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N$, initially with 0 variables.

Forward procedure:

$E_{min}^{CV} \leftarrow +\infty$
Continue $\leftarrow 1$
While Continue
  For $i = 1$ to n
    If x_i **does not belong** to the set of variables of auxiliary data points $\hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N$
      $\check{\mathbf{x}}^1, \ldots, \check{\mathbf{x}}^N \leftarrow \hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N \oplus x_i$, where $\oplus$ corresponds to adding x_i to the set of variables of the auxiliary data points $\hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N$
      Fit the RBF interpolation model $g_i(\check{\mathbf{x}})$ to the points $(\check{\mathbf{x}}^k, y(\mathbf{x}^k))$ for $1 \leq k \leq N$
      and using Eq. (2) calculate the corresponding CV error, E_i^{CV}
    End If
  End For
  If $\mathrm{argmin}_{1\leq \mathrm{i} \leq \mathrm{n}} \mathrm{E}_\mathrm{i}^{\mathrm{CV}} < \mathrm{E}_{\mathrm{min}}^{\mathrm{CV}}$
    $E_{min}^{CV} \leftarrow \mathrm{argmin}_{1\leq \mathrm{i} \leq \mathrm{n}} \mathrm{E}_\mathrm{i}^{\mathrm{CV}}$
    $\hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N \leftarrow \hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N \oplus x_i$
  Else
    Continue $\leftarrow 0$
  End If
End While

Backward procedure:

Continue $\leftarrow 1$
While Continue
  For $i = 1$ to n
    If x_i **belong** to the set of variables of auxiliary data points $\hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N$
      $\check{\mathbf{x}}^1, \ldots, \check{\mathbf{x}}^N \leftarrow \hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N \ominus x_i$, where $\ominus$ corresponds to removing x_i from the set of variables of the auxiliary data points $\hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N$
      Fit the RBF interpolation model $g_i(\check{\mathbf{x}})$ to the points $(\check{\mathbf{x}}^k, y(\mathbf{x}^k))$ for $1 \leq k \leq N$
      and using Eq. (2) calculate the corresponding CV error, E_i^{CV}
    End If
  End For
  If $\mathrm{argmin}_{1\leq \mathrm{i} \leq \mathrm{n}} \mathrm{E}_\mathrm{i}^{\mathrm{CV}} < \mathrm{E}_{\mathrm{min}}^{\mathrm{CV}}$
    $E_{min}^{CV} \leftarrow \mathrm{argmin}_{1\leq \mathrm{i} \leq \mathrm{n}} \mathrm{E}_\mathrm{i}^{\mathrm{CV}}$
    $\hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N \leftarrow \hat{\mathbf{x}}^1, \ldots, \hat{\mathbf{x}}^N \ominus x_i$
  Else
    Continue $\leftarrow 0$
  End If
End While

At the first iteration, data points with a single input variable are fitted using RBF interpolation models. All input variables are tested, one at a time, and the variable leading to the RBF model with best predicting feature (lowest CV error) is selected. At the following iterations, variables not yet included are tested, one at a time, and the best is added until the CV error cease to improve. It is worth to point out that at the end of a given iteration, p, of the forward procedure, the subset of variables selected probably does not correspond to be best subset of p variables. The reason is simply because this procedure does not test all possible combinations of p variables – $\binom{n}{p} = \frac{n!}{p!(n-p)!}$. Therefore, after the forward selection, a backward screening is performed aiming to further decrease the CV error. The backward selection follows a strategy similar to the forward procedure but now variables are removed instead of added.

3. Computational results

Computational tests were performed using MATLAB(R2018a) on a Intel Core PC @ 2.60Ghz. The optimization of the CV error (Eq. 2) was performed using the optimization toolbox of MATLAB, concretely *fminsearch*, an implementation of the Nelder-Mead derivative-free optimization procedure [17]. The basis functions, and corresponding RBF optimal θ parameters, were selected based on the optimal CV error obtained as this measure was used as surrogate of the RBF model prediction ability [28].

A typical measure used to assess the quality of the yield forecast is the normalized root mean squared error [5, 11]. The normalized root mean squared error of crop yield predictions (P) can be calculated based on deviations from actual yields (A) accumulated over time represented by M seasons:

$$nRMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(P_i - A_i)^2} \times \frac{100}{\bar{A}},$$

where $\bar{A}$ is the mean value of the actual yields. The normalized root mean squared error gives a measure of the relative difference of simulated versus observed data. The lower the value of nMRSE, the higher the ability of the models to capture the interannual variability of the yields. The simulation is considered excellent when nRMSE is inferior to 10%, good when nRMSE is between 10% and 20%, fair when nRMSE is between 20% and 30%, and poor when nRMSE is greater than 30% [5, 11].

Another measure used to assess the quality of the yield forecast is the mean prediction error (or mean relative absolute error) calculated as

$$\frac{1}{M}\sum_{i=1}^{M}\frac{|P_i - A_i|}{P_i} \times 100.$$

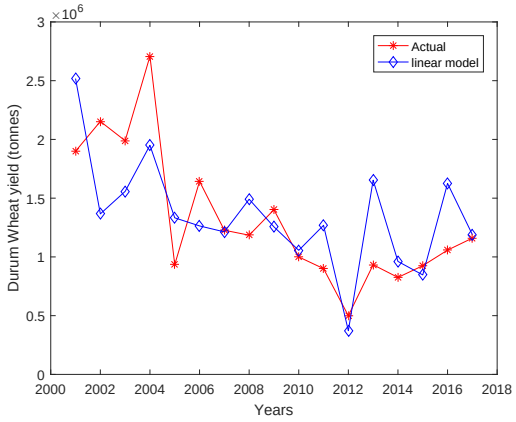
This measure provides quantitative information regarding the differences between estimated and observed yields.

The tailored strategy for early prediction of durum wheat yield in Spain for each of the $M = 17$ seasons in study, 2000/2001–2016/2017, was performed as follows:

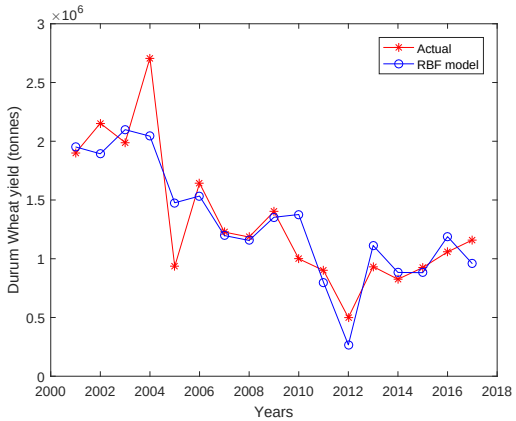
- Eliminate all the data related to the season being predicted including sown areas, climatic variables and yields.
- Set the month the forecast will be made. Eliminate all the climatic data after that month for all the seasons. For each province, considering the remaining data:
  - compute a subset of explanatory variables using Algorithm 2 for each basis function considered;
  - determine the best RBF model that corresponds to the lowest CV error obtained in previous step;
  - using the best RBF model estimate the durum wheat yield for each of the five provinces;
- Compute five durum wheat yield predictions for Spain by scaling each of the provinces' estimate by their mean contribution to the overall durum wheat yield of Spain;
- Sort the five estimates and take the middle value (median) as the durum wheat yield forecast for Spain in that season.

This strategy uses cross-validation to assess the models prediction performance which is a common procedure including for wheat yield prediction (see, e.g. [15, 18]). As most of the models used to forecast durum wheat yield are multivariate linear models (see, e.g. [12, 13, 15, 31]), the strategy sketched for obtaining the best set of variables and the RBF model with highest predictive ability was also used to obtain a multivariate linear model for benchmark purposes. MATLAB implementation *lscov* was used to obtain the ordinary least squares solution, x, of the linear system of equations $Ax = b$, where columns of A correspond to the selected variables, b corresponds to the Spanish durum wheat yield for the $M-1$ seasons used to fit the model with the goal of predicting the removed season.

Fig. 5 displays the forecast results obtained by linear models and RBF models for near-harvest yield estimate, i.e. the month of forecast is June, meaning that only information available until June is used in for prediction. The nMRSE obtained by RBF models was 18.9% which is considered good while the linear models obtained a poor forecast (nMRSE=32.2%). The mean prediction error obtained by RBF models was 16.3% clearly outperforming the linear models that obtained a mean prediction error of 26.6%. Furthermore, RBF models were able to capture most of the trend of the durum wheat yield series, which is quite difficult for such irregular series, while linear models clearly fail to do so. We should stress the decisive role of variable screening that clearly enhanced forecasts' quality. It should be also reminded that each of the framework steps



(a)



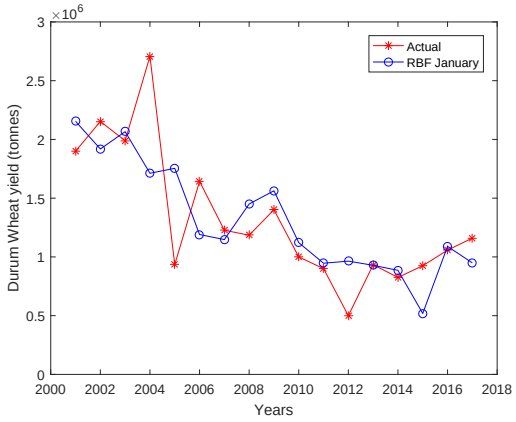
(b)

Figure 5: Durum wheat yield previsions obtained by linear models (a) and durum wheat yield previsions obtained by RBF models (b) considering June as the month of forecast.

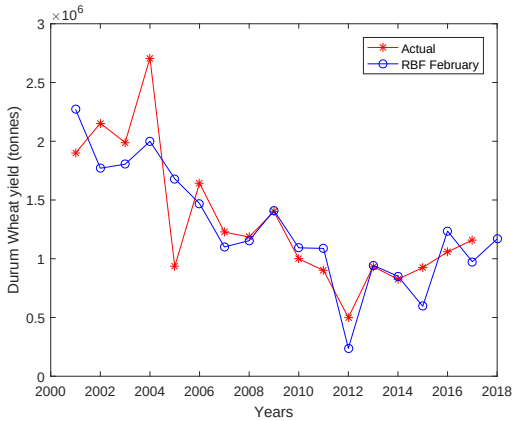
including variable screening, basis selection and RBF model fitting are done in an automated way for an out-of-sample forecast. Thus, different subsets of variables were obtained for different regions and for different seasons to be predicted. While displaying all the subsets of predictors obtained might be confusing as they are too many and depend on geographical region and season, it might be relevant to know that the variables that appear more often were the sown area (has), the accumulated rainfall (mm) in January and the maximum temperature (^{o}C) in October. It is interesting to note that the two main climatic variables emerge: rainfall and air temperature.

In order to make early predictions of durum wheat yield, the previous strategy was used considering the months preceding June. As the estimated sown area for each province for the current season is only available during the month of January, predictions are only possible at the end of that month. The nMRSE obtained by RBF models at the end of January up to the end of May were 30.4%, 25.1%, 24.8%, 21.9%, 20.3%, respectively. This means that the forecast obtained at the end of January is poor while the forecast obtained by the end of February is the first fair forecast. The remaining forecasts are also fair but it is worth to point out that there is only a small deterioration of the results until February. Estimates obtained with RBF models at the end of January and at the end of February are displayed in Fig. 6. It can be observed that forecast at the end of January fails to capture most of the trend of the yield series while the first fair forecast obtained by the end of February already captures most of the series trend. As the climatic data for this season is available up to the end of February, a forecast for the current season, 2017/2018, considering climatic data up to the end of February is already displayed in Fig. 6. Despite the estimated decrease in the sown area, the prevision foretell a slight yield increase for the current season as provisional yield for 2016/2017 is 1159900 ton and forecast for 2017/2018 is 1168200 ton.

A RBF model considering only the three variables that emerged the most was also computed and the forecast results are displayed in Fig. 7. Note that this model can obtain a forecast in the end of January as accumulated rainfall (mm) in January is its late climatic variable. The nMRSE obtained was 17.2% while the mean prediction error 11.9%. Despite the very good results, such model is less reliable for future forecasts than the previously reported models. In fact, the yield estimates provided by this model cannot be considered as out-of-sample estimates because the variable selection incorporates information from all series. Actually, for the seasons where that variables did not emerge, the estimates are quite bad. E.g., for season 2011/2012 the prediction error is superior to 70%. An estimate for the current season, 2017/2018, was also computed using this model and the prevision foretell a slight yield decrease (1101300 ton) compared to the last season. It is possible to obtain RBF models with almost perfect in-sample accuracy but they are unreliable for future forecasts. Nevertheless, in this particular case, estimates obtained are quite similar.



(a)



(b)

Figure 6: Durum wheat yield previsions obtained by RBF models at the end of January (a) and durum wheat yield previsions obtained by RBF models at the end of February (b).

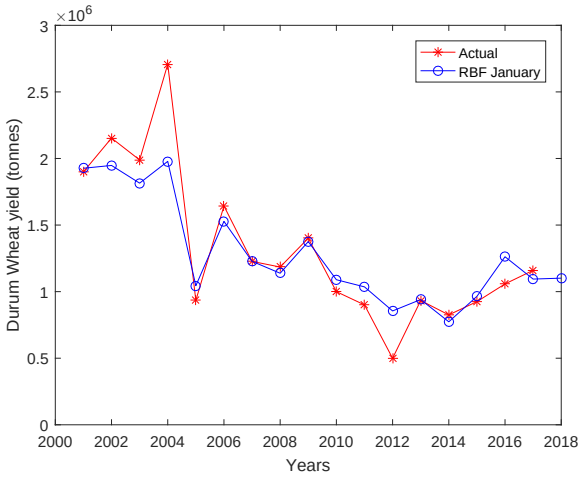


Figure 7: Durum wheat yield previsions obtained by RBF models considering the sown area (has), the accumulated rainfall (mm) in January and the maximum temperature (^{o}C) in October.

4. Discussion and conclusions

In this paper, an alternative methodology to deterministic crop models for within-season early prediction of durum wheat yield is proposed. Global warming and increasing occurrence of extreme weather events are only two of the factors that make crop yield forecast extremely difficult as they lead to an increased interannual yield variability. Nevertheless, the RBF models proposed presented good predictions clearly outperforming the commonly used multivariate linear models. The strategy drafted for variable selection was decisive for obtaining RBF models with improved predicting ability. Although some variables emerge more often than others, their role should be carefully interpreted in this context. While in complex system deterministic crop models it is possible to identify climatic variables that have an important role during the more susceptible crop stages, here the main emerged meteorological variables should simply be seen as good yield predictors rather than having a decisive role in a particular crop development stage.

One of the difficulties faced in this study was related to the goal of estimating national yields instead of local predictions. Aggregated yields are obviously more challenging as the different yields' error are added as well as the error associated to scale. E.g., if actual durum wheat yield from the five larger producer communities in Andalusia are used to compute the Spanish yield instead of the RBF model estimates, the mean prediction error is 4.2%. Thus, comparison of the results achieved is 16.3% against 4.2% which is a better perspective than a comparison 16.3% against 0%.

Typically, estimates obtained by mechanistic crop models near-harvest are

much more accurate than results obtained by RBF models for the month of June, at least for a specific location. However, long-term forecasts quality results deteriorate markedly using deterministic crop models which do not occur for the framework proposed. Thus, this mathematical framework gathering RBF models and variable screening show great potential to provide early aggregated yield predictions. Although near-harvest forecast are not sharp accurate, long-term estimates are able to capture the trend of the yield series, making this framework a valid alternative to be combined/merged with deterministic crop models. Furthermore, the use of RBF models, as alternative to linear models, could enhance the ability to capture the soil–climatic–plant interactions of dynamic and complex system deterministic crop models.

Acknowledgements

This work has been supported by the Fundação para a Ciência e a Tecnologia (FCT) under project grant UID/MULTI/00308/2013.

References

[1] Abedi, T., Alemzadeh, A., & Kazemeini, S. A. (2011). Wheat yield and grain protein response to nitrogen amount and timing. *Australian Journal of Crop Science, 5*, 330–336.

[2] Aggelopoulos, S., Pavloudi, A., Manolopoulos, I., & Kamenidou, I. (2008). The attitudes and views of farmers on the new common agricultural policy and the restructuring of crops: the case of Greece. *American-Eurasian Journal of Agricultural & Environmental Science 4*, 397–404.

[3] Anuário de Estadistica del Ministerio de Agricultura, Alimentación y Medio Ambiente, http://www.mapama.gob.es/

[4] Bakker, M. M., Govers, G., Ewert, F., Rounsevell, M., & Jones, R. (2005). Variability in regional wheat yields as a function of climate, soil and economic variables: Assessing the risk of confounding. *Agriculture, Ecosystems & Environment, 110*, 195–209.

[5] Bannayan, M., & Hoogenboom, G. (2009). Using pattern recognition for estimating cultivar coefficients of a crop simulation model. *Field Crops Research, 111*, 290–302.

[6] Blum, A., & Pnuel, Y. (1990). Physiological attributes associated with drought resistance of wheat cultivars in a Mediterranean environment. *Australian Journal of Agricultural Research, 41*, 799–810.

[7] Cossani, C. M., Slafer, G. A., & Savin, R. (2011). Do barley and wheat (bread and durum) differ in grain weight stability through seasons and waternitrogen treatments in a Mediterranean location? *Field Crops Research, 121*, 240–247.

[8] Diacono, M., Castrignano, A., Troccoli, A., De Benedetto, D., Basso, B., & Rubino, P. (2012). Spatial and temporal variability of wheat grain yield and quality in a Mediterranean environment: a multivariate geostatistical approach. *Field Crops Research, 131*, 49–62.

[9] Garrido-Lestache, E., Lopez-Bellido, R.J., & Lopez-Bellido, L. (2005). Durum wheat quality under Mediterranean conditions as affected by N rate, timing and splitting, N form and S fertilization. *European Journal of Agronomy, 23*, 265–278.

[10] Hoogenboom, G. (2000). Contribution of agrometeorology to the simulation of crop production and its application. *Agricultural and Forest Meteorology, 103*, 137–157.

[11] Jamieson, P. D., Semenov, M. A., Brooking, I. R., & Francis, G. S. (1998). Sirius: A mechanistic model of wheat response to environmental variation. *European Journal of Agronomy, 8*, 161–179.

[12] Marta, A.D., Orlando, F., Mancini, M., Guasconi, F., Motha, R., Qu, J., & Orlandini, S. (2015). A simplified index for an early estimation of durum wheat yield in Tuscany (Central Italy). *Field Crops Research, 170*, 1–6.

[13] Ferrise, R., Toscano, P., Pasqui, M., Moriondo, M., Primicerio, J., Semenov, M. A., & Bindi, M. (2015). Monthly-to-seasonal predictions of durum wheat yield over the Mediterranean Basin. *Climate Research, 65*, 7–21.

[14] García del Moral, L. F., Rharrabti, Y., Villegas, D., & Royo, C. (2003). Evaluation of Grain Yield and Its Components in Durum Wheat under Mediterranean Conditions. *Agronomy Journal, 95*, 266–274.

[15] Gouache, D., Bouchon, A., Jouanneau, E., & Brisc, X. (2015). Agrometeorological analysis and prediction of wheat yield at the departmental level in France. *Agricultural and Forest Meteorology, 210*, 1–10.

[16] Instituto de investigación y formación agraria y pesquera, www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/

[17] Nelder, J., & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal, 7*, 308–313.

[18] Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L., & Mouazen, A.M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture, 121*, 57–65.

[19] Petr, J. (1991). *Weather and Yield*. Amsterdam:Elsevier.

[20] Porter, J. R., & Semenov, M. A. (2005). Crop responses to climatic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*, 2021–2035.

[21] Powell, M. (2002). Radial Basis Function Methods for Interpolation to Functions of Many Variables. *HERMIS: An International Journal of Computer Mathematics and its Applications, 3*, 1–23.

[22] Ritchie, J. T., & Otter, S. (1985). Description and performance of CERES-wheat. A user oriented wheat yield model. ARS Wheat Yield Project. *ARS, 38*, 159–177.

[23] Rocha, H., & Dias, J.M. (2018). Honey Yield Forecast Using Radial Basis Functions. In *International Workshop on Machine Learning, Optimization, and Big Data (MOD 2017)*, Volterra, Tuscany, Italy.

[24] Rocha, H., Dias, J.M., Ferreira, B.C., & Lopes, M.C. (2013). Selection of intensity modulated radiation therapy treatment beam directions using radial basis functions within a pattern search methods framework. *Journal of Global Optimization, 57*, 1065–1089.

[25] Rocha, H., Dias, J.M., Ferreira, B.C., & Lopes, M.C. (2013). Beam angle optimization for intensity-modulated radiation therapy using a guided pattern search method. *Physics in Medicine & Biology, 58*, 2939.

[26] Rocha, H., Li, W., & Hahn, A. (2006). Principal Component Regression for Fitting Wing Weight Data of Subsonic Transports. *Journal of Aircraft, 43*, 1925–1936.

[27] Rocha, H. (2008). Model parameter tuning by cross validation and global optimization: application to the wing weight fitting problem. *Structural and Multidisciplinary Optimization, 37*, 197–202.

[28] Rocha, H. (2009). On the selection of the most adequate radial basis function. *Applied Mathematical Modelling, 33*, 1573–1583.

[29] Romero, J. R., Roncallo, P. F., Akkiraju, P. C., Ponzoni, I., Echenique, V. C., & Carballido, J. A. (2013). Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. *Computers and Electronics in Agriculture, 96*, 173–179.

[30] Silva Jr, E. M., Maia, R. D., & Cabacinha, C. D. (2018). Bee-inspired RBF network for volume estimation of individual trees. *Computers and Electronics in Agriculture, 152*, 401–408.

[31] Toscano, P., Gioli, B., Genesio, L., Vaccari, F. P., Miglietta, F., Zaldei, A., Crisci, A., Ferrari, E., Bertuzzi, F., La Cava, P.,Ronchi, C., Silvestri, M., Peressotti, A., & Porter, J. R. (2014). Durum wheat quality prediction in Mediterranean environments: From local to regional scale. *European Journal of Agronomy, 61*, 1–9.

[32] Tu, J. (2003). Cross-validated Multivariate Metamodeling Methods for Physics-based Computer Simulations. Proceedings of the IMAC-XXI. In *Proceedings of the IMAC-XXI: A Conference on Structural Dynamics*, Kissimmee, Florida.

[33] Tu, J., & Jones, D.R. (2003). Variable Screening in metamodel design by cross-validated moving least squares method. In *44th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, Norfolk, Virginia.

[34] William, J. R., Jones, C. A., Kiniry, J. R., & Spanel, D. A. (1989). The EPIC crop growth model. *Transactions of the ASAE, 32*, 497–511.

[35] Xiao, G., Zhang, Q., Yao, Y., Zhao, H., Wang, R., Bai, H., & Zhang, F. (2008). Impact ofrecent climatic change on the yield of winter wheat at low and high altitudes insemi-arid northwestern China. *Agriculture, Ecosystems & Environment, 127*, 37–42.

[36] Zilinskas, A. (2010). On similarities between two models of global optimization: statistical models and radial basis functions. *Journal of Global Optimization, 48*, 173–182.