Authors response

Title: Using the leave-two-out method to determine the optimal statistical crop model
Author(s): Thi Lan Anh Dinh and Filipe Aires
MS No.: gmd-2021-218
MS type: Methods for assessment of models

Dear referees,

We thank the three referees for their highly constructive comments and suggestions. Our responses (in blue) are listed below. The updated manuscript (i.e., tracked changes version) is also included underneath our responses.

We hope that our corrections will make the manuscript clearer and pleasing to the referees and other readers.

Thanks & best regards,
Lan Anh

Referee comment 1 :

The authors show the value of a two out of sample approach for statistical modelling. The formal approach is convincingly motivated and presented.
Thank you for your appreciation.

The substance of the modelling, however, is completely inadequately presented. One would expect a brief characteristic of the modelled crops and their cropping specifics (annual, perennial; sensitive phases for weather dependence), a characterisation of the climate at the site considered, information about the crop relevant weather variability in the periods considered, a motivation of the predictors, and the margins for determining the predictors, a discussion of model errors.
We agree that the problem should be presented in a better way. To improve the manuscript, we added more information in Sect.2.1 about the characteristics of the model, crop information, characteristics of study areas (climate, agricultural practices), and crop-weather relation. These factors had been already considered, but we did not show them in the manuscript, so we added them in the updated version. We believe that the substance of the modelling is now adequately introduced.

The superficiality of the problem analysis becomes particularly clear in the classification of the model quality:
'It was shown that monthly mean precipitation and temperature could explain more than 30 % of the coffee yield anomaly variability. The 70 % remaining variability is due to non-climatic factors (agricultural practices, diseases, or political and social context). '
This conclusion presupposed that the authors had fully explained the weather-related yield variability with their approaches. However, this cannot be assumed given the selection of climate variables, their coarse resolution, the static view on the underlying processes, the negligence of their interplay, the linearity restrictions applied, the exclusion of spillover effects from previous seasons and of soil characteristics.
We agree with the referee that the (fundamental) problem is much more complex than our weather-yield impact model. In reality, many factors (e.g., soil characteristics, the spillover effects from previous seasons, or extreme events) could affect the yield. For instance, the extremes during the growing season can largely influence the yield of that year (Beillouin et al., 2020; Mathieu and Aires 2018; Vogel et al., 2019). However, in practice, it is not possible for a statistical model to take into account such complexity. In fact, the crop database is, most of the time, very limited in time (often about 10 to 20 years). This means that there are not enough samples to calibrate a very complex statistical model with many input factors and a description of their interaction.

Actually, the main objective of this paper is to introduce the leave-two-out technique that measures, in a robust way, the true capacity of a statistical crop model. For the two studied crops,

this technique shows us that we simply cannot introduce more input parameters in the statistical model; this would be misleading and wrong.

You are right; it is a bit misleading when we say that weather can explain more than 30% of the yield anomaly variance, it could explain more than that if enough samples were available for the calibration of the impact model. We mean that considering the available historical yield record, we can only set up a statistical model that can explain 30 %. This is a lower estimate, and climate could explain more than that, but we would require many more samples to go into deeper details of the plant physiology. Please note that we are extremely rigorous in our statistical modelling practice (this is why we introduced this leave-two-out method) and that many other studies are not so rigorous and artificially claim they can explain a more significant part of the variance. We were probably not clear enough in the first version of the paper and we hope that the new version will clarify the overall meaning and strategy of our analysis.

Concerning the resolution of climate data, we believe that the 0.1° x 0.1° resolution data (~10 x 10 km in the Equator) should be considered as an adequate input for this type of statistical model over different administrative levels (i.e., district level and department level). Please note that monthly-mean values are smooth in nature and that increasing resolution would not necessarily offer more details. Furthemore, the considered districts in Vietnam cover an area of about 822 km^2 and the average area of a department is about 6 000 km^2. Also, the data are compatible with what can be obtained from climate models (e.g., CMIP6 (Eyring et al., 2016)) and thus are adapted to the climate change impact study that we want to perform later on.

Nevertheless, the presented approach shows which formal possibilities exist to obtain a 'best' model without an in-depth examination of the object to be modelled. This is quite interesting for practical statistical modelling. However, the examples presented here (maize, coffee) should be motivated much more comprehensively and classified in a more differentiated way. This does not mean that the authors have to make a comprehensive modelling claim. But they should be able to justify and classify the range of the model configurations considered in their modelling attempt.
As we mentioned above, we added descriptions about the details of the two considering crops and models. Once again, we want to highlight the point that data scarcity is the main problem of statistical models for such applications. You are right, we should introduce better that potentially, a more complex model could be considered. However due to the limited amount of data available to constrain such a model, we have no alternative than to use a simpler one. This is very clear in the abstract and the introduction.

The general quality of the manuscript is currently at the border between ‘fair’ and ‘poor’, but has potential for moving to the better direction when the authors motivate more explicitly their modelling approach for the selected climates and crops.

We were not clear enough about the general context of this modelling, and we hope that our corrections improve the manuscript clarity.

We changed the title: “Using the **nested** leave-two-out **cross-validation** method to determine the **optimal complexity** of the statistical crop model” to be more explicit about the target of the study. This is not to optimise a very sophisticated model that would explain all the crop physiology but instead to find the model's optimal complexity for a very short time-record database.

Although the term “leave-two-out”/ LTO will be used throughout the manuscript for simplicity purposes, we mentioned the “**nested** leave-two-out **cross-validation**” in the title to avoid confusion for the reader at first glance.

References:

Beillouin D, Schauberger B, Bastos A, Ciais P, Makowski D. Impact of extreme weather conditions on European crop production in 2018. Philos Trans R Soc Lond B Biol Sci. 2020 Oct 26;375(1810):20190510. doi: 10.1098/rstb.2019.0510. Epub 2020 Sep 7. PMID: 32892735; PMCID: PMC7485097.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development, 9(5):1937–1958.

Mathieu, J. A., & Aires, F. (2018). Using neural network classifier approach for statistically forecasting extreme corn yield losses in Eastern United States. Earth and Space Science, 5, 622–639. https://doi.org/10.1029/2017EA000343

Vogel, E., Donat, M. G., Alexander, L. V., Meinshausen, M., Ray, D. K., Karoly, D., Meinshausen, N., & Frieler, K. (2019). The effects of climate extremes on global agricultural yields. Environmental Research Letters, 14(5), [054010]. https://doi.org/10.1088/1748-9326/ab154b

Referee comment 2 :

General comments

This paper addresses the very important question of model complexity. How can one best choose the level of model complexity when the objective is to minimize error of prediction for out of sample cases. The paper presents two practical cases (prediction of coffee yield in Viet Nam, prediction of maize yield in France) and two prediction methods (linear regression, artificial neural network). The main objectives of the paper are to explain the leave two out cross validation approach (LTO) as a method of evaluating and choosing between models of different levels of complexity, and to compare it with the leave one out cross validation approach (LOO).

We thank you a lot for your constructive comments.

The presentation of LTO is useful and interesting, but I have some issues with the way the linear regression examples are formulated. (I do not comment on the use of neural networks, with which I am not very familiar).

Our response to your specific comments and technical corrections are listed below.

Specific comments

The authors consider models with a fixed number of explanatory variables and take as the measure of model complexity, the number of potential explanatory variables from which the explanatory variables in the model are chosen. This is not usually the way the problem is formulated, at least in linear regression. In general, the list of potential explanatory variables is fixed, and the question is how many and which to include in the final model. Then the larger the number of explanatory variables chosen, the more complex the model. Comparing LTO and LOO as a function of the number of potential explanatory variables may then not be very relevant to the problem of determining model complexity for a linear model. A more relevant question would be: How do LTO and LOO compare, when the number of potential explanatory variables is fixed, and the objective is to choose the best ones to include in the regression model?

You are right, establishing the optimal number of inputs in a fixed list of potential predictors is often the problem a modeller is facing. We actually consider this case in Sect.4.1: when the number of potential predictors is fixed to 30 and then we increase the number of inputs from 2 to 7 (shown in Fig.6(b)).

As the main target of the paper is to determine the optimal complexity of the statistical crop model, we want to investigate different factors that control model complexity. Model complexity is a wide field of research (e.g., Vapnik–Chervonenkis dimension), in this manuscript we consider several parameters controlling this complexity: the number of potential predictors, the number of inputs, and the two model types (LIN versus NN).

The question of how long should be the list of potential predictors is also a true concern for the modeller: The crop experts have the tendency to increase this list a lot because many parameters can have an impact on the crop. Furthermore, the modeller needs to choose the temporal resolution of the predictors (we used monthly, but weekly or daily values could have been chosen. In Mathieu and Aires (2018), a list of six climate indices {T_{Jul}, T_{Aug}, P_{Jul}, T_{Jun}, T_{May}, P_{Aug}} have been proposed by many experts to perform statistical crop yield modelling. When doing a "plan of experience", the way statisticians define the list of potential predictors is therefore a strong decision, and we show here that it has a significant impact on the model performance. For instance, Fig.5(b1) shows that selecting three inputs out of 38 potential predictors gives a much higher LTO RMSE testing error than selecting three inputs out of 18 potential predictors.

In general, more inputs make the model more complex. Our study shows that the larger number of potential predictors also defines the model complexity. Thus, the number of potential predictors should be considered as an essential factor when doing any statistical analysis. But any parameter controlling the model complexity should be considered: It is actually the goal of this study to propose a tool to optimise these factors.

We added more information in the manuscript (e.g., Sect. 2.3 and Sect. 4.1) to better explain our performed experiences and our main targets. We hope that these changes can accommodate your concerns.

In addition, though LOO can be used for model selection (choosing best explanatory variables) in linear regression, there are other approaches which are probably more common, such as forward regression, stepwise regression, the Akaike Information Criterion etc. How does LTO compare with other methods of model selection?

Yes, there are other common approaches for model selection. We actually used the forward selection (Sect. 2.3.1). These tools are not incompatible with the LTO. It is actually necessary to have a way to test for model errors in the forward selection: LOO is often used (Mark and Goldberg, 2001) for that purpose. Here our main message is that we must use LTO instead of LOO to determine the optimal model, including the choice of forward regression.

Reference:

Mark, Jonathan, & Goldberg, Michael A. (2001). Multiple regression analysis and mass assessment: A review of the issues. The Appraisal Journal, Jan., 89-109.

Mathieu, J.A., Aires, F., 2018. Assessment of the agro-climatic indices to improve crop yield forecasting. Agric. For. Meteorol. 253-254, 15–30.

Technical corrections

L138 Problem with English
We changed the sentence to:
"This leads to the overfitting (or overtraining) problem, i.e., the model fits the training dataset artificially well but it cannot predict well data not present in the training dataset. Thus, using this type of model is not reliable."

L145 Should be "to choose"
Yes. We corrected it.

L163-164 I don't understand the sentence
We changed the sentence to:
"It will be seen in the following that using only the testing dataset instead of the testing and validation datasets can be misleading."

Section 3, introduction. It seems to me that much of this is said elsewhere. Maybe combine with section 2.3?
Thanks for the comment. We shortened the introduction of Sect. 3 and added some ideas into Sect. 2.3.

Fig 3. I find the portrayal of distribution functions confusing. First of all, in the figure the distribution functions are continuous, while in practice they are necessarily discrete. Secondly, as far as I can tell, the distribution functions are totally irrelevant. Only the average error is of interest.
Thanks for the comment. Here, the distribution functions intend to better illustrate the full process of the folding scheme: the validation and testing scores are first obtained for all folds of id_{test}; these scores give the blue and red distributions; after that, the final average scores are computed from these two distributions.
The PDF functions can also give information about the uncertainties of the final score. For instance, if the spread of the blue distribution is small, it means that our estimation of the validation score is characterised by a low uncertainty.
Also, we think there is no need to have detailed discrete values in a scheme like this. We synthesise them as continuous values.
Therefore, we would like to keep this figure as it is.

Section 3.2.2. Talking of a "testing dataset" is a bit confusing (is this the full set of testing data, which is identical to the available data, or is this a single value for each fold). Perhaps refer to the "testing value" or "testing datum" when talking about a single fold.
We used the "testing value" to avoid confusion.

L266 “implemented” is probably the wrong word.
We changed it to “exploited”

L362 “request” is probably the wrong word
It should be “prefer”, but we removed this sentence due to the repetition of the previous sentence (at the beginning of this paragraph).

L366 I don’t understand this sentence.
We made changes for this sentence and the following sentences:
“... .The first one concerns the coffee yield modelling over a major Robusta coffee-producing district in Vietnam. It was shown that considering the available historical yield record, we can only set up a statistical model that explains about 30 % of the coffee yield anomaly variability.”

Referee comment 3 :

The authors discuss an important aspect of statistical analysis with limited data availability and exemplify the shortcomings of the more frequently used ‘leave one out’ cross validation compared to the ‘leave two out’. Two case studies are used, an annual crop in Europe and a perennial crop in Vietnam. Two different statistical approaches are used, linear regression and neural network. The two case studies convincingly demonstrate how to select the best model for forecasting crop yield with limited data availability, while considering the problem of overfitting. In general the study is well designed and clearly communicated, although substantial English revision is required.

Thank you very much for your appreciation on this work. The manuscript has been improved. We hope that this updated version accommodates all your concerns.

Some general observations:

- Line 132: the model architecture is explained as consisting of i) number of potential predictors, and ii) the number of inputs. Please specify clearer what the difference is between predictors and inputs. The authors also mention model types: Some models require more parameters to estimate even with equal numbers of predictors. This should be described in clearer terms.

  We added more explanations in this section, i.e., Sect. 2.3.1:

  “Various factors control the complexity level of a statistical model: the model architecture (the number of potential predictors, the number of inputs, the number of parameters or the model types (e.g., linear or non-linear)) or the training process (e.g., the number of epochs in NN or the loss function).In theory, it is challenging to define the exact definition of a model complexity: even the number of parameters in the models is only a proxy because a model with a low number of parameters can be highly complex, e.g., Vapnik–Chervonenkis dimension (Hastie et al., 2009). This study thus investigates some of the factors that control part of the model complexity.

  Number of inputs

  The inputs are variables that are necessary for model execution through algorithms. The inputs are selected among the potential predictors. We often have a big set of potential predictors (e.g., all-weather variables during the crop growing season), but we select only some variables from this set as the model inputs. The number of inputs defines the model complexity: the higher the number of inputs is, the more complex the model is (supposed that other factors are fixed).

  Number of potential predictors

  The potential predictors here refer to all possible variables that can potentially impact the yield. Our study considers 38 weather variables for Robusta coffee and 12 variables for grain maize (Sect. 2.1), but these numbers could be much larger. For instance, in addition to selected weather variables, we could consider other variables (e.g., water deficit, soil moisture), agro-climatic

indices (e.g., degree-days, free frost period (Mathieu and Aires, 2018b)). Here, we use monthly variables, but weekly or daily variables could have been considered. Therefore, establishing the list of potential predictors is not fixed: it is a crucial modelling step. The following sections (Sect. 4.1 and 5.1) will show that the number of potential predictors drives the model complexity: having too many potential predictors is dangerous, in particular, if the tools are not right.

Model types

Model complexity can be shown in two model types that we presented in Sec. 2.2. For example, with n_{input} inputs, a simple LIN model requires (n_{input} + 1) parameters (Eq. (2)), while a feedforward NN model with one hidden layer and one output requires much more parameters: ($n_{input} \times n_{neuron} + n_{neuron}$) + n_{neuron} + 1, where n neuron is the number of neurons in the hidden layer."

Reference:

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference and prediction, Springer, http://www-stat.stanford.edu/~tibs/ElemStatLearn/, 2009.

- The value of using LIN3 and LIN5 are not clear to me, which is related to the previous comment: the authors explain them as linear regression models with three and five inputs, respectively. Could you explain what the value is of this experiment? What added insights do we gain when comparing LIN3 and LIN5? Do we not already get all the insights from adding the number of potential predictors to LIN3?

  In this part (Sect.4.1), we would like to test different model complexity levels by changing several factors: the number of potential predictors, the number of inputs, and the two model types (LIN, NN). Various models have been tested to resolve the problems of the number of potential predictors, but only LIN3 and LIN5 are shown (together with the LIN models in Fig.6(b)). Also, LIN5 is chosen here as it illustrates very well the robustness of the LTO method compared to the LOO: the LOO training/testing and LTO training/validation RMSE values follow an evident decreasing tendency, while this tendency is not so clear in the LIN3 example. On the other hand, LIN3 and LIN5 also are examples of model complexity defined by the number of inputs. With a more complex model (LIN5), the observed behaviour of training/validation/testing RMSE values is much stronger than a simpler model (LIN3).

  We added some changes in this section to better explain the problem.

- There is quite a bit of redundancy – the authors explain the problem several times, thereby repeating themselves. For example, section 3 (lines 168-181) already has been elaborated in previous sections. More concise presentation of the problem and methods would benefit the manuscript.

Yes. We totally agree about this. We shortened some parts of the manuscript and combined them into other sections. You can trace these changes in the tracked changes manuscript.

- Even though the main aim of the study is to compare the LOO and LTO, presenting the chosen predictors and final model would be appreciated. Any reader familiar with the case study crops will be interested to understand what climate descriptors were tested and selected in the final model. This transparency is further necessary for making the study reproducible.

  Thanks for this useful suggestion. We added more information about the chosen predictors and final model (Sect. 4.2, Sect. 5.3).

  For instance, in Sect. 4.2:

  "The final model includes {$P_{Nov(t-1)}$, $P_{Nov(t)}$, $T_{Mar(t)}$} and these selected variables coincide with the key moments of Robusta coffee. For example, there is the need for a dry period for the buds to develop into dormancy at the end of the development stage, i.e., Nov(t-1). Therefore, $P_{Nov(t-1)}$ impacts directly the buds, thus the potential yield. Similarly, the fruit maturation stage (Nov(t)) benefits from weather conditions with less precipitation. At the beginning of the fruit development period (Mar(t)), too low temperature slows maturation rate to the detriment of yield, while the higher temperature is beneficial."

  In Sect. 5.3.:

  "... In other words, the final model should consider only variables from April to August. As in our case, statistical model selects {T_{Jul}, P_{May}, P_{Apr}} as the final inputs for grain maize in the eatern region (Bas-Rhin, Haut-Rhin); {T_{Jul}, P_{Jul}, T_{Apr}} for the southern region (e.g., Landes, Pyrénées-Atlantiques, Gers); and {P_{Jul}, P_{Apr}, P_{Jun} or T_{Jun}} for central part (Vendée, Charente-Maritime, Vienne). It is reasonable to have different inputs for different regions (or even departments) due to their distinct environmental conditions. In general, weather variables in July—the flowering period—are among the most influential variables. During this time, a high temperature affects the photosynthesis process, thus reducing the potential yield; in contrast, positive precipitation anomalies are preferable (Ceglar et al., 2016; Mathieu and Aires, 2018b). Precipitations in April and May also show significant impacts on grain maize as a water deficit during this vegetative stage decreases plant height (Çakir, 2004)."

Using the nested leave-two-out cross-validation method to determine the optimal complexity of the statistical crop model

Thi Lan Anh Dinh[1] and Filipe Aires[1]

[1]Sorbonne Université, Observatoire de Paris, Université PSL, CNRS, LERMA, 75014 Paris, France

Correspondence: Thi Lan Anh Dinh (lan-anh.dinh@obspm.fr)

Abstract. The use of statistical models to study the impact of weather on crop yield has not ceased to increase. Unfortunately, this type of application is characterised by datasets with a very limited number of samples (typically one sample per year). In general, statistical inference uses three datasets: the training dataset to optimise the model parameters, the validation datasets to select the best model, and the testing dataset to evaluate the model generalisation ability. Splitting the overall database into three datasets is impossible in crop yield modelling. The leave-one-out cross-validation method or simply ~~leave-one-out~~ Leave-One-Out (LOO) has been introduced to facilitate statistical modelling when the database is limited. However, the model choice is made using only the testing dataset, which can be misleading by favouring unnecessarily complex models. The nested cross-validation approach was introduced in machine learning to avoid this problem by truly utilising three datasets~~, especially problems~~ even with limited databases. In this study, we propose~~d~~ one particular implementation of the nested cross-validation, called the nested leave-two-out cross-validation method or simply the Leave-Two-Out (LTO), to choose the best model with an optimal model complexity (using the validation dataset) and estimate the true model quality (using the testing dataset). Two applications are considered: Robusta coffee in Cu M'gar (Dak Lak, Vietnam) and grain maize over 96 French departments. In both cases, LOO is misleading by choosing too complex models; LTO indicates that simpler models actually perform better when a reliable generalisation test is considered. The simple models obtained using the LTO approach have reasonable yield anomaly forecasting skills in both study crops. This LTO approach can also be used in seasonal forecasting applications. We suggest that the LTO method should become a standard procedure for statistical crop modelling.

1 Introduction

Many approaches are available to study the impact of climate/weather variables on crop yield. Statistical modelling, which aims to find relations between a set of explanatory variables and crop yields variability, is a widely used approach (Lobell and Burke, 2010; Mathieu and Aires, 2016; Gornott and Wechsung, 2016; Kern et al., 2018). This approach ~~with various forms~~ has many advantages, such as identifying crop production ~~constraints~~ sensitivities (Mathieu and Aires, 2018a), complementing field experiments (Gaudio et al., 2019), and helping in adaptation strategies (Iizumi et al., 2013), but it is often complex to understand and to use for several reasons.

Unfortunately, the crop model is often characterised by datasets with a very limited number of samples. For instance, Prasad et al. (2006) built a crop yield estimation model with 19 years of ~~crop~~ yield data. ~~and~~ Ceglar et al. (2016) studied the impact

of meteorological drivers over 26 years on grain maize and winter wheat yield in France. One year of data represents one sample in these applications, and about 20 samples are small for a data-driven approach. The small sample size issue makes model selection very challenging. It is not easy to assess the true quality of an obtained model. ~~The small sample size issue makes model selection very challenging.~~ In principle, increasing the model complexity ~~can~~ should increase the model quality. However, it can lead to "overfitting" if the model is too complex ~~considering the~~ and if we have a limited information included in the database. Overfitting occurs when the model fits the training dataset artificially well, but it cannot predict well on unseen data. To avoid this issue, in statistical modelling, ~~In statistical models,~~ the overall database is divided into three datasets: the training dataset to optimise the model parameters, the validation datasets to select the best model, and the testing dataset to evaluate the model generalisation ability (Ripley, 1996). Splitting a small number of samples into three datasets is not an easy task.

Cross-validation (Allen, 1974; Stone, 1974) was introduced as an effective method for both model assessment and model selection when the data is relatively small. A common type of cross-validation is the ~~leave-one-out~~ Leave-One-Out cross-validation (LOO) that has been used in many crop models (Kogan et al., 2013; Zhao et al., 2018; Li et al., 2019; Dinh et al., 2021) ~~(Kogan et al., 2013; Zhao et al., 2018; Dinh et al., 2021)~~. This approach relies on two datasets: a training dataset is used to calibrate the model, and a testing dataset is used to assess its quality. The testing is also used to select the best model, which is not a good practice and introduces difficulties~~, as seen in the following~~. ~~Thus,~~ Since the chosen model is not independent of the testing dataset, ~~and~~ the obtained testing score may be unreliable. This is not a problem if there are many available samples (e.g., in remote sensing applications).~~, but~~ However, a small sample size can cause many issues: the model can overfit the training dataset; thus, the complexity of the chosen model is not adequate, and our assessment of its generalisation ability is false. ~~This is often a mistake in crop yield modelling that uses over-complex models that cannot be calibrated with a limited number of samples.~~ This mistake is often seen in crop modelling when over-complex models are used with a limited number of samples. Some regularisation techniques (e.g., information content techniques or ~~dimesion~~ dimension reduction techniques) can help to constrain ~~constraint~~ models toward lower complexity to limit the overfitting problem (Dinh et al., 2021). However, these approaches can become more technical and more challenging to master from non-statisticians.

To solve the issues of LOO, another more complex approach has been introduced: the nested cross-validation (Stone, 1974), also known as double cross-validation or k*l-fold cross-validation, is able to use three datasets: training, validation, and testing. In ~~details~~ detail, this approach considers one inner loop cross-validation nested in an outer cross-validation loop. The inner loop is to select the best model (validation dataset), while the outer loop is to estimate its generalisation score (testing dataset). To our knowledge, this ~~This~~ approach has~~, however,~~ never been used in statistical crop modelling. This study proposes one particular implementation of this nested cross-validation (or k*l-fold cross-validation when l=k-1) called the ~~leave-two-out~~ Leave-Two-Out (LTO). The LTO will be used here to obtain a reliable assessment of the model generalisation ability, to compare the performances of different predictive models, and thus to determine the optimal complexity of the statistical crop models. This approach will be tested in two real-world applications: Robusta coffee in Cu M'gar (a district of Dak Lak province in Vietnam) from 2000 to 2018 and grain maize over 96 departments (i.e., administrative units) in France for the 1989-2010 period. The following sections of this study will (1) introduce the materials and databases used for statistical crop models, (2) describe the

role of three datasets in statistical inference, (3) introduce the two cross-validation approaches, (4) evaluate and select the "best model" by using LOO and LTO approaches, (5) estimate the Robusta coffee yield anomalies in Cu M'gar (Dak Lak, Vietnam), and (6) assess the seasonal yield anomaly forecasts for grain maize in France.

2 Modeling crop yield using machine learning

2.1 ~~Databases~~ Materials

2.1.1 ~~Coffee yield database~~ Robusta coffee

Overview

Robusta (Coffea canephora) is among the two most common coffee species (i.e., Robusta and Arabica). About 40 % of Robusta coffee is produced in the Central Highlands of Vietnam (USDA, 2019; FAO, 2019) due to its adequate conditions in terms of elevation (200-1500 m), soil type (basalt soil), and climate (an annual average temperature of about 22 °C). In addition, agricultural practices (e.g., fertilisation, irrigation, shade management, and pruning) are very intense in these coffee farms (Amarasinghe et al., 2015; Kath et al., 2020). This region includes four main coffee-producing provinces, and each province is divided into several districts. Here, we focus on Robusta coffee in Cu M'gar, one major coffee-producing district in the Central Highlands.

A coffee tree is a perennial, which is highly productive for about 30 years (Wintgens, 2004) but can be much longer (more than 50 years) with good management practices. Mature coffee trees undergo several stages before harvesting, including the vegetative stage (bud development) and the productive stage (flowering, fruit development, and maturation) (Dinh et al., 2021). It requires about eight months (May to December) for the vegetative stage and about 9-11 months (January to September/November) from flowering until fruit ripening for Robusta coffee. Although climate during the productive stage is sensitive to coffee (Craparo et al., 2015b; Kath et al., 2020), it has been shown that a prolonged rainy season favours vegetative growth and thus increases the potential coffee yield (Dinh et al., 2021). As a result, it is necessary to consider the weather variables during both vegetative and productive stages when studying the weather impact on coffee yield. This study thus analyses the weather of 19 months (from May of the previous year to November) preceding the harvest.

Yield database

The Robusta coffee yield data were obtained from the General Statistics Office of Vietnam for the 2000-2018 period ($n_{samp} = 19$). ~~The data are available at the district and provincial levels. Here, we focus on Robusta coffee in Cu M'gar,~~ We focus on Cu M'gar district as it is one major coffee-producing district ~~of Dak Lak (Vietnam), as~~ in Vietnam, but also this district is most sensitive to weather (Dinh et al., 2021). Our goal is to forecast the weather sensitivity of crop yield.

The long-term trend represents the slow evolution of the crop yield; it often describes the changes in management like fertilisation or irrigation. Thus, suppressing this trend from the yield time series allows removing the influence of non-weather related factors and this is the common practice. For Robusta coffee, a simple linear function is used to define the yield trend: $\overline{y}(t) = y_0 + \alpha \cdot t$, where $\overline{y(t)}$ is the long-term trend, y_0 is the yield in 2000, and α is the constant annual rate of improvement.

Once the yield trend is defined, the coffee yield anomalies are calculated by removing this trend from the raw yield data. The Robusta coffee yield for year t is noted as $y(t)$~~, the long-tem trend value as $\overline{y(t)}$,~~ and the coffee yield anomaly $a(t)$ (in %) is calculated as:

$$a(t) = \frac{y(t) - \overline{y(t)}}{\overline{y(t)}} \times 100 \in [-100, 100]. \quad (1)$$

If $a(t) > 0$, then the yield in year t is higher than in a regular year, and vice versa. For ~~instance~~ example, an anomaly of $a(t) = -16$ ~~means~~ implies that the yield for year t is 16 % lower than the annual trend.

2.1.2 Grain maize ~~yield database~~

Overview

Grain maize (Zea mays L.) is among the most common annual crops in Europe. Our study will focus on French regions—the leading grain maize producer in Europe (EUROSTAT, 2021). The study area has been improved a lot in agro-management and irrigation practices after 1960, e.g., irrigation acres was about 50 % at the beginning of the 21^{st} century (Siebert et al., 2015; Schauberger et al., 2018; Ceglar et al., 2020). Although there is a change in time from sowing to maturity (Olesen et al., 2012), the average growing season of French grain maize ranges from April to September (Ceglar et al., 2017; Agri4cast, 2021). Many previous studies showed that grain maize yield is sensitive to weather conditions (Ceglar et al., 2016, 2017; Lecerf et al., 2019), especially during crop growing season. Therefore, we will analyse the weather of the 6-month growing period of grain maize in France.

Yield database

The French crop data (area, production, and yield) on the regional level (i.e., department which is an administrative unit in France) were collected from Agreste website (https://agreste.agriculture.gouv.fr; "Statistique agricole annuelle") for a period of 22 years (from 1989 to 2010). The data are available for several crops such as soft wheat, durum wheat, maize, oats, etc.~~This study considers an application for~~ Here we consider an application of grain maize over 96 French departments (Fig. 1). Some specific tests (in Sect. 5) will focus on ten departments (as presented in Fig. 1(d)) where the average grain maize production is higher than 4×10^5 tons (or the area is higher than 40 thousand hectares). Other available crop data will be considered in future studies.

Similar to ~~the~~ Robusta coffee ~~ease~~, the grain maize anomalies are calculated by removing the long-term yield trend. Here, a 10-year moving average window is used because the trend is slightly more complex than ~~for~~ Robusta coffee.

2.1.3 Weather database

The monthly-mean total Precipitation (P) and 2 m Temperature (T) variables were collected for the period 1981-2018 from the ERA5-Land, i.e., a replay of the land component of ERA5 re-analysis of the European Center for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al., 2018). This database is at a spatial resolution of $0.1° \times 0.1°$ (about 10 km $\times$ 10 km in the Equator). The monthly data are then projected from its original $0.1° \times 0.1°$ regular grid into the crop administrative levels to match the yield data.

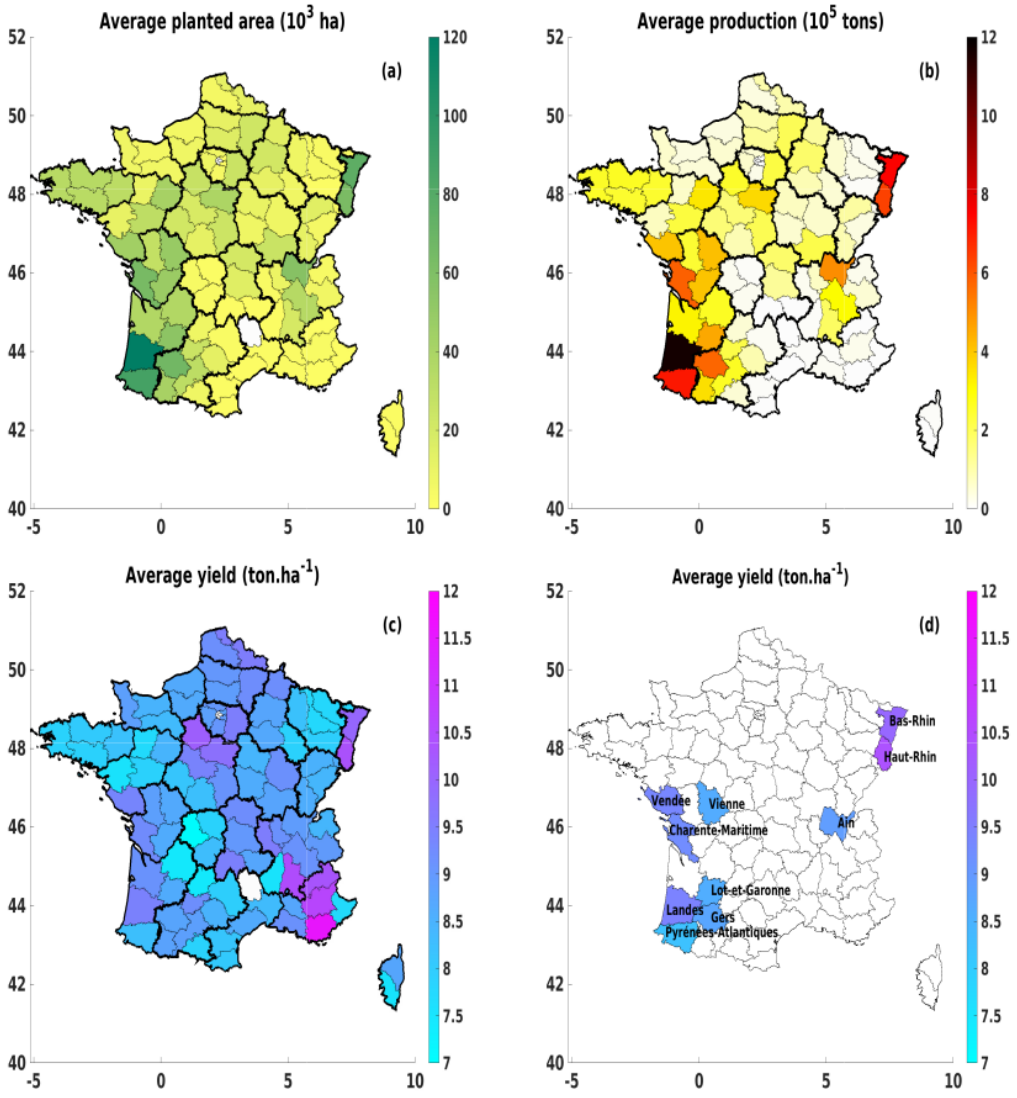

Figure 1. Grain maize database: (a) the average planted area (in 10^3 ha), (b) the average production (in 10^5 tons), (c) the average yield (in ton·ha^{-1}) over 96 French departments; (d) same as (c) but presenting over only ten major grain maize-producing departments. All data are averaged from 2000-2010.

125 ~~In this study,~~ This study considers the $2 \times n$ monthly weather anomaly variables (representing P and T for n months) ~~are considered~~. The number of months n varies for each crop:

- For Robusta coffee: we evaluated n=19 corresponding to the period from the bud development process to the harvest season's peak ~~(Dinh et al., 2021)~~ (Sect. 2.1.1). Thus, 2×19 monthly weather data (P and T from May of year $(t-1)$

to November of year t: $P_{May(t-1)}$, $\cdots$, $P_{Dec(t-1)}$, $P_{Jan(t)}$, $\cdots$, $P_{Nov(t)}$ and $T_{May(t-1)}$, $\cdots$, $T_{Dec(t-1)}$, $T_{Jan(t)}$, $\cdots$, $T_{Nov(t)}$) are used as potential explanatory variables for Robusta coffee yield anomalies.

- For grain maize: six months of growing period (from sowing to harvest) will be studied (Sect. 2.1.2). Thus, $n = 6$ results into 2×6 weather variables: P and T from April to September (P_{Apr}, P_{May}, $\cdots$, P_{Sep} and T_{Apr}, T_{May}, $\cdots$, T_{Sep}).

Weather anomalies could be considered as for crop yield data. However, the climate trend of the 10 to 20 years is relatively low compared to the inter-annual variations. Thus, the long-term trend can be neglected, and the relative anomalies will be estimated based on the long-term average. This average value is computed for each of the n months before the harvest time. In addition, we applied a 3-month moving average centred on the particular month (instead of the monthly data) to reduce the variability at the monthly scale. This ~~since this~~ variability would introduce instabilities in our analysis due to the short database time length. (It ~~This~~ is actually a regularisation technique).

We also analysed other weather variables (not shown) , e.g., maximum/minimum temperature, solar radiation. However, these variables were finally excluded due to several reasons: (1) These variables show relatively low correlations to the crop yield anomalies. (2) Or they are highly correlated to P and T variables, especially for the case of Robusta coffee. (3) It will be seen in the following that considering the available yield database size, it is more reasonable to consider a limited number of explanatory variables to avoid overfitting (see more in Sec. 2.3).

2.2 Statistical yield models

The statistical models intend to measure the impact of weather on crop yield anomalies, which can be noted as: $a(t) = f_w(X)$, where ~~$a(t)$ is the crop yield anomaly for year t,~~ f_w is the parametric statistical model, w is the model parameters, and X is the set of weather inputs $\{X_i \text{ for } i = 1,\ 2, \cdots,\ n_{input}\}$. The function f_w can be based on multiple statistical methods depending on the complexity of the application:~~, for instance,~~ linear regression (Prasad et al., 2006; Kern et al., 2018; Lecerf et al., 2019), partial least-squares regression (Ceglar et al., 2016), random forest (Beillouin et al., 2020), neural network (Mathieu and Aires, 2016, 2018a), or mixed-effects (Mathieu and Aires, 2016)~~, etc~~.

In this study, two statistical models are considered:

- Linear regression (LIN) is the simplest model and the most frequently used. The relationship between the crop yield anomalies a ~~$a(t)$~~ and the weather variables X_i ~~X_{ti} ($i = 1,\ 2, \cdots,\ n_{input}$ is the number of input variables)~~ is formulated as:

$$a = \alpha_0 + \alpha_1 \cdot X_1 + \cdots + \alpha_n \cdot X_{n_{input}}, \tag{2}$$

  where α_i are the regression ~~coeffiecents~~ coefficients. Detailed description of the LIN model can be found, for ~~instance~~ example, in Dinh et al. (2021).

- Neural Network (NN) is a non-linear statistical model. The simplest type of NN is the feedforward model (Bishop, 1995; Schmidhuber, 2015), where there is only one direction—forward—from the input nodes, through the hidden nodes and

to the output nodes. Only one hidden layer with n_{neuron} neurons is considered in the architecture here. The output crop yield anomaly a is ~~modeled~~ modelled by the following equation:

$$a = \sum_{j=1}^{n_{neuron}} w_j \times \sigma \left(\sum_{i=1}^{n_{input}} w_{ji} X_i + b_{hidden} \right) + b_{output} \quad (3)$$

where w are the weights, ~~x_i are the weather variables,~~ b are the NN biases. A detailed description of the NN model (applied for impact models) is described, for instance, in Mathieu and Aires (2016).

The least-squares criterion, which measures the discrepancies between the targets and estimated crop yield anomalies, is used to optimise the model during the calibration process for both LIN and NN models. ~~For instance, it~~ It is used to obtain the coefficients α_i in Eq. (2) and the NN parameters w in Eq. (3) during the training stage.

Two diagnostics are considered here to measure the quality of the yield anomaly estimations. (1) ~~One is the~~ The correlation COR (unitless) between the estimated a_{est} and observed a_{obs} yield anomalies. (2) The Root Mean Square Error is defined as: RMSE = $\sqrt{\frac{1}{n_{samp}} \sum_{i=1}^{n_{samp}} (a_{est}(i) - a_{obs}(i))}$. It includes systematic and random errors of the model. The ~~unit of~~ RMSE unit is the same as $a(t)$; RMSE=40 represents ~~a 40% error.~~ an anomaly error of 40 %.

2.3 Model complexity and overfitting

2.3.1 Model complexity

Various factors control the complexity level of a statistical model: the model architecture (the number of potential predictors ~~on which the inputs are chosen~~, the number of inputs, the number of parameters or ~~simply~~ the model types (e.g., linear or non-linear)) or the training process (e.g., the number of epochs in NN or the loss function). In theory, it is challenging to define the exact definition of a model complexity: even the number of parameters in the models is only a proxy because a model with a low number of parameters can be highly complex, e.g., Vapnik–Chervonenkis dimension (Hastie et al., 2009). This study thus investigates some of the factors that control part of the model complexity.

Number of inputs

The inputs are variables that are necessary for model execution through algorithms. The inputs are selected among the potential predictors. We often have a big set of potential predictors (e.g., all-weather variables during the crop growing season), but we select only some variables from this set as the model inputs. The number of inputs defines the model complexity: the higher the number of inputs is, the more complex the model is (supposed that other factors are fixed).

Number of potential predictors

The potential predictors here refer to all possible variables that can potentially impact the yield. Our study considers 38 weather variables for Robusta coffee and 12 variables for grain maize (Sect. 2.1), but these numbers could be much larger. For instance, in addition to selected weather variables, we could consider other variables (e.g., water deficit, soil moisture), agro-climatic indices (e.g., degree-days, free frost period (Mathieu and Aires, 2018b)). Here, we use monthly variables, but weekly or daily variables could have been considered. Therefore, establishing the list of potential predictors is not fixed: it is a crucial modelling

step. The following sections (Sect. 4.1 and 5.1) will show that the number of potential predictors drives the model complexity: having too many potential predictors is dangerous, in particular, if the tools are not right.

Model types

Model complexity can be shown in two model types that we presented in Sec. 2.2. For example, with n_{input} inputs, a simple LIN model requires $(n_{input}+1)$ parameters (Eq. (2)), while a feedforward NN model with one hidden layer and one output requires much more parameters: $(n_{input} \times n_{neuron} + n_{neuron}) + n_{neuron} + 1$, where n_{neuron} is the number of neurons in the hidden layer.

2.3.2 Overfitting

~~In principle, it~~ It is possible to increase the model quality by increasing its complexity because a more complex model can fit better a database. However, such a simple seasoning is dangerous: ~~it is not always the case:~~ the model complexity can be too high compared to the limited information included in the training database. This limitation leads to the overfitting (or overtraining) problem, i.e., the model fits the training dataset artificially well~~,~~ but it cannot predict well data not present in the training dataset~~, meaning that the model is not reliable not to be used.~~ . Thus, using this type of model is not reliable. There is no general rule determining the model complexity based on the number of samples. An empirical tool needs to be used to check the adequacy of the model. In the following, by studying the sensitivity of the model quality to different complexity levels, we want to determine the optimal statistical crop model that truly estimates the yield anomalies as best as possible~~, considering the limited database by avoiding overfitting~~.

2.4 Training, validation and testing datasets

One of the main challenges in statistical inference is that the model is set up using a samples database, but it must perform well on new~~,—~~previously unseen~~—~~samples. ~~This is a complex task, and methods have been designed to perform good training, chose the suitable model, and measure the model generalisation ability realistically.~~ For that purpose, the overall database $\mathcal{B}$ ~~is~~ needs to be divided into three datasets: $\mathcal{B} = \mathcal{B}_{Train} + \mathcal{B}_{Val} + \mathcal{B}_{Test}$~~, each one of these three datasets undertakes a specific task~~ (Ripley, 1996):

- The **training dataset** $\mathcal{B}_{Train}$ is used to calibrate the model parameters once the model structures has been chosen.
- The **validation dataset** $\mathcal{B}_{Val}$ is a sample of data held back from the training dataset, which is used to find the best model. For instance, it helps tune the model hyper-parameters: choose the more adequate inputs (i.e., feature selection), determine the number of predictors, find the best model type (LIN, random forest, NN), determine some training choices~~,~~ ~~etc~~.
- The **testing dataset** $\mathcal{B}_{Test}$ is held back from the training and the validation datasets to estimate the true model generalisation ability ~~to generalise~~.

The process of partitioning $\mathcal{B}$ will be called in the following as the "folding" process. For ~~instance~~ example, the folding choice can be chosen using $\mathcal{B}_{Train} = 50\ \%$, $\mathcal{B}_{Val} = 25\ \%$, and $\mathcal{B}_{Test} = 25\ \%$. ~~50 %, 25 %, and 25 % of the whole database $\mathcal{B}$ for $\mathcal{B}_{Train}$, $\mathcal{B}_{Val}$, and $\mathcal{B}_{Test}$ respectively.~~

The need for the validation dataset is not always understood. The training dataset is used to fit the parameters; the testing dataset is often used to estimate the model quality but also to choose the best model (as in the LOO approach). However, using only this testing dataset without a validation dataset brings a risk of choosing the model that ~~is~~ best ~~suited~~ suits to this particular testing dataset. This issue is a special kind of overfitting, which is not on the model calibration but on the model choice. If the database is big, many samples in the testing dataset will be representative enough; ~~thus~~ therefore, choosing the best model based on it is acceptable. If the database is small (as in crop modelling tasks), the model selection can be too specific for the particular samples of the testing dataset; thus, an overfitting problem can appear (Sect. 2.3.2). It will be seen in the following that ~~the missing of validation dataset~~ using only the testing dataset instead of the testing and validation datasets can be misleading. We avoid this difficulty by having a dataset to calibrate the model (training) and another one to choose the best model (validation). The truly independent testing dataset is then used to measure the model generalisation ability to process ~~well~~ truly unseen data.

3 Measuring the quality of statistical yield models

~~In practice, statistical yield models often face the problem of "data scarcity" (i.e., having a limited number of crop yield data). One sample corresponds to one year of data; a 19-year database thus provides only 19 samples. This scarcity of data introduces two problems:~~

~~First, it is impossible to infer a complex model from such a limited number of samples. The model complexity needs to be limited by the information provided by the samples. If samples are not enough, there is a considerable risk of overfitting the model towards the limited number of samples of the training dataset. There is no general rule determining the complexity of the model based on the number of samples.~~

~~Second, it is necessary to divide the database into training, validation, and testing datasets (Sect. 2.4). With a limited number of samples, the training process may need every possible data point to determine model parameters adequately (Kuhn and Johnson, 2013), and it might be impossible to choose the best model or assess its generalisation ability with the remaining samples. It is challenging to keep a significant percentage of the database for the validation and the testing datasets.~~

With a limited number of samples, the training process may need every possible data point to determine model parameters (Kuhn and Johnson, 2013). It is thus impossible to keep a significant percentage of the database for the validation and the testing datasets. To choose a model with an adequate complexity level and avoid overfitting, a robust way to measure the generalisation ability is necessary, using as few samples as possible. Cross-validation (Allen, 1974; Stone, 1974) was ~~developed~~ introduced as an effective method for both model selection and model assessment when having a small number of samples.

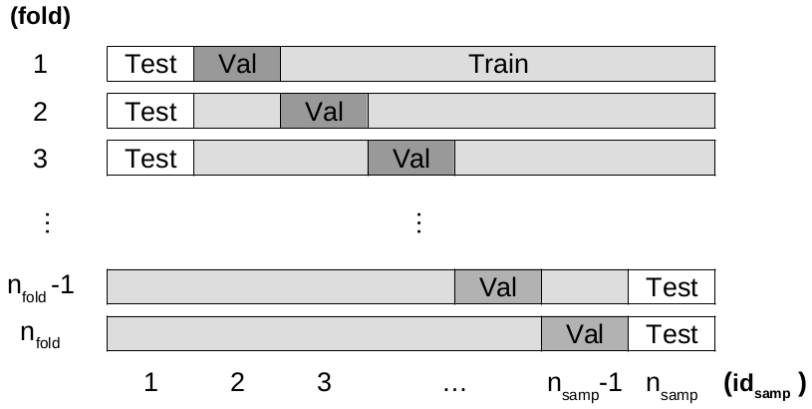

Figure 2. Folding strategy for the LTO procedure with $n_{fold} = n_{samp} \times (n_{samp} - 1)$ folds (~~correponding~~ corresponding to the n_{fold} rows). In each fold, there are one testing, one validation, and $(n_{samp} - 2)$ training samples.

3.1 Traditional Leave-One-Out

The LOO method is one common type of cross-validation~~, in which the model trained, chosen, and evaluated using only two datasets.~~ in which the model uses only two datasets: one to train, another to choose the model and evaluate the result. The main idea of LOO is that given n_{samp} available samples in $\mathcal{B}$; the model is calibrated n_{samp} times using $(n_{samp} - 1)$ samples in the training dataset $\mathcal{B}_{Train}$ (leaving one sample out). The resulting model is then tested ~~$(n_{samp} - 1)$~~ on the left sample ($\mathcal{B}_{Test}$). There ~~is~~ are n_{samp} testing score estimations, one for each sample. In this case, $\mathcal{B} = \mathcal{B}_{Train} + \mathcal{B}_{Test}$ and $\mathcal{B}_{Val}$ is empty. The averaging of these n_{samp} testing scores is expected to be a robust assessment of the model ability to generalise on new samples. However, since no validation dataset is used to select the best model, the choice of the best model ~~is made using the testing dataset; thus, it~~ may be biased towards this testing dataset (Cawley and Talbot, 2010). The chosen model is not independent of the testing dataset, and thus, the obtained testing score is not reliable.

3.2 Proposed Leave-Two-Out

LOO is very useful in many cases (Kogan et al., 2013; Li et al., 2019; Dinh et al., 2021) ~~(Kogan et al., 2013; Dinh et al., 2021),~~ but as described in Sect. 2.4, the overall database needs to be divided into three datasets~~: $\mathcal{B} = \mathcal{B}_{Train} + \mathcal{B}_{Val} + \mathcal{B}_{Test}$~~. A LTO approach, adapted from the nested cross-validation ~~(Stone, 1974),~~ is then proposed in the following.

3.2.1 Folding scheme

A folding process is used to generate the training, validation, and testing scores. Each fold divides the database $\mathcal{B}$ into a training dataset $\mathcal{B}_{Train}$ of $(n_{samp} - 2)$ samples, a validation $\mathcal{B}_{Val}$~~,~~ and a testing $\mathcal{B}_{Test}$ datasets~~,~~ with one sample each. Two samples are considered out of the training dataset instead of one in the LOO procedure. This folding process is presented in Fig. 2, with the number of folds $n_{fold} = n_{samp} \times (n_{samp} - 1)$. This is why this approach is also called k*l-fold cross-validation when l=k-1.

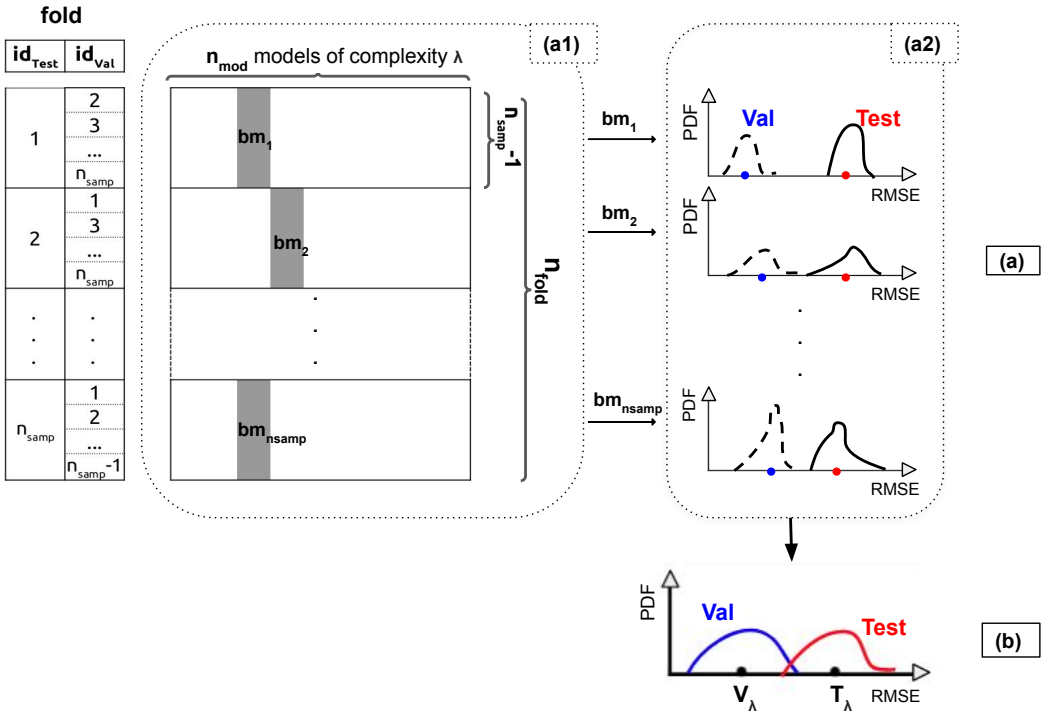

Figure 3. Illustration of the ~~The~~ LTO procedure to estimate a model quality for a fixed complexity level λ with n_{mod} candidate models (horizontal axis). (a) The model errors obtained for each candidate model and each fold of the database $\mathcal{B}$ (vertical axes); (b) The obtained RMSE values for the validation and testing datasets. (See detailed description in Sect. 3.2.)

3.2.2 Validation and testing scores

Figure 3 illustrates how the LTO evaluation procedure is conducted. In part (a1) ~~of this figure~~, the number of candidate models n_{mod} (represented in the horizontal axis) is defined with a fixed complexity λ of the model. For instance, for the LIN3 model (i.e., LIN model with three inputs) with 12 potential predictors, we obtain $n_{mod} = C_{12}^3 = 220$ models. These models are used to perform the yield anomaly estimations. In the vertical axis, for each of the n_{samp} choices of the testing ~~dataset~~ value $id_{test} \in \{1, 2, \cdots, n_{samp}\}$, there are $(n_{samp} - 1)$ possible validation datasets, and thus training datasets. These $(n_{samp} - 1)$ training datasets correspond each to the training of the models in the horizontal axis, ~~(~~i.e., to fit model parameters~~)~~. So $(n_{samp} - 1)$ validation and $(n_{samp} - 1)$ testing estimations are obtained for each one of the n_{mod} models. The averaged validation score is used to choose the best model bm_i for $i = 1, 2, \cdots, n_{samp}$; this is the role of the validation dataset.

Each choice of the testing ~~dataset~~ value (each id_{test}) corresponds to a selected best model bm_i and two distributions (i.e., ~~probability density functions~~ Probability Density Functions (PDFs)) for $(n_{samp} - 1)$ validation errors and $(n_{samp} - 1)$ testing

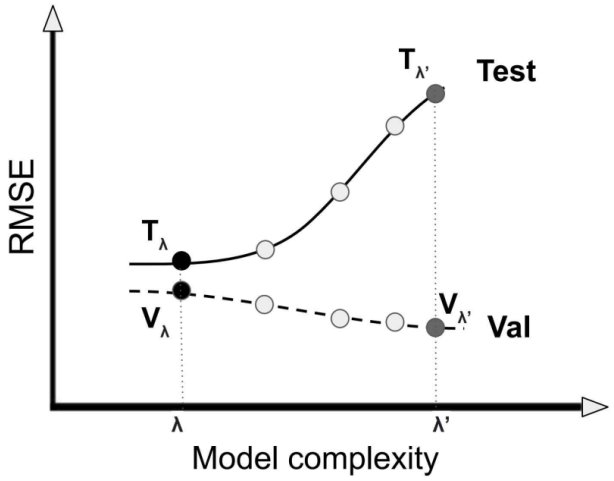

Figure 4. Schematic illustration of validation and testing RMSE values of predicted yield anomalies for an increasing model complexity obtained from the LTO ~~proceduce~~ procedure. For a fixed complexity level λ, two RMSE values are obtained: V_λ for validation and T_λ for testing datasets.

errors, ~~showed~~ shown in Fig. 3(a2). These two distributions result in a validation score (blue dot) and a testing score (red dot). The shape of these distributions give the average score and its uncertainty.

Finally, the n_{samp} testing choices give n_{samp} validation and n_{samp} testing scores that form a validation PDF in blue line, a testing PDF in red line, and thus the two scores V_λ and T_λ in Fig. 3(b).

A pseudo-code is provided in "Appendix A " to facilitate the implementation of the LTO procedure in any language.

3.2.3 Generalisation ability versus model complexity

The process represented in Fig. 3 is used to obtain the validation (V_λ) and testing (T_λ) scores from the LTO approach for a given model complexity λ. A different complexity level (different λ) results into different V_λ and T_λ values. The V_λ and T_λ evolution curves obtained for validation and testing RMSE values of yield anomalies for an increasing model complexity are presented in Fig. 4. For simplicity, only validation and testing scores will be discussed since the training error should ~~be~~ consistently decrease with ~~decreasing when increasing the~~ model complexity. When increasing the complexity level ($\lambda' > \lambda$), the validation error is smaller ($V_{\lambda'} < V_\lambda$) but the testing error is bigger ($T_{\lambda'} > T_\lambda$); this is typical from overfitting (Sect. 2.3.2). ~~It is also known as overfitting: the complexity level is too high; the model can highly fit the validation dataset but does not generalise well.~~ In the following applications (Sect. 4 and 5), we will study these evolution curves for different models with various complexity levels in order to identify the appropriate yield models for Robusta coffee and grain maize.

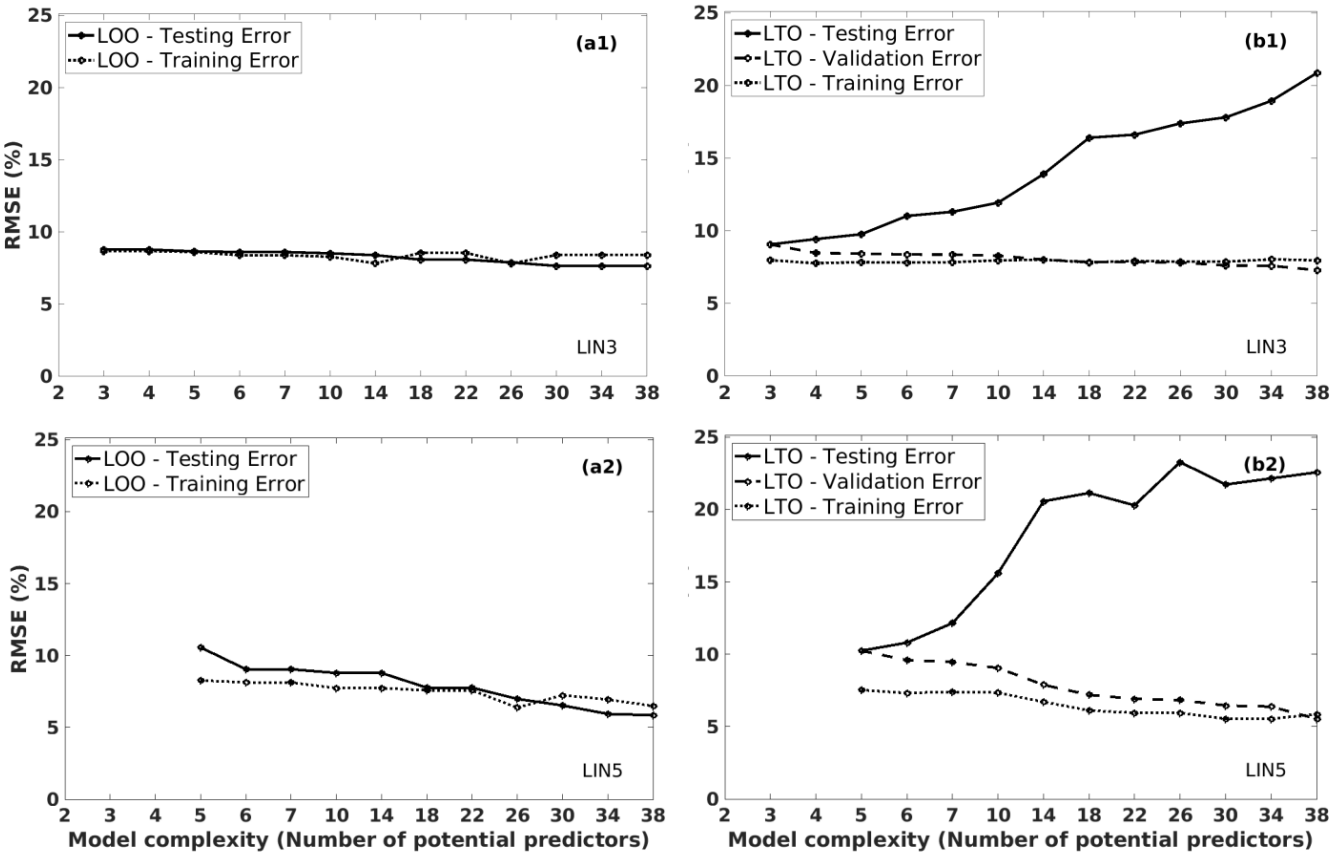

Figure 5. The training/validation/testing RMSE values of the predicted coffee yield anomalies using different models by adjusting the model complexity (increasing the number of potential predictors) in Cu M'gar (Dak Lak, Vietnam): (a1) and (a2) are induced from LOO procedure, (b1) and (b2) are induced from LTO procedure.

4 Robusta coffee in Cu M'gar

The first application concerns the statistical yield modelling of Robusta coffee in Cu M'gar (Dak Lak, Vietnam). The goal is to define a model that can predict the yield anomalies and then estimate its true applicability measured by a reliable generalisation score. ~~A model is defined by several factors, including the number of potential predictors on which the inputs are chosen, the actual number of inputs, or the model type.~~ We first assess several models with varying complexities to find the appropriate ~~model~~ complexity using LOO (Sect. 3.1) and LTO (Sect. 3.2) approaches.

4.1 Yield model selection

Two methods of estimating the model quality (LOO and LTO) are considered to choose the appropriate model complexity. As discussed in Sect. 2.3, this study will analyse several factors controlling the model complexity.

Number of potential predictors and number of inputs

Figure 5 shows the RMSE values of the predicted Robusta coffee yield anomalies for the ~~LIN3 and LIN5 models, which are the linear regression models with three and five inputs, respectively.~~ LIN models with the number of potential predictors ranging from 3 to 38 (on the horizontal axis). These values are computed using the LOO and LTO procedures for the training, validation, and testing datasets. Several models have been tested, we presente here two particular examples of LIN3 and LIN5 models, which are the linear regression models with three and five inputs, respectively. These inputs are selected among the considered potential predictors. For instance, for LIN3 model with six potential predictors, LOO and LTO aim at choosing three inputs among $\{P_{Nov(t-1)}, P_{Nov(t)}, T_{Mar(t)}, T_{Jan(t)}, T_{May(t)}, P_{Oct(t-1)}\}$. ~~The horizontal axis shows 13 models with a different number of potential predictors ranging from 3 to 38.~~ In addition, the comparison of LIN3 and LIN5 is representative for the examples when the number of inputs defines the model complexity.

The LOO procedure suggests that the more complex the model is, the better results are. Both training and testing RMSE values decrease gradually (Fig. 5(a1)) with the increase of the number of potential predictors for LIN3 models (although the training error shows fluctuations). It is even more obvious for LIN5 models: the testing RMSE value decreases by 5 % when increasing the model complexity from ~~five potential predictors~~ 5 to 38 potential predictors (Fig. 5(a2)). Thus, models with more inputs and more potential predictors would appear adequate when using the LOO procedure.

The LTO procedure is considered in Fig. 5(b1) and (b2), the training and validation RMSE values decrease with the model complexity, in a similar way as the training and testing errors in the LOO procedure. This similarity is because the LTO validation dataset has the same role as the LOO testing dataset: to find the best model! However, the testing errors do increase with the increase of the number of potential predictors (Fig. 5(b1)). This behaviour is ~~ever~~ even stronger for the LIN5 model ~~of~~ (Fig. 5(b2)). ~~Because~~ because the potential of overtraining is higher with a more complex model.~~, we~~ We observe a more significant difference between the testing errors and validation/training errors in this case (Fig. 5(b2)) than the LIN3 model (Fig. 5(b1)). The LTO procedure clearly indicates that a simpler model (i.e., a lower number of potential predictors) is more suitable. This conclusion makes sense ~~because~~ since it is inappropriate to use a very complex model (as the LOO model choice) when having only 19 samples.

The LOO procedure is actually misleading because it suffers from overfitting: it chooses the best model and assesses the generalisation ability on the same testing dataset. This overfitting issue is suppressed in the LTO procedure since we chose the model on the validation dataset and assessed its generalisation score on an independent testing dataset.

Model types

In another example, we increase the model complexity by not only the number of potential predictors but also the model type. A more complex model, a feedforward NN model (NN3 with three inputs and seven neurons in the hidden layer), is considered instead of a simple LIN model. ~~Another example using the NN models (NN3 with three inputs and seven neurons in the hidden layer) in Fig~~ Figure. 6(a) shows the same behaviour: the more complex the model is, the higher the testing error becomes due to the overtraining (The model ~~is stopped~~ stops at six potential predictors due to the computationally cost. More NN examples will be discussed in Sect. 5). For the same number of potential predictors, the testing errors in NN3 models (Fig. 6(a)) are much higher than those in LIN3 models (Fig. 5(b1)). The significant difference between training errors and validation/testing errors

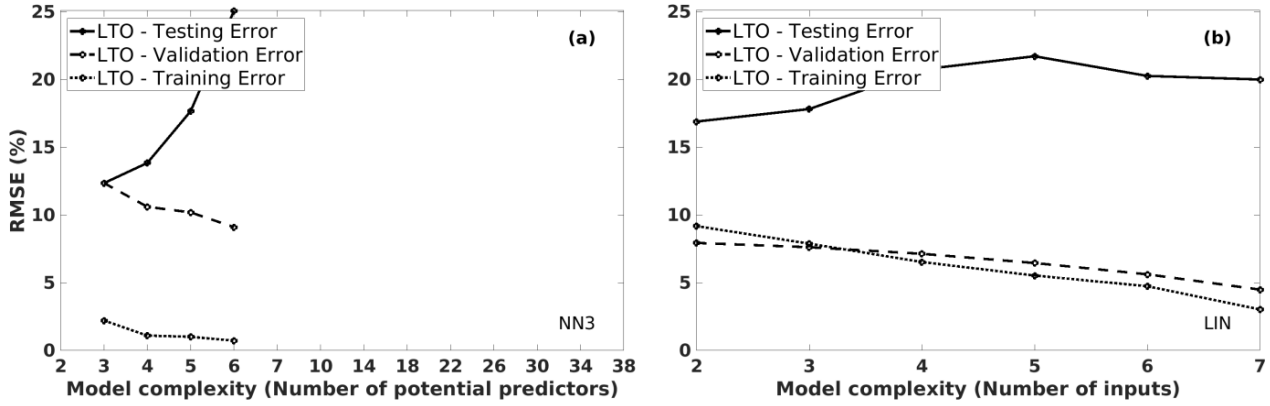

Figure 6. The training/validation/testing RMSE values of the predicted coffee yield anomalies using different models by adjusting the model complexity in Cu M'gar (Dak Lak, Vietnam): (a) - NN3 models (with seven neurons in the hidden layer) by increasing the number of potential predictors, (b) - LIN models (with 30 potential predictors) by increasing the number of inputs.

in NN3 models is related to the overfitting problem (compared to the LIN3 models). Using a NN model that is too complex for a limited database is highly dangerous.

Number of inputs

Finally, we test a case when the number of inputs defines model complexity. In this example, the number of potential predictors is fixed to 30 (n_{pre}=30). The number of inputs is chosen from two to seven as shown on the horizontal axis in Fig. 6(b). ~~We also tested the LTO procedure with an increased complexity level by keeping the same number of potential predictors (n_{pre}=30) but increasing the number of inputs from two to seven (on the horizontal axis in Fig. 6(b)).~~ In this case, the same behaviour is observed: the testing errors show an opposite trend with the training/validation errors and gradually increase with the model complexity.

In short, considering the limited information in the available database—that is used to train, select the model, and evaluate its quality—it is not possible to use more than a very simple and limited model. Therefore, for this 19-samples coffee yield modelling case, using a simple LIN model is better than a complex one (NN model, for instance), and it will be illusory to think that complex plant relations can be ~~implemented~~ exploited with such a limited number of samples.

4.2 Yield anomaly estimation

The previous section shows that ~~As shown in the previous section,~~ the LTO procedure allows us to ~~chose~~ choose a reasonable model, simple enough, with fewer potential predictors and inputs. Thus, the crop yield estimations of the LTO method will be assessed here to see how good the selected model (LIN3 model with three predictors) is. The final model includes {$P_{Nov(t-1)}$, $P_{Nov(t)}$, $T_{Mar(t)}$} and these selected variables coincide with the key moments of Robusta coffee. For example, there is the need for a dry period for the buds to develop into dormancy at the end of the development stage, i.e., Nov(t-1). Therefore, $P_{Nov(t-1)}$ impacts directly the buds, thus the potential yield. Similarly, the fruit maturation stage (Nov(t)) benefits from weather conditions

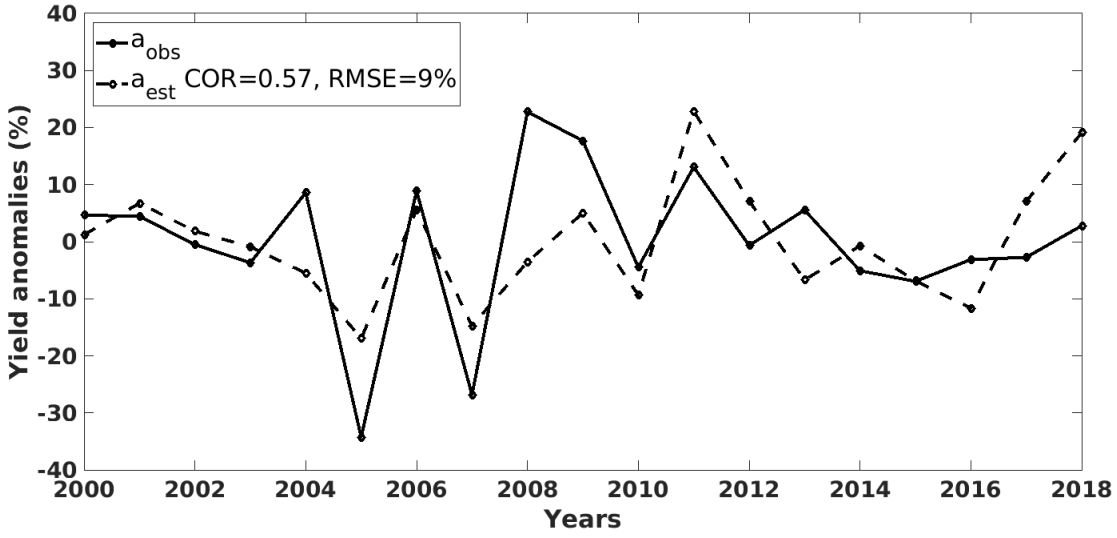

Figure 7. The observed (solid line) and LTO estimated (dashed line) coffee yield anomalies time series in Cu M'gar (Dak Lak, Vietnam).

with less precipitation. At the beginning of the fruit development period (Mar(t)), too low temperature slows maturation rate to the detriment of yield, while the higher temperature is beneficial.

Figure 7 presents the estimated yield anomalies time series for Robusta coffee in Cu M'gar from 2000 to 2018. The estimation (a_{est} in the dashed line) describes quite well the observations (a_{obs} in the solid line) with a correlation of 0.57. With ~~the~~ precipitation and temperature variables, the selected model is able to identify many extreme years (e.g., 2005-2009, 2010, 2011) or a decreasing trend from 2011 to 2015. Also, the correlation score means that the model can explain more than 30 % (0.57^2) of the variation in coffee yield anomalies, which is in agreement, for instance, with Dinh et al. (2021). This value is reasonable as the weather is among several factors (e.g., agricultural practices, diseases, irrigation ~~topography~~) ~~that affect the~~ affecting coffee yield. Climate could potentially explain a higher percentage of variability, with a more complex model. However, for that, we would need a longer historical record. It is possible to apply ~~such a~~ the resulting statistical crop yield model to future climate simulations and then study the impact of climate change on coffee (Bunn et al., 2015; Craparo et al., 2015a; Läderach et al., 2017). This would be the subject of a forthcoming study.

5 Grain maize over France

This application considers several aspects of grain maize over France. First, the sensitivity of the forecasting quality to the model complexity is studied, using the LOO and LTO approaches over the Bas-Rhin department, one of the major grain maize-producing departments in France. Then, the forecasting scores are investigated over the major grain maize-producing departments.

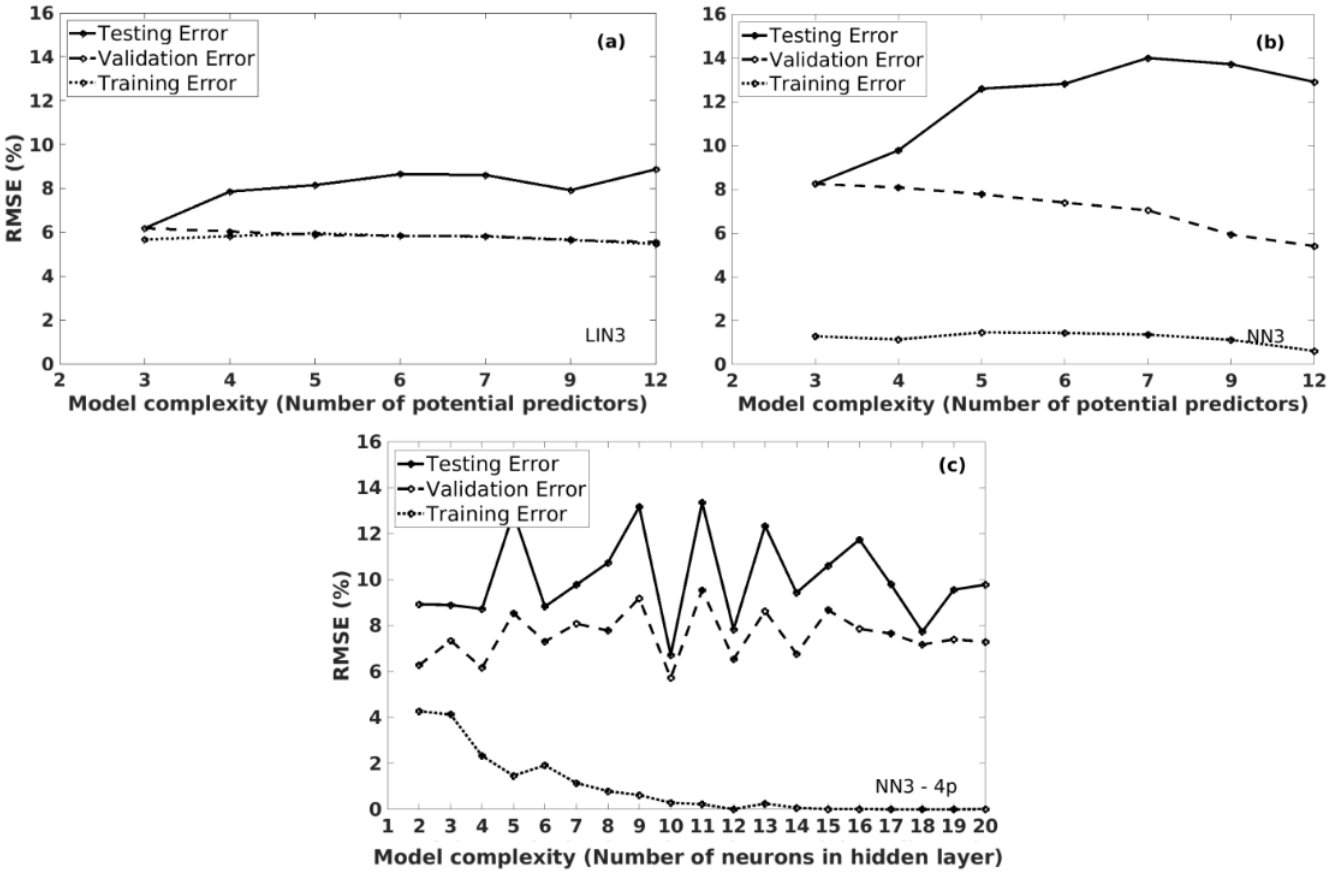

Figure 8. The training/validation/testing RMSE values of the predicted grain maize yield anomalies using different models by adjusting the model complexity in Bas-Rhin (France): (a) and (b) - the comparison between LIN3 and NN3 (with seven neurons in the hidden layer) models by increasing the number of potential predictors; (c) the NN3 with four potential predictors by increasing the number of neurons in the hidden layer.

5.1 Yield model selection - Focus on Bas-Rhin

This section aims to define an appropriate statistical model for grain maize using 22 years of yield data. This test is done over Bas-Rhin (i.e., one major grain maize-producing department in France). As shown in Sect. 4.1, the LOO approach is misleading ~~by~~ in choosing too complex models; we focus here on the LTO results for different models with various complexity levels. Figure 8 describes the RMSE values of the predicted grain maize yield anomalies for three datasets (training, validation, and testing) of the LTO procedure. The ~~for~~ LIN3 models with various complexity levels and several architectures of NN3 models are considered. The results of LIN3 models are presented in Fig. 8(a), and NN3 models (with seven neurons in the hidden layer) are in Fig. 8(b), with a different number of potential predictors ranging from ~~three~~ 3 to 12 in the horizontal axis. In the two cases, the LTO procedure shows a similar behaviour as for the Robusta coffee application ~~of the previous section~~

(Sect. 4.1): the validation/training errors decrease with the number of potential predictors, while the testing errors show an opposite trend. These overtraining ~~results~~ behaviours suggest that a simple model (e.g., LIN3 with three potential predictors) is more adequate: the testing RMSE value is small and close to the RMSE values over the two other datasets.

More complex models are tested in Fig. 8(c):~~, in which~~ NN3 models with four potential predictors ~~are considered~~. Here, the ~~The~~ model complexity ~~here~~ corresponds to the number of neurons in the hidden layer (from ~~two~~ 2 to 20 neurons) in the horizontal axis. The impact of overfitting (Sect. 2.3) is noticeable when the model is too complex. For instance, in Fig. 8(c), the training errors get smaller for more neurons in the hidden layer, as expected. However, the testing and validation errors show large fluctuations when increasing the number of neurons. The overfitting problem appears at the first step with two neurons in the hidden layer, ~~show~~ shown by the high testing error in Fig. 8(c). Same results (not shown) are obtained for NN3 models with n potential ~~preditors~~ predictors, where $n = 3,\ 7,\ 12$. Thus, the NN models are unreliable ~~in this case~~ due to the limited number of samples to train a non-linear model.

5.2 Reliability model assessment

In this section, a statistical yield model is applied first over 96 French departments to assess the true model quality. Then, we will focus on ten major departments to assess how the selected models perform for yield anomaly predictions.

Figure 9 shows the true testing RMSE maps of predicted grain maize yield anomalies in France. Here, the testing errors induced from the LTO procedure are used on the models chosen by the LOO and ~~the~~ LTO approaches. In other words, both methods (LOO and LTO) can be ~~used~~ considered to identify optimal crop models, but only the LTO method is used (as a reliable tool) to estimate the model generalisation ability. For example, when considering only LIN3 models, LOO chooses models with 12 potential predictors, while LTO chooses ~~those with three potential predictors~~ three. From these choices, the true model generalisation scores (i.e., testing errors) are estimated using the LTO approach, ~~showed~~ shown in the RMSE maps of Fig. 9(a1) and (b1). Another example focuses on LIN5 models (presented in Fig. 9(a2) and (b2)). The true errors obtained from the LOO choice are clearly higher than those from the LTO choice for LIN3 models. For instance, the testing RMSE values range from 10 % to 18 % in many departments in Fig. 9(a1), while these values are often lower than 10 % in Fig. 9(b1). This difference shows that the LOO approach under-estimates these true errors, as seen in Fig. 9(a1). Thus, the model choice of the LOO approach is misleading. For more complex models like LIN5 models—that is prefered by the LOO choice—in the second row of Fig. 9, the higher errors are observed, especially for LOO model errors of many northern departments with up to 22 % of RMSE (Fig. 9(a2)). This grain maize application confirms the benefit of LTO to select and assess the true quality of statistical yield models, while LOO is misleading by under-estimating the true errors of its selected models. A simple LIN3 model with three potential predictors is adequate for this application considering the limited amount of data.

We now analyse how good the LTO testing estimations are compared to the observations over ten major grain maize-producing departments (as ~~showed~~ shown in Fig. 1(d)). Figure 10 presents the boxplots of residuals for these departments, which are the differences between the observed and estimated yield anomalies (Residual=$a_{obs} - a_{est}$ in %). The medians of the residuals lie near zero. It means that the selected models can predict the yield anomalies with acceptable coverage and

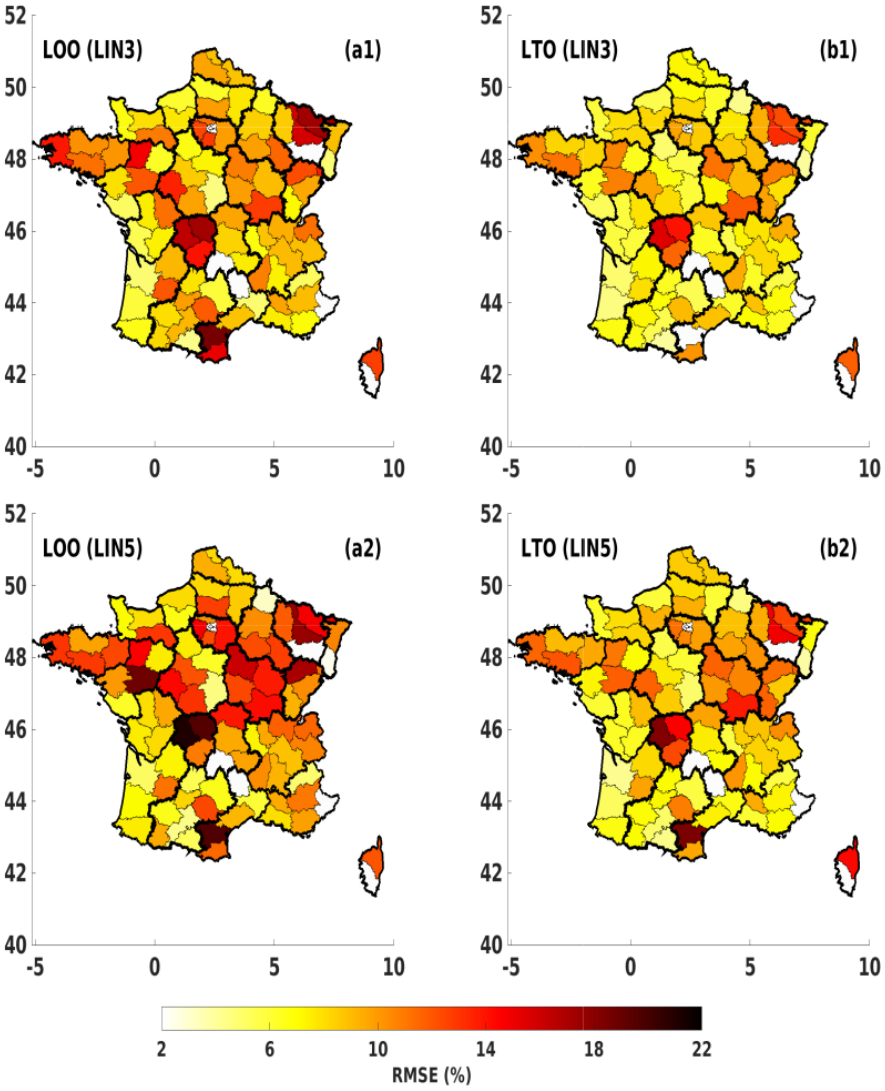

Figure 9. The true testing RMSE maps of predicted grain maize yield anomalies in France for LOO (first column) and LTO (second column) approaches, induced from two LIN models with a different number of inputs: LIN3 (first row) and LIN5 (second row).

420 precision. Although there are some extreme values (Lot-et-Garonne) and some outliers, the interquartile, which ranges from about -8 % to 8 %, shows slight differences between the observations and estimations over study departments.

5.3 Seasonal yield forecasting

The LTO approach is helpful for selecting an adequate model with better forecasting. Here ~~In the following~~, the model chosen by the LTO procedure is tested for seasonal forecasting, from the sowing time (April) to the forecasting months (i.e., from June

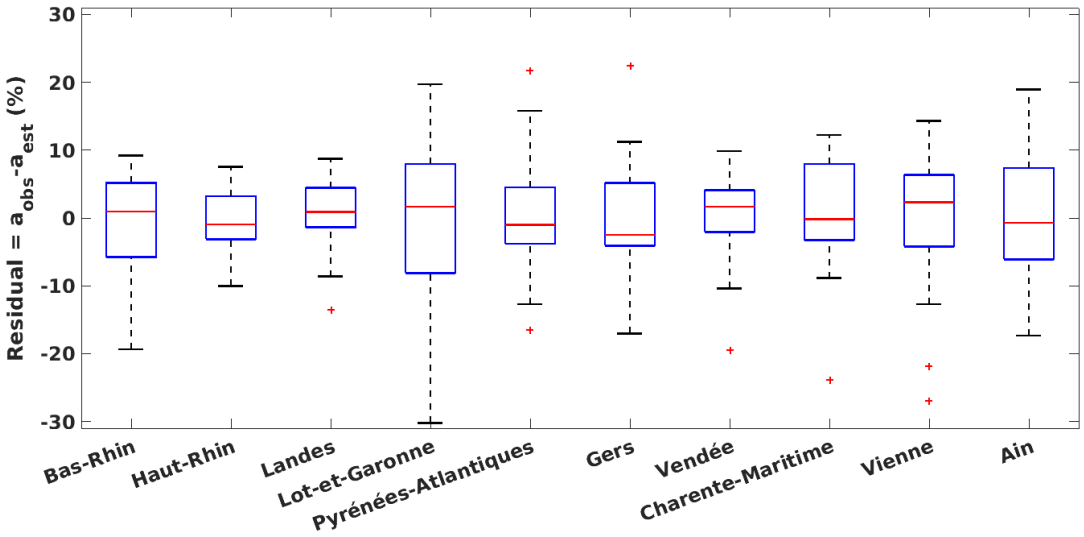

Figure 10. Boxplots of residuals (the difference between the observed and estimated yield anomalies) for ten major grain maize-producing departments: red horizontal bars are medians, boxes show the 25th-75th percentiles, error bars depict the minimum and maximum values, and red + signs are suspected outliers.

Departments	Forecasting months			
	June	July	August	September
Bas-Rhin	0.46	0.47	0.47	0.47
Haut-Rhin	0.35	0.53	0.53	0.53
Landes	0.63	0.64	0.66	0.67
Lot-et-Garonne	0.02	0.22	0.22	0.29
Pyrénées-Atlantiques	0.34	0.60	0.60	0.60
Gers	0.33	0.61	0.60	0.43
Vendée	0.63	0.63	0.63	0.63
Charente-Maritime	0.21	0.52	0.53	0.62
Vienne	0.39	0.40	0.40	0.40
Ain	0.17	0.52	0.52	0.52
Average	**0.35**	**0.51**	**0.52**	**0.52**

Table 1. The correlation between the observed and estimated yield anomalies for different forecasting months (from June to September), over ten major grain maize-producing departments.

425 to September).~~In this scenario, for example, for the June model,~~: all-weather variables (including P and T) from April to June can be selected for the June forecasting. Table 1 represents the correlations between the observed and estimated yield anomalies

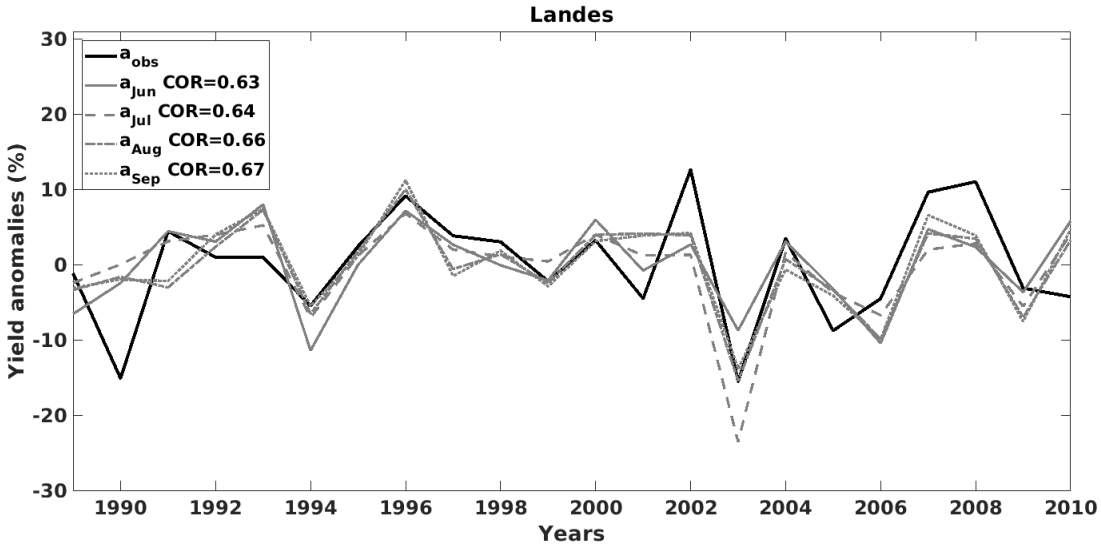

Figure 11. The observed (a_{obs}) and the estimated ~~grain maize~~ yield anomalies time series, for different ~~forecating~~ forecasting months from June to September (e.g., a_{Jun} means June forecasting), for grain maize in Landes (France).

of the forecasts from June to September. The quality of the seasonal forecasting models gradually increases when approaching the harvest because more information is provided. With the weather information at the beginning of the season (April, May, and June), the June forecasting model obtains an average correlation of 0.35 between the observations and estimations. This score is significantly improved when adding information of July (correlation of 0.51). This improvement means that the weather in July strongly influences grain maize yields. The improvement from July to August is much less than from June to July, with an average increase of ~~0.16 and 0.01~~ 0.01 and 0.16, respectively. No information is added in the September forecasting model since it coincides with the harvest time. In other words, the final model should consider only variables from April to August. As in our case, statistical model selects {T_{Jul}, P_{May}, P_{Apr}} as the final inputs for grain maize in the eatern region (Bas-Rhin, Haut-Rhin); {T_{Jul}, P_{Jul}, T_{Apr}} for the southern region (e.g., Landes, Pyrénées-Atlantiques, Gers); and {P_{Jul}, P_{Apr}, P_{Jun} or T_{Jun}} for the central part (Vendée, Charente-Maritime, Vienne). It is reasonable to have different inputs for different regions (or even departments) due to their distinct environmental conditions. In general, weather variables in July—the flowering period—are among the most influential variables. During this time, a high temperature affects the photosynthesis process, thus reducing the potential yield; in contrast, positive precipitation anomalies are preferable (Ceglar et al., 2016; Mathieu and Aires, 2018b). Precipitations in April and May also show significant impacts on grain maize as a water deficit during this vegetative stage decreases plant height (Çakir, 2004).

In addition, Fig. 11 shows time series plots of the yield anomaly observations and estimations for different forecasting months in Landes (France). In this case, the June forecasting results show a high correlation with the observed yield anomalies (0.63). This score slightly increases when approaching the harvest. It also indicates that the weather can explain more than

40 % ($0.67^2 = 44.89$ %) of variations in grain maize yield anomalies in this region, which is in line with other crop studies (Ray et al., 2015; Ceglar et al., 2017). However, the forecasting models cannot predict all the extremes (e.g., negative yield anomaly in 1990) that are probably influenced by the climate extremes ~~of climate~~ (Hawkins et al., 2013; Ceglar et al., 2016). The statistical models could be improved by adding the indices that focus on extreme weather events.

6 Conclusions and perspectives

Crop yield modelling is very useful in agriculture as it can help increase the yield, improve the production quality, and minimise the impact of adverse conditions. Statistical models are among the most used approaches with many advantages ~~(Lobell and Burke, 2010; Iizumi et al., 2013; Mathieu and Aires, 2018; Gaudio et al., 2019)~~. The main difficulty in this context is the limitation of the available crop databases to calibrate such statistical models. Applications typically rely on only two or three decades of data ~~(Prasad et al., 2006; Ceglar et al., 2016; Kern et al., 2018)~~. This small sample size issue directly impacts the complexity level that can be used in the statistical model: a model too complex cannot be fit with such limited data, and assessing the true model quality is also challenging. In practice, statistical inference requires three datasets: one for calibrating the model, a second one for choosing the right model (or tuning the model hyper-parameters), and another for assessing the true model generalisation skills ~~(Ripley, 1996)~~. Dividing a very small database into such three datasets is very difficult.

The LOO method has been used as a cross-validation tool to calibrate, select, and assess the model (Kogan et al., 2013; Zhao et al., 2018; Dinh et al., 2021). It was shown in this paper that LOO ~~could be~~ is highly misleading because it uses only one dataset to ~~chose~~ choose the best model and estimate its generalisation skills simultaneously. This is a true problem as LOO is one of the main statistical tools to obtain ~~good~~ crop yield models. This study proposes a particular form of nested cross-validation approach ~~in the form of what~~ that we call a LTO method. This method uses a complex folding scheme to estimate independent training, validation, and testing scores. ~~Results show that LOO is truly misleading and can artificially request prefer complex models that overfit the problem.~~ In contrast to LOO, LTO shows that only very simple models can be used when the database is limited in size. The LTO implementation proposed here is very general and can be applied to any statistical crop modelling application.

Two applications have been considered ~~here~~. The first one concerns the coffee yield modelling over a ~~district in Vietnam's~~ major Robusta coffee-producing ~~region~~ district in Vietnam. It was shown that considering the available historical yield record, we can only set up a statistical model that ~~monthly mean precipitation and temperature could~~ explains ~~more than~~ about 30 % of the coffee yield anomaly variability. The ~~70 %~~ remaining variability is due to non-climatic factors (agricultural practices, diseases, or political and social context). It also could come from climate; however, the model would require much more samples to go into deeper details of the climate-crop yield relationship. In addition, ~~E~~explaining a third of the coffee yield variability is in line with the literature (Ray et al., 2015; Craparo et al., 2015b; Dinh et al., 2021). LTO was able to identify the suitable model complexity ~~of the statistical model that can be~~ trained on the historical record and estimate the true model ability to predict yield on independent years. The final model includes $\{P_{Nov(t-1)}, P_{Nov(t)}, T_{Mar(t)}\}$, which corresponds to the key

moments of Robusta coffee: the end of the bud development, the fruit maturation, and the beginning of the fruit development, respectively.

The second application is related to grain maize over France. The LTO was used here to ~~chose~~ choose between simple linear models and more complex neural network models. Our findings also show~~ed~~ that LOO was misleading in overestimating the testing scores. LTO indicated that a simple linear model should be used and estimated the model generalisation ability correctly. This approach can also be helpful in seasonal forecasting applications (during the growing and the beginning of harvest seasons). In this application, the weather can explain more than 40 % of the yield anomaly variability, which is a reasonable score (Ray et al., 2015; Ceglar et al., 2017). This score can vary depending on study regions~~, e.g.,~~ because some regions are more sensitive to the climate than others. Generally, grain maize yield anomalies are mainly influenced by weather variables during the flowering period (July) and the early season (April).

In the future, the mixed-effects model can be considered instead of a straightforward ~~statical~~ statistical model~~s~~. This approach—which intends to use samples in several regions (e.g., gathering samples into groups) to compensate for the lack of historical data—could help us obtain more complex crop models (Mathieu and Aires, 2016). Such a mixed-effect could benefit from the LTO scheme. In terms of applications, the crop models that we derived here could be used on climate simulations (from an ensemble of climate models for the next 50 years) to investigate the crop yield sensitivity to ~~changing climate conditions~~ climate change. Other crops will be investigated, over France (e.g., wheat, oats, sunflower (Ceglar et al., 2016; Schauberger et al., 2018; Ceglar et al., 2020), over Europe (e.g., wheat, grain maize, barley (Lecerf et al., 2019), or globally (e.g., coffee (Bunn et al., 2015)). Furthermore, statistical crop models should benefit the definition of adaptation and mitigation strategies. For instance, it is expected that the climate runs could help us identify the change in optimality for the crop culture in the world.

Code availability. The Matlab code used to run an example of the leave-two-out method is available at the following Zenodo link for the revision process of GMD: https://zenodo.org/record/5159363 (Anh and Filipe, 2021).

Data availability. The coffee data were provided by ~~the~~ Vietnam's General Statistics Office (GSO) ~~of Vietnam~~ for the 2000-2018 period. These data are available from GSO on reasonable request. For any inquiries, please send an email to banbientap@gso.gov.vn. The data on French grain maize (and other French crops) are available at http://agreste.agriculture.gouv.fr from 1989 on. In addition, the weather data, i.e., ERA5-Land data, can be downloaded from https://cds.climate.copernicus.eu (last access: 22 Apr 2021).

Appendix A: Appendix

n_{samp} = number of samples; %years
n_{pre} = number of potential predictors;
n_{mod} = number of models;
n_{fold} = number of folds of the dataset;
Score(2,n_{fold},n_{mod}); %representing RMSE or COR; 2 for [Test,Val];

```
bm = best model ϵ {1,⋯,n_mod};
%Step 1: Build scores for each fold, each model
for inp = 1 to n_fold
    %Define the folding process
    Test = 1 sample ϵ {1,⋯,n_samp};
    Val = 1 sample ϵ {1,⋯,n_samp} - Test;
    Learn = {1,⋯,n_samp} - Test - Val;
    for imod = 1 to n_mod
        %Train models
        model = train(model, Learn);
        Score(1,inp,imod) = RMSE(model, Test);
        Score(2,inp,imod) = RMSE(model, Val);
    end
end
%Step 2: Choose best model for all folds; estimate its score
for isamp = 1 to n_samp
    Mean_Val = mean(Score(2,n_fold{isamp},:)); %(1,1,n_mod)
    ibm(isamp) = argmin_i(Mean_Val);
    Score_Test(isamp) = mean(Score(1,n_fold{isamp},ibm(isamp))); Test score
    Score_Val(isamp) = mean(Score(2,n_fold{isamp},ibm(isamp))); Val score
end
FinalScore_Test = mean(Score_Test)
FinalScore_Val = mean(Score_Val)
```

Author contributions. All authors conceptualized the research and formulated the model. LADT implemented the model in Matlab and analyzed the output with FA. All authors contributed to writing the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work is a part of Lan Anh's Ph.D., which benefited from the French state aid managed by the ANR under the "Investissements d'avenir" program with the reference ANR-16-CONV-0003, and the Australian Centre for International Agricultural Research, Small Research Activity (SLAM/2018/209) conducted by the International Center for Tropical Agriculture (CIAT).

References

Agri4cast: Crop Calendar, https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx?o=, Accessed 20 Jun 2021, 2021.

Allen, D. M.: The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction, Technometrics, 16, 125–127, https://doi.org/10.1080/00401706.1974.10489157, 1974.

Amarasinghe, U. A., Hoanh, C. T., D'haeze, D., and Hung, T. Q.: Toward sustainable coffee production in Vietnam: More coffee with less water, Agricultural Systems, 136, 96–105, https://doi.org/10.1016/j.agsy.2015.02.008, 2015.

Anh, D. T. L. and Filipe, A.: Code and Data for the Leave-Two-Out Method, https://doi.org/10.5281/zenodo.5159363, 2021.

Beillouin, D., Schauberger, B., Bastos, A., Ciais, P., and Makowski, D.: Impact of extreme weather conditions on European crop production in 2018, Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 375, 20190510, https://doi.org/10.1098/rstb.2019.0510, 2020.

Bishop, C. M.: Neural Networks for Pattern Recognition, Oxford University Press, Inc., USA, 1995.

Bunn, C., Laderach, P., Ovalle Rivera, O., and Kirschke, D.: A bitter cup: climate change profile of global production of Arabica and Robusta coffee, Climatic Change, 129, 89–101, https://doi.org/10.1007/s10584-014-1306-x, 2015.

Cawley, G. C. and Talbot, N. L.: On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, J. Mach. Learn. Res., 11, 2079–2107, 2010.

Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., and Dentener, F.: Impact of meteorological drivers on regional inter-annual crop yield variability in France, Agricultural and Forest Meteorology, 216, 58–67, https://doi.org/10.1016/j.agrformet.2015.10.004, 2016.

Ceglar, A., Turco, M., Toreti, A., and Doblas-Reyes, F. J.: Linking crop yield anomalies to large-scale atmospheric circulation in Europe, Agricultural and Forest Meteorology, 240-241, 35–45, https://doi.org/10.1016/j.agrformet.2017.03.019, 2017.

Ceglar, A., Zampieri, M., Gonzalez-Reviriego, N., Ciais, P., Schauberger, B., and Van Der Velde, M.: Time-varying impact of climate on maize and wheat yields in France since 1900, Environmental Research Letters, https://doi.org/10.1088/1748-9326/aba1be, 2020.

Craparo, A., Asten, P. V., Laderach, P., Jassogne, L., and Grab, S.: Coffea arabica yields decline in Tanzania due to climate change: Global implications, Agricultural and Forest Meteorology, 207, 1–10, https://doi.org/10.1016/j.agrformet.2015.03.005, 2015a.

Craparo, A., Asten, P. V., Läderach, P., Jassogne, L., and Grab, S.: Coffea arabica yields decline in Tanzania due to climate change: Global implications, Agricultural and Forest Meteorology, 207, 1–10, https://doi.org/10.1016/j.agrformet.2015.03.005, 2015b.

Dinh, T. L. A., Aires, F., and Rahn, E.: Statistical analysis of the weather impact on Robusta coffee yield in Vietnam, Agricultural and Forest Meteorology (under review for publication), 2021.

EUROSTAT: Database in Agriculture, forestry and fisheries, https://ec.europa.eu/eurostat/web/products-datasets/-/tag00093, Accessed 22 Sep 2021, 2021.

FAO: FAOSTAT Crops production database, http://www.fao.org/faostat/en/#home, Accessed 22 Apr 2020, 2019.

Gaudio, Escobar-Gutiérrez, A. J., Casadebaig, P., Evers, J. B., Gérard, F., Louarn, G., Colbach, N., Munz, S., Launay, M., Marrou, H., Barillot, R., Hinsinger, P., Bergez, J. E., Combes, D., Durand, J. L., Frak, E., Pagès, L., Pradal, C., Saint-Jean, S., van der Werf, W., and Justes, E.: Current knowledge and future research opportunities for modeling annual crop mixtures : A review, arXiv, 2019.

Gornott, C. and Wechsung, F.: Statistical regression models for assessing climate impacts on crop yields: A validation study for winter wheat and silage maize in Germany, Agricultural and Forest Meteorology, 217, 89–100, https://doi.org/10.1016/j.agrformet.2015.10.005, 2016.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference and prediction, Springer, http://www-stat.stanford.edu/~tibs/ElemStatLearn/, 2009.

Hawkins, E., Fricker, T. E., Challinor, A. J., Ferro, C. A., Ho, C. K., and Osborne, T. M.: Increasing influence of heat stress on French maize yields from the 1960s to the 2030s, Global Change Biology, 19, 937–947, https://doi.org/10.1111/gcb.12069, 2013.

Hersbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Alonso-Balmaseda, M., Balsamo, G., Bechtold, P., Berrisford, P., Bidlot, J.-R., de Boisséson, E., Bonavita, M., Browne, P., Buizza, R., Dahlgren, P., Dee, D., Dragani, R., Diamantakis, M., Flemming, J., Forbes, R., Geer, A. J., Haiden, T., Hólm, E., Haimberger, L., Hogan, R., Horányi, A., Janiskova, M., Laloyaux, P., Lopez, P., Munoz-Sabater, J., Peubey, C., Radu, R., Richardson, D., Thépaut, J.-N., Vitart, F., Yang, X., Zsótér, E., and Zuo, H.: Operational global reanalysis: progress, future directions and synergies with NWP, https://doi.org/10.21957/tkic6g3wm, 2018.

Iizumi, T., Sakuma, H., Yokozawa, M., Luo, J. J., Challinor, A. J., Brown, M. E., Sakurai, G., and Yamagata, T.: Prediction of seasonal climate-induced variations in global food production, Nature Climate Change, 3, 904–908, https://doi.org/10.1038/nclimate1945, 2013.

Kath, J., Byrareddy, V. M., Craparo, A., Nguyen-Huy, T., Mushtaq, S., Cao, L., and Bossolasco, L.: Not so robust: Robusta coffee production is highly sensitive to temperature, Global Change Biology, https://doi.org/10.1111/gcb.15097, 2020.

Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., Bognár, P., and Lichtenberger, J.: Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices, Agricultural and Forest Meteorology, 260-261, 300–320, https://doi.org/10.1016/j.agrformet.2018.06.009, 2018.

Kogan, F., Kussul, N., Adamenko, T., Skakun, S., Kravchenko, O., Kryvobok, O., Shelestov, A., Kolotii, A., Kussul, O., and Lavrenyuk, A.: Winter wheat yield forecasting in Ukraine based on Earth observation, meteorologicaldata and biophysical models, International Journal of Applied Earth Observation and Geoinformation, 23, 192–203, https://doi.org/10.1016/j.jag.2013.01.002, 2013.

Kuhn, M. and Johnson, K.: Applied predictive modeling, Springer, 2013.

Läderach, P., Ramirez–Villegas, J., Navarro-Racines, C., Zelaya, C., Martinez–Valle, A., and Jarvis, A.: Climate change adaptation of coffee production in space and time, Climatic Change, 141, 47–62, https://doi.org/10.1007/s10584-016-1788-9, 2017.

Lecerf, R., Ceglar, A., López-Lozano, R., Van Der Velde, M., and Baruth, B.: Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe, Agricultural Systems, 168, 191–202, https://doi.org/10.1016/j.agsy.2018.03.002, 2019.

Li, Y., Guan, K., Yu, A., Peng, B., Zhao, L., Li, B., and Peng, J.: Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the U.S, Field Crops Research, 234, 55–65, https://doi.org/https://doi.org/10.1016/j.fcr.2019.02.005, 2019.

Lobell, D. B. and Burke, M. B.: On the use of statistical models to predict crop yield responses to climate change, Agricultural and Forest Meteorology, 150, 1443–1452, https://doi.org/https://doi.org/10.1016/j.agrformet.2010.07.008, 2010.

Mathieu, J. A. and Aires, F.: Statistical weather-impact models: An application of neural networks and mixed effects for corn production over the United States, Journal of Applied Meteorology and Climatology, 55, 2509–2527, https://doi.org/10.1175/JAMC-D-16-0055.1, 2016.

Mathieu, J. A. and Aires, F.: Using Neural Network Classifier Approach for Statistically Forecasting Extreme Corn Yield Losses in Eastern United States, Earth and Space Science, 5, 622–639, https://doi.org/10.1029/2017EA000343, 2018a.

Mathieu, J. A. and Aires, F.: Assessment of the agro-climatic indices to improve crop yield forecasting, Agricultural and Forest Meteorology, 253-254, 15–30, https://doi.org/https://doi.org/10.1016/j.agrformet.2018.01.031, 2018b.

Olesen, J., Børgesen, C., Elsgaard, L., Palosuo, T., Rötter, R. P., Skjelvåg, A., Peltonen-Sainio, P., Börjesson, T., Trnka, M., Ewert, F., Siebert, S., Brisson, N., Eitzinger, J., Asselt, E., Oberforster, M., and Van der Fels-Klerx, H. I.: Changes in time of sowing, flowering and maturity of cereals in Europe under climate change, Food Additives & Contaminants: Part A, 29, 1527–42, https://doi.org/10.1080/19440049.2012.712060, 2012.

Prasad, A. K., Chai, L., Singh, R. P., and Kafatos, M.: Crop yield estimation model for Iowa using remote sensing and surface parameters, International Journal of Applied Earth Observation and Geoinformation, 8, 26–33, https://doi.org/10.1016/j.jag.2005.06.002, 2006.

Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C.: Climate variation explains a third of global crop yield variability, Nature Communications, 6, 1–9, https://doi.org/10.1038/ncomms6989, 2015.

Ripley, B. D.: Pattern Recognition and Neural Networks, Cambridge University Press, https://doi.org/10.1017/CBO9780511812651, 1996.

Schauberger, B., Ben-Ari, T., Makowski, D., Kato, T., Kato, H., and Ciais, P.: Yield trends, variability and stagnation analysis of major crops in France over more than a century, Scientific Reports, 8, 1–12, https://doi.org/10.1038/s41598-018-35351-1, 2018.

Schmidhuber, J.: Deep learning in neural networks: An overview, Neural Networks, 61, 85–117, https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.003, 2015.

Siebert, S., Kummu, M., Porkka, M., Döll, P., Ramankutty, N., and Scanlon, B. R.: A global data set of the extent of irrigated land from 1900 to 2005, Hydrology and Earth System Sciences, 19, 1521–1545, https://doi.org/10.5194/hess-19-1521-2015, 2015.

Stone, M.: Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society: Series B (Methodological), 36, 111–133, https://doi.org/10.1111/j.2517-6161.1974.tb00994.x, 1974.

USDA: Coffee: World Markets and Trade, https://downloads.usda.library.cornell.edu/usda-esmis/files/m900nt40f/sq87c919h/8w32rm91m/coffee.pdf, accessed 22 Apr 2020, 2019.

Wintgens, J. N.: Coffee: Growing, Processing, Sustainable Production: A Guidebook for Growers, Processors, Traders, and Researchers, 2004.

Zhao, Y., Vergopolan, N., Baylis, K., Blekking, J., Caylor, K., Evans, T., Giroux, S., Sheffield, J., and Estes, L.: Comparing empirical and survey-based yield forecasts in a dryland agro-ecosystem, Agricultural and Forest Meteorology, 262, 147–156, https://doi.org/https://doi.org/10.1016/j.agrformet.2018.06.024, 2018.

Çakir, R.: Effect of water stress at different development stages on vegetative and reproductive growth of corn, Field Crops Research, 89, 1–16, https://doi.org/https://doi.org/10.1016/j.fcr.2004.01.005, 2004.